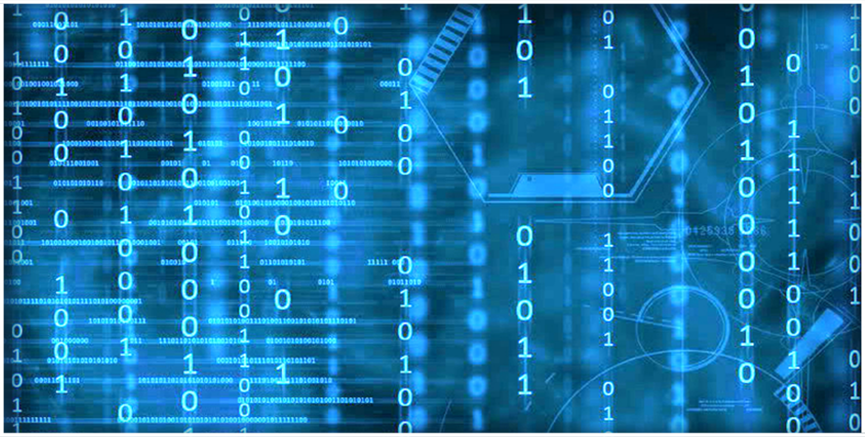
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 12 Issue 5

May 2021



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 12 Issue 5 May 2021
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Rain Attenuation in 5G Wireless Broadband Backhaul Link and Develop (IoT) Rainfall Monitoring System

Konstantinos Zarkadas[1],George Dimitrakopoulos[2]
Department of Informatics and Telematics, Harokopio University, Athens, Greece

***Abstract*—Climate change is the cause of more frequent and intense rainfall where they affect wireless communications because they cause severe weakening of the power of the emitted signal. These losses reduce network coverage and, therefore, system availability. The proposed solution is to integrate an Internet of Things (IoT) rainfall monitoring system where it will be able to collect real-time data on the height of rain that erupts in a particular place. This data will help areas where base stations install and the distance of the link that may need to be changed to reduce rainfall's harmful effects. So, the prediction of attenuation due to rain is an essential parameter in both terrestrial and satellite connections. The present study uses the ITU-R P 838 and ITU-R P 530 models to theoretically calculate losses in a 5G wireless broadband link with 99.9% link availability. This study conducts three frequency bands, 24 GHz, 28 GHz, and 38GHz, in Palo Alto, California. The travel distance is 5km, while the rainfall rate for the analyzed area is in zone D. The results show that the attenuations are proportional to the frequency, polarization, and rainfall rate.**

Keywords*—*Rain attenuation; Internet of Things; wireless broadband

I. Introduction

Heavy rains occur more frequently, one of the many global warming [1], [2]. It is no longer a scenario that was to show its effects in the distant future. The situation is now difficult, and its first consequences are appearing. Studies have shown that global warming has a significant effect on precipitation. As early as 2011, the atmosphere's ability to retain water from water vapor increased by 7% for every 1 ° C of heating [3]. An essential factor in landslides in a geographical area is the seasonal period where the percentage of water is different, as revealed by research that has been studied [4]. More specifically, for the North American region, the rains during the summer months are more in the summer than in the winter. It also noted that energetic precipitation has been increasing both in frequency and rhythm since 1901.

Rainfall varies from year to year and is distinguished by the rate, duration, and amount of water that sometimes negatively impact crops, telecommunications, energy, and many other human activities and infrastructure. There are losses generated by electromagnetic waves transmitted during wireless communication in the telecommunications sector due to the absorption and scattering [5]. As a result, the coverage area and the efficiency of the wireless communication network are limited. The high frequencies in the microwave zone, millimeter-wave already used for 5G technology, are significantly affected by signal attenuation due to rain [6]. It studies that frequencies above 10 GHz are significantly affected [7]. It is evident that climatic conditions significantly affect systems exposed or affected under the right weather conditions and cause malfunctions in the network. From the research so far, it cannot be accurately calculated that the amount of water will evolve on a rainy day. Even meteorological stations with their technological advancement in mathematical models may not have accurate data on the intensity or duration of rainfall or a storm. In recent years, smart meteorological stations have been created that calculates the intensity of the rain but not its height. So, the primary goal of the document is to design and implement a rainfall monitoring system. This solution promotes an IoT system that detects rainfall in real-time and provides information through a cloud service called Adafruit IO. This measured inferred whether the rate of rain would increase or not—this communication base on the MQTT protocol and the development of interfaces from the platform.

This article is structured as follows. Section 2 refers to the related work done so far and the theoretical background of rain attenuation. Section 3 shows the measurement setup and the theoretical measurements of the microwave link. In Section 4 present the experimental setup and analyze the IoT system of rain. The results will provide and discussed in Section 5. Finally, Section 6 contains concluding remarks and an outlook on future work potentials.

II. Theoretical Background

A. Related Work

All Scientists and engineers are concerned about limiting or controlling the damage caused by heavy rainfall. They are developing search and alert applications for years with the help sensors and microcontrollers connect on the internet and transfer information to and from the user. In [8], an application is developed, which is essentially a meteorological station that informs about the meteorological changes such as the measurement of temperature, humidity, rainfall, carbon monoxide, altitude, and LPG in the environment. It consists of a microcontroller Arduino Mega where it connects to the individual sensors that, when stimulated, will give the message on the LCD screen. Besides, the device sends SMS to the user and informs them about the prices collected with a GSM module's help. In the end, the results compared with the data of the national meteorological company.

In [9], a system creates to control the rate of rainfall. It is a system that has the structural features of a rain gauge. It consists of a funnel that collects the water where it falls into a sensor—also connected to the EZ430-RF2500 where it contains the MSP430F2274 microcontroller and a CC2500 transceiver for data transmission, which, in turn, is sent to another transceiver which is the central station connected to a laptop (Laptop computer unit) in C language in GUI graphical interface.

Even in [10], a system works as a rain gauge created; it consists of the detection unit with a sensor that works as a rain gauge. It consists of two parts, a funnel and a container that collects water. There is an Arduino Uno microcontroller that recovers data from the sensors. This system is even environmentally friendly as solar collector power it, and on days when there is no sunshine, the battery that has already charged use. There are three ways to communicate with the server: The station communicates through the 3G network with an Arduino Ethernet card and a 3G modem. The second through a 433 MHz free band, which use for industrial or scientific purposes. Finally, it will be easier to use the GSM network with a GSM shield with an Arduino card.

Increasingly, the technological advancement of sensors in both field of application, cost, the size has contributed to the development of many constructions at both research and mechanical level. In particular, projects to detect floods or weather conditions in areas such as crops, livestock units, renewable energy units (e.g., solar parks), and even smart cities. Also, 3G and 4G technologies have further promoted the implementation of many such applications. Today, Internet of Things (IoT) technology is making its presence felt, which is greatly aided by the infrastructure, efficiency, and flexibility offered by 5G technology, equipped with high speeds, capacity, low latency, which offers reliability in telecommunications systems. Therefore, due to this technological boom and the reliability of wireless systems that should provide uninterruptedly, this article gives weight to studying the effects of rain on a point-to-point microwave connection. This connection has fifth technology features, and the data collected by calculations is from three different frequencies with an ordinary rainfall rate (mm / h). The difference with the research so far is that the rain detection systems have to do with developing a meteorological station or the creation exclusively of a rain meter where the rain rate calculates directly. This need in the present study was by developing an IoT system for calculating the amount of rain (mm) that cleans itself every 1 second and informs in real-time about the rainfall course. Besides, a microcontroller uses where it collects information. The rain sensor and the help of a servo motor clean the flat surface to receive the next measurement. Finally, this system can interconnect the hardware through a cloud platform, accessible to the user. It will distinguish the sensor output and act immediately for any failures in terms of signal coverage or, in general, to analyze data on the availability of the broadband network.

B. Background

The signal attenuation calculates with appropriate models that help to calculate in theoretical attenuations [11]. The International Telecommunication Union (ITU) has helped facilitate international interconnection, spectrum sharing, and technical standards to ensure networks and technologies' smooth operation. Based on international recommendations (ITU-R P.838-3) [12], the specific attenuation of the γR (dB / km) signal obtained by the rainfall rate R (mm / h) from the following formula:

$$\gamma R \ = \ KRa \ (dB/km) \qquad (1)$$

The values k and a are coefficients that set depending on the operating frequency and the field polarization, defined by ITU-R P. 838 -3 in Table I.

The rate of rainfall comes in two ways, either from meteorological stations through individual measurements that calculate the average intensity per year, or from the categorization made by ITU - R. 837.1 [13] in acceptable climate zones in Table II.

In more detail, if a radio cluster with 99.99% availability is studied, it should be ensured that the connection will be available even at high rainfall rates that cause attenuation, with a probability of occurrence exceeding 0.01% of the time.

Therefore, after calculating the specific attenuation of the γR (dB / km) signal for horizontal and vertical polarization with the help of Table I, the ground rate r factor is defined by the standard (ITU-R P 530.), should be calculated [14] and given by the following formula:

$$r = \frac{1}{1+\frac{d}{d_0}} \qquad (2)$$

TABLE I. FACTORS K AND ACCORDING TO THE ITU-R P.838-3 STANDARD

Frequency (GHz)	K_H	a_H	K_V	a_V
20	0.09164	1.0568	0.09611	0.9847
21	0.1032	1.0447	0.1063	0.9771
22	0.1155	1.0329	0.1170	0.9700
23	0.1286	1.0214	0.1284	0.9630
24	0.1425	1.0101	0.1404	0.9561
25	0.1571	0.9991	0.1533	0.9491
26	0.1724	0.9884	0.1669	0.9421
27	0.1884	0.9780	0.1813	0.90349
28	0.2051	0.9679	0.1964	0.9277
32	0.2778	0.9302	0.2646	0.8981
34	0.3171	0.9129	0.3026	0.8834
36	0.3580	0.8967	0.3427	0.8690
38	0.4001	0.8816	0.3844	0.8552

TABLE II. THE RATE OF RAINFALL IN (MM /H) DEEPENING ON THE PERCENTAGE OF TIME PER ZONE

Percentage of time (%)	B	C	D	E	F	G	H	J	K	L
1.0	0.5	0.7	2.1	0.6	1.7	3	2	8	15	2
0.3	2	2.8	4.5	24	45	7	4	13	42	7
0.1	3	5	8	6	8	12	10	20	12	15
0.03	6	9	13	12	15	20	18	28	23	33
0.01	12	15	19	22	28	30	32	35	42	60
0.003	21	26	29	41	54	45	55	45	70	105

where d (km), the actual length of the path, while d0 is defined as follows:

$$\begin{cases} 35e^{-0.015R_{0.01}}\, R_{0.01<100mm/h} \\ \quad 35^{-1.5}R_{0.01\geq 100mm/h} \end{cases} \psi \qquad (3)$$

If R0.01> 100 mm / h, place the value 100 mm / h in place of R0.01 The length of the current path calculated:

$$d_{eff} = r * d\ (km) \qquad (4)$$

Thus, the total value of the losses is calculated as follows:

Horizontal polarization

$$L^H = \gamma^H_{R0.01} * d_{eff}\ (dB) \qquad (5)$$

Vertical polarization

$$L^V = \gamma^V_{R0.01} * d_{eff}(dB) \qquad (6)$$

C. Measurement Setup

1) Description of area: The selected area is Palo Alto (37 26 13 N - 122 07 48W), located northwest of Santa Clara County, California, USA. It locates in the center of Northern California, and the choice of this location is not accidental. Heavy rains in northern California expect to increase in the coming years [15]. Also, the geographical area belongs to zone D, according to the ITU - R. standard 837.1.

2) Experimental data: The scenario concerns a wireless broadband link with Line of Sight (LOS), where the route length is 5 km. Both the transmitter and the receiver operate at 24, 28, and 38 GHz. These frequencies are ideal. After all, they work in the 5G spectrum because they have lower oxygen uptake rates [16]. The equipment is iPASOLINK NEC VR4 based on fifth-generation (5G) technical specifications Fig. 1. This equipment is already used by telecommunications providers as a reliable wireless backhaul solution to support a fifth-generation (5G) network. It consists of an external ODU type IAG unit with channel bandwidth up to 112MHz, low power consumption - 13 ~ 38GHz: 42 dBm (30% lower than the current version), modulation 2048 QAM, with a capacity of 1Gbps and a bandwidth of 112 MHz, elements that are equally compatible with NR 5G Fig. 2.

Fig. 1. Location of Wireless Broadband Backhaul Link.

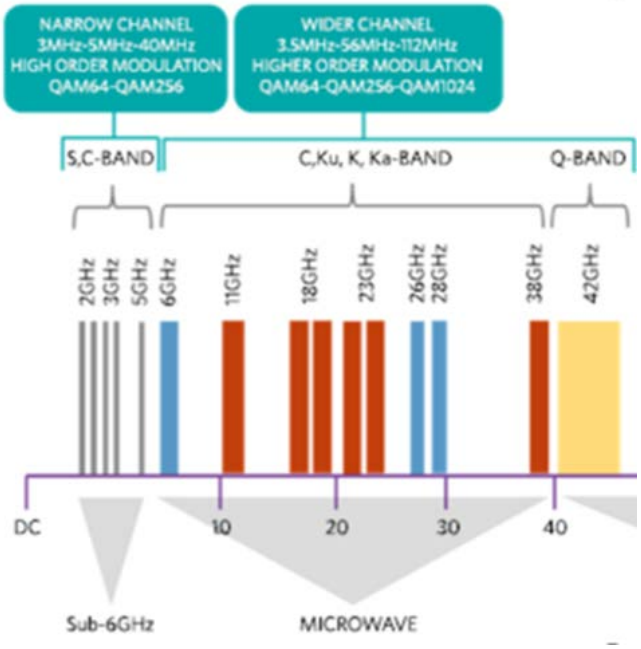


Fig. 2. Characteristics of Microwave Wireless Backhaul [17].

D. Theoretical Measurements

At this point, rainfall attenuation with coupling availability of 99.99% is calculated, with a distance of 5km for each of three operating frequencies:

- 38 GHz

Horizontal polarization

According to Table II, the coefficients K and α at 38GHz receive the following values.

KH = 0.4001, α = 0.8816

With the help of relation (1) the losses per km are:

$$\gamma^H_{R0.01} = K_H * R^{\alpha}_{0.01} = 0{,}4001 * 19^{0{,}8816} = 5{,}36\ (\frac{dB}{km})$$

Vertical polarization

Similarly, as before, the losses for vertical polarization are calculated

$$\gamma^V_{R0.01} = K_V * R^{\alpha}_{0.01} = 0{,}3844 * 19^{0{,}8552} = 4{,}76(\frac{dB}{km})$$

To calculate the total losses of horizontal and vertical polarization, the losses per km (as done earlier) and the current deff path's length must be measured. It is useful because it shows that along with the link, there is heterogeneity of meteorological conditions (e.g., more torrential rain elsewhere, not elsewhere), it is always less than the actual length. The current deff path calculate by multiplying the factor r by the actual length d.

After first calculating the sufficient path length through rain, for $R_{0.01}$<100mm/h:

$$d_0 = 35e^{-0{,}015*R0{,}01} = 35e^{-0{,}015*19} = 26{,}32$$

Thus, the reduction factor r:

$$r = \frac{1}{1+\frac{d}{d_0}} = \frac{1}{1+\frac{5}{26{,}32}} = 0{,}84$$

Therefore, the length of the active path can now set:

$$d_{eff} = r * d = 0{,}84 * 5 = 4{,}2km$$

So the total losses for each polarization are calculated separately:

Horizontal polarization

$$L^H_{R0,01} = \gamma^H_{R0,01} * d_{eff} = 5{,}36\left(\frac{dB}{km}\right) * 4{,}2(km) = 22{,}51dB$$

Vertical polarization

$$L^V_{R0,01} = \gamma^V_{R0,01} * d_{eff} = 4{,}76\left(\frac{dB}{km}\right) * 4{,}2(km) = 20dB$$

- 24 GHz

Horizontal polarization

According to Table II, the coefficients K and α at 24GHz take the following values. With the help of relation (1) the losses per km are:

$$\gamma^H_{R0.01} = K_H * R^a_{0.01} = 0{,}142 * 19^{1{,}01} = 2{,}77\ (\frac{dB}{km})$$

Vertical polarization

Similarly, as before, the losses for vertical polarization are calculated

$$\gamma^V_{R0.01} = K_V * R^a_{0.01} = 0{,}140 * 19^{0{,}95} = 2{,}29(\frac{dB}{km})$$

To calculate the total losses of horizontal and vertical polarization, the losses per km (as was done earlier) and the current deff path's length must be measured—the current deff path calculated by multiplying the factor r by the actual length d.

After first calculating the sufficient path length through rain, for $R_{0.01}$<100mm/h:

$$d_0 = 35e^{-0{,}015*R0{,}01} = 35e^{-0{,}015*19} = 26{,}32$$

Thus, the reduction factor r:

$$r = \frac{1}{1+\frac{d}{d_0}} = \frac{1}{1+\frac{5}{26{,}32}} = 0{,}84$$

Therefore, the length of the active path can now set:

$$d_{eff} = r * d = 0{,}84 * 5 = 4{,}2km$$

So, the total losses for each polarization are calculated separately:

Horizontal polarization

$$L^H_{R0,01} = \gamma^H_{R0,01} * d_{eff} = 2{,}77\left(\frac{dB}{km}\right) * 4{,}2(km) = 11{,}63dB$$

Vertical polarization

$$L^V_{R0,01} = \gamma^V_{R0,01} * d_{eff} = 2{,}29\left(\frac{dB}{km}\right) * 4{,}2(km) = 9{,}61dB$$

- 28GHz

Horizontal polarization

According to Table II, the coefficients K and α at 28GHz take the following values.

With the help of relation (1) the losses per km are:

$$\gamma^H_{R0.01} = K_H * R^a_{0.01} = 0{,}20 * 19^{0{,}96} = 3{,}37\ (\frac{dB}{km})$$

Vertical polarization

Similarly, as before, the losses for vertical polarization are calculated

$$\gamma^V_{R0.01} = K_V * R^a_{0.01} = 0{,}19 * 19^{0{,}92} = 2{,}86(\frac{dB}{km})$$

To calculate the total losses of horizontal and vertical polarization, the losses per km (as was done earlier) and the current deff path's length must be measured. The current deff path calculate by multiplying the factor r by the actual length d.

After first calculating the sufficient path length through rain, for $R_{0.01}$<100mm/h:

$$d_0 = 35e^{-0{,}015*R0{,}01} = 35e^{-0{,}015*19} = 26{,}32$$

Thus the reduction factor r:

$$r = \frac{1}{1+\frac{d}{d_0}} = \frac{1}{1+\frac{5}{26{,}32}} = 0{,}84$$

Therefore, the length of the active path can now set:

$$d_{eff} = r * d = 0{,}84 * 5 = 4{,}2km$$

So the total losses for each polarization are calculated separately:

Horizontal polarization

$$L^H_{R0,01} = \gamma^H_{R0,01} * d_{eff} = 3{,}37\left(\frac{dB}{km}\right) * 4{,}2(km) = 14{,}15dB$$

Vertical polarization

$$L^V_{R0,01} = \gamma^V_{R0,01} * d_{eff} = 2{,}86\left(\frac{dB}{km}\right) * 4{,}2(km) = 12{,}01dB$$

III. Experimental Setup

A. System Architecture

IoT is a device system that interconnects via the Internet. It does with the help of many platforms and protocols on the market. The development of this rainfall detection system (IoT) aims to inform the user about the course of the precipitation, as it is an essential factor for the degree of intensity with which the rain erupts in a certain period (per hour or minute) [18]. The system analyzed consists of a low-cost microcontroller (ESP32) with integrated Wifi and dual-function Bluetooth

(Fig. 3). This connection to all the sensors and many parts of circuits such as (power supplies, analog inputs-outputs, digital outputs, and breadboard). Its primary function is to collect the sensors' data, process them, and inform about its condition through the Adafruit IO service with the internet unit's help. Also used is a SUNKEN rain sensor with dimensions of 13 x 8.4 x 1.3 inches and a Miuzei SG90 9G micro servo motor (Fig. 4).

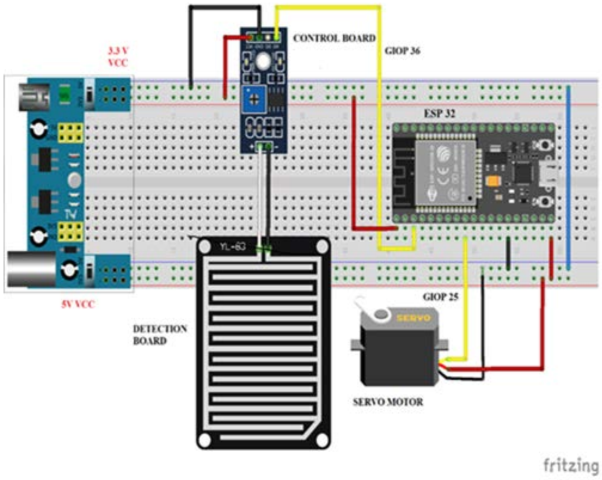


Fig. 3. Simulation of Project.

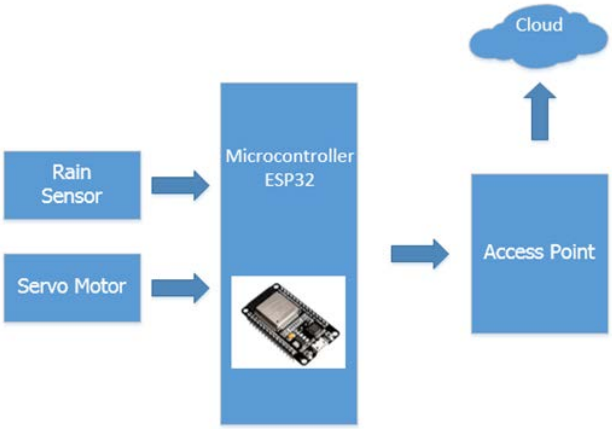


Fig. 4. Block Diagram of Projects.

B. Input System

This system works by receiving data from a rain sensor, where the engine activates after a specific range of values. The output data obtained through ESP 32, where analyzed below:

- Input Data of Rain Sensor

This device uses to control rain or flood in various applications. It included a detection board, a control board where the sensor's sensitivity is regulated—his logic based on a variable resistance. The sensor is a resistor that shows less resistance when wet and more resistant when dry. When there is no water drop on the surface, then the resistance increases, and so at the output, there is a high voltage according to V = IR (Ohm's Law). Correctly, the rain sensor can be used both as a digital input, i.e., distinct values of 0 or 1 can characterize it. (e.g., a switch that can be on (Off / Off) or it can use as an analog input, i.e., it can be characterized as from a price range (e.g., a potentiometer where it can get a price range depending on where it turned). So when there is an analog device, the reading of an analog value with ESP 32 allows us to measure different voltage levels between 0 V and 3.3 V, where this can translate into 0 to 1023 integers. The range is determined by the built-in ADC, which in the case of ESP 32, is ADC 10 bit (2 ^ 10 = 1024). The specifications of rain sensor set out in the Table III.

- Servo motor

Servo motors are small devices that have a shaft protruding on their outer casing. It is a device that can rotate an axis from 0 to 180 or 360 degrees. He could change the angle (move) if a coded signal sents. As long as this signal is present on the Servo input line, it will keep its axis in a specific position. When the signal changes, it causes the servo to change the angle of the shaft. This device is used for many projects, such as controlled aircraft, cars, and robotics. In this case, the engine uses as an analog output.

- Process System

Make clear to the real-time user that the values collected by the microcontroller have set up, a scale showing the total value of the analog value range set by Table IV is displayed, depending on the amount of water falling on the panel. This table has created with the help of a cylindrical tube graded in millimeters [9]. Finally, it should note that the panel place at a slope of 40, which results in the expulsion of a large amount of water more significant than 50 mm.

- Micro controller esp 32

The ESP32-DEVKIT V1 is an MCU Wi-Fi + BT + BLE unit with a wide range of applications, from low-power network sensors to the most demanding tasks such as voice coding, music streaming, and MP3 decoding. Two CPU cores can control individually, and the CPU clock frequency is adjustable from 80 MHz to 240 MHz. The controller has a data rate of up to 150 Mbps and an output power of 20.5 dBm on the antenna to ensure wireless transmission. Its integration with Bluetooth and WIFI makes it flexible for many applications. The controller can connect digital input/output units and analog inputs/outputs, useful for many applications. These units are nothing more than electronic data that, when excited, send a signal to the controller, then it gives the corresponding command with which the user programs it. At this stage, the program code writes in MicroPython language.

TABLE III. SPECIFICATIONS OF THE RAIN SENSOR

Pin	Description
Voltage Rating	5V
GND	Negative power source.
Power supply voltage	3.3-5V
Sensor Board Size	54mm x 40mm
Shield PCB Size	30mm x 16mm
Sensitivity Adjustment	Clockwise is more sensitive.

TABLE IV. RULE BASE SYSTEM

Analog value	Direct reading gauge (mm)
980 - 890	5
890 - 780	10
780 - 670	20
670 - 580	30
580 - 450	40
450 - 400	50

- Software IDE

In this study, use Thonny, where is the Python Integrated Development Environment (IDE) used. It is especially useful for code entry. It has a built-in error detection program. This editor also allows to program the ESP32 and ESP8266 boards with MicroPython and is compatible with Windows, Mac OS X, and Linux. MicroPython came from the Python 3 language and developed to program microcontrollers and electronic devices connected to it. While MicroPython is not yet as well-known as C and C ++ for electronic device scheduling, it is slowly becoming popular with more and more microcontrollers and IDE. It is also effortless to use, and its interface makes it very easy for a novice user. It also features a read-evaluate-print (REPL). This feature helps the user connect to the board and execute it directly without compiling it. In the present phase, the interpreter, MicroPython (ESP32), and the corresponding USB prey that the board has connected to select the code to load on it select.

- Adafriut IO

Adafruit.IO is a cloud service - this means that it does not need to be managed by the user. This platform uses to connect to the Internet. To store and retrieve data, but can do much more than that. Adafruit.Io can handle and visualize many data streams. The platform environment is quite user-friendly (Fig. 5). Also, the Adafruit online platform works with many microcontrollers (including ESP32) and, with the support of the MQTT protocol, allows easy and secure management of IoT systems.

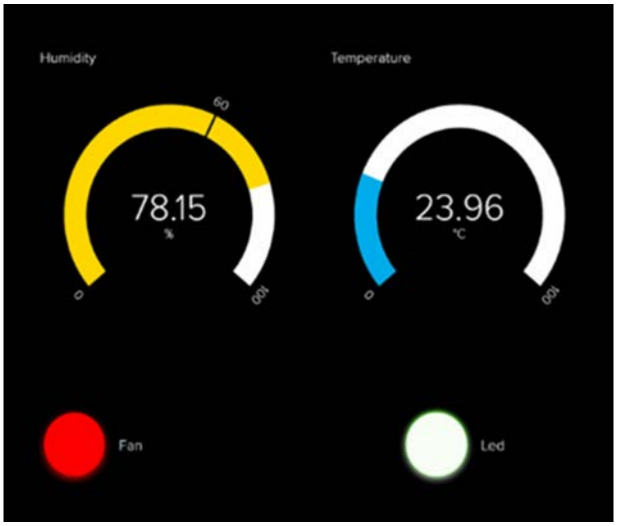


Fig. 5. Interface Adafruit IO.

C. Working of IoT based Rainfall Monitoring System

The construction shows that the rain height's size can place in a base station, a residential complex, lighting columns, and different environments (urban or rural areas) (Fig. 6). The basic idea is to inform the user in real-time via a local network from the platform. An ESP32 device can play the role of an Access Point or a Station. The data output thus passes to the Adafruit IO platform via the MQTT protocol.

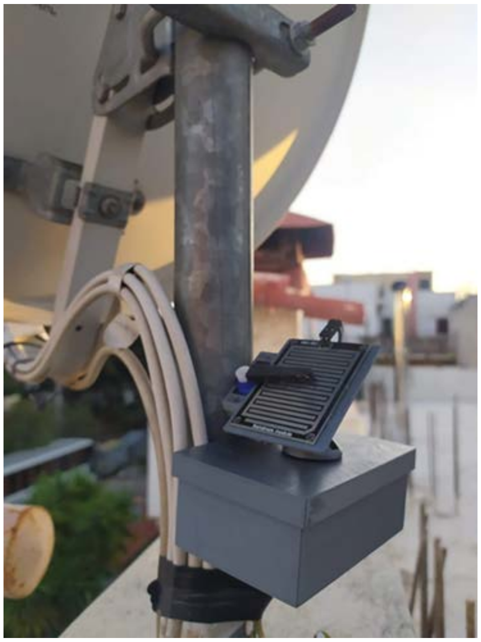


Fig. 6. IoT based Rainfall Monitoring System with 3D Case.

The sender is the (publisher), and the recipient is the (subscriber). Both do not come into direct contact with each other. However, a third party (MQTT broker) directs the publisher's messages to any endpoint, that is, the subscriber. The following Fig. 7 shows the IoT system architecture, where the sensor is located on the transmitter and through the MQTT broker sends the appropriate messages to the monitor that controls them in real-time.

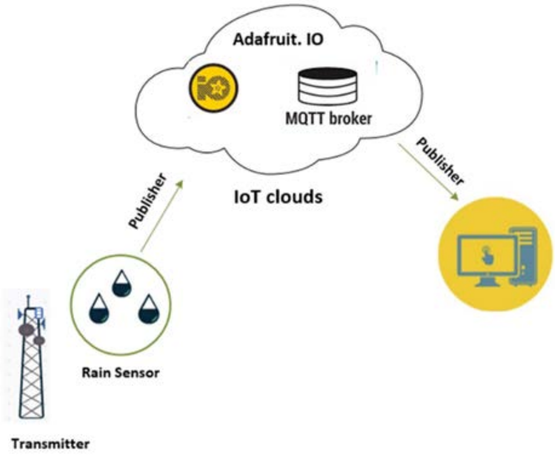


Fig. 7. MQTT Broker Network Architecture in Wireless Broadband.

The board has programmed so that the dry surface indicates the value of 1023. The values that can be displayed come from many experimental tests with a cylindrical tube graded in millimeters. There are no other values from one point onward because the rain panel is placed at a slope of 40°, resulting in not enough water to remain when it falls on it due to gravity.

IV. Results and Discussion

A. *Theoretical Measurement Results*

Theoretical weather calculations on wireless communication systems would be more straightforward if there were no rain or humidity effects. As it is known from the bibliography so far, the rain can weaken the radio waves' power depending on its rhythm and operation frequency. The present study presents the scenarios presented at Palo Alto-based on three different frequencies of 24 GHz, 28 GHz, and 38 GHz, where they are ideal for 5G wireless broadband connections.

From the theoretical measurements made, the rainfall attenuation calculates with a coupling availability of 99.9%, with a distance of 5km for each of the three operating frequencies (Fig. 8). It found that the lower the operating frequency, the lower the radio signal's attenuation, and that different loss values are presented based on polarization. Horizontal polarization is more prone to damping than vertical polarization. As raindrops increase in size, they get more extended in the horizontal direction. They, therefore, will attenuate horizontal polarization more than vertical polarization. The shape of the drops explains this as they fall from the atmosphere. In particular, during the fall, they change shape showing the tendency to distribute their volume, mainly along the horizontally polarized electric field. In this regard, the field interaction with the droplet is more critical in horizontal polarization than in vertical. Therefore, the scenario where it is less affected by rainfall is the operation at 24 GHz. The frequency range from 24 GHz and above offers low latency and high data rate [19]. Such a wireless connection will continue to provide high-quality services to users, even in severe weather conditions.

The specific attenuation depends on the value of the frequency. As the frequency increases, so do the specific attenuation based on the international recommendations (ITU-R P.838-3) [12].

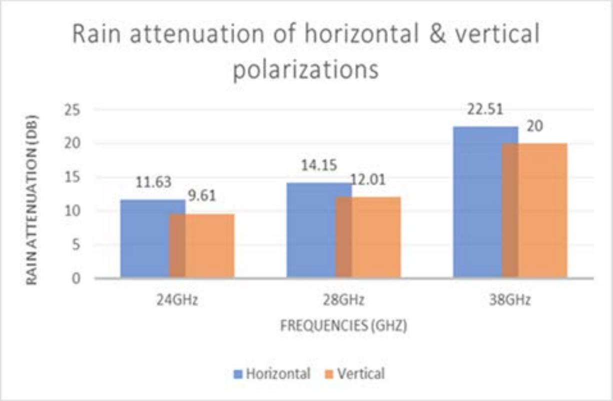
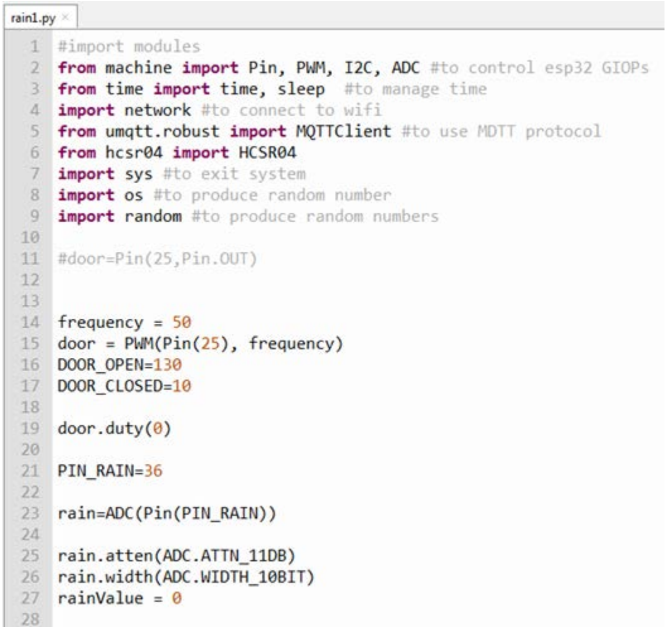


Fig. 8. Rain Attenuation of Horizontal and Vertical Polarizations.

B. *Esp 32 Thonny IDE Output*

As studied in the scenario that took place, the parameter of the rate of the rainfall R (mm / hr) is critical because the original size that characterizes it is the height of the rain, where it measures in millimeters. In the theoretical measurements that followed, the rate was the same due to its geographical location. At present paper, an IoT system develops, which will present in real-time the rainfall in the standards of a scale that create for the display of the sizes. The planning divided into three sections:

In the first one, the required libraries have defined, the calls of the Pin, PWM, I2C, ADC classes, and the initializations of the analog input for the rain sensor and the initializations for the analog output, which is the servo motor (Fig. 9).



```
 1  #import modules
 2  from machine import Pin, PWM, I2C, ADC #to control esp32 GIOPs
 3  from time import time, sleep  #to manage time
 4  import network #to connect to wifi
 5  from umqtt.robust import MQTTClient #to use MDTT protocol
 6  from hcsr04 import HCSR04
 7  import sys #to exit system
 8  import os #to produce random number
 9  import random #to produce random numbers
10
11  #door=Pin(25,Pin.OUT)
12
13
14  frequency = 50
15  door = PWM(Pin(25), frequency)
16  DOOR_OPEN=130
17  DOOR_CLOSED=10
18
19  door.duty(0)
20
21  PIN_RAIN=36
22
23  rain=ADC(Pin(PIN_RAIN))
24
25  rain.atten(ADC.ATTN_11DB)
26  rain.width(ADC.WIDTH_10BIT)
27  rainValue = 0
28
```

Fig. 9. Libraries and Initialization of Variables.

In the second part of the code is defined as the communication MQTT client publisher, with the MQTT broker (server), where this, in turn, communicates with client subscribers. A Client can be both a publisher and a subscriber. The sensor's measurement is in line 119, and the value of the measurement is sent with the Send_Data () function on line 122. Essentially the line of code that sends the measurement to the interface is line 86. The above line sends the information str (rainValue) through the client created above (line 48), “mqtt_feedname1”. Specifically, the path, “io.adafruit.com/kostaszar/feeds/rain”. Rain is the name of the feed on the interface side, i.e., the value rainValue will send to the interface block connected to the feed rain.

The third and last show all those commands that are the program's central structure and are the ones that have the central role. It is essentially a repetition where it is checked every 10 seconds for the values it takeç. If the values listed in the platform interface are below 800, then the most significant amount of water is collected in the panel, so there is a need to clean where after two seconds, the rain sensor is ready to receive water again for the next recording. Within these two seconds, the engine has made two 180° turns. With a drop of 5 mm, the water drop, the program settings so that the range of values it can get is between 980 and 890. Then the 10 mm of water corresponds to 890 - 780 and the 20 mm of water in 780 - 670. The next Price range is 670 - 580 were with 30 mm of water. Finally, with 40 mm of water space, precipitation is 580 - 450, and with 50 mm of water, a value appears between 450 – 400.

Every element presents a size or an output to a feed. Each feed is a communication channel between ESP 32 (client) and Adafruit (broker). Dashboards can place the icons (blocks) by the developer. In this analysis, a rain gauge defines, from which the value displayed appears at a specific time, and the user can easily see it.

V. Conclusion

As studied for the 24 GHz, 28 GHz, and 38 GHz frequency bands, damping due to rain may seem vital because they create network coverage problems. Especially at 38 GHz, the damping for horizontal polarization is 22.51 dB. At 28 GHz, the damping is less by 8.36 dB, while at 24 GHz, the losses are smaller than the other two frequencies where they reach 11.63 dB for horizontal polarization.

In the era of the fifth generation of wireless broadband networks in urban areas due to the increased coverage and capacity requirements, the mobile communication networks distributed in cells. Where within a macro cell, there may be many smaller ones to serve more broadband services. The phenomenon of attenuation will be more evident in rural areas because, in addition to the prevailing weather conditions, the distance that the electromagnetic wave must travel as it travels taken into account. In contrast to urban areas where the distances are short, it seems that the 24 GHz frequency is ideal from the other two, although they are suitable for the development of fifth-generation networks.

Thus, mobile companies' ultimate solution will be to continually update the weather conditions prevailing in a base station or a wireless broadband backhaul installation—ideal for the places where the climate is changing rapidly and is affected by heavy rainfall. According to this document, sensor systems will be useful because they can interconnect with the World Wide Web and other devices that will allow them to exchange data and provide the required information.

References

[1] Chen, C., Harvey, J.A., Biere, A., & Gols, R. (2019). Rain downpours affect survival and development of insect herbivores: the specter of climate change? Ecology, 100.

[2] Yin, J., Gentine, P., Zhou, S., Sullivan, S.C., Wang, R., Zhang, Y., & Guo, S. (2018). Large increase in global storm runoff extremes driven by climate and anthropogenic changes. Nature Communications.

[3] Trenberth, K.E. (2011). Changes in precipitation with climate change.

[4] Easterling, D.R., K.E. Kunkel, J.R. Arnold, T. Knutson, A.N. LeGrande, L.R. Leung, R.S. Vose, D.E. Waliser, and M.F. Wehner, 2017: Precipitation change in the United States. In: Climate Science Special Report: Fourth National Climate Assessment, Volume I [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 207-230, doi: 10.7930/J0H993CC.

[5] Singh, H., Bonev, B., Petkov, P.Z., & Kumar, R.R. (2018). A Novel Approach for Predicting Attenuation of Radio Waves caused by Rain. 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 1193-1198.

[6] M. Ghanim, M. Alhilali, J. Din and H. Y. Lam, "Rain Attenuation Statistics over 5G Millimetre Wave Links in Malaysia," 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Malang, Indonesia, 2018, pp. 266-269.

[7] Shayea, I., Rahman, T.A., Azmi, M.H., & Islam, M.R. (2018). Real Measurement Study for Rain Rate and Rain Attenuation Conducted Over 26 GHz Microwave 5G Link System in Malaysia. IEEE Access, 6, 19044-19064.

[8] M. I. Haque, A. H. MD. Shatil, A. N. Tusar, M. Hossain and M. H. Rahman, "Renewable Powered Portable Weather Update Station," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2019, pp. 374-377.

[9] O.Omoruyi, S. N. John, O. Chinonso, O. Robert, A. A. Adewale and K. O.Okokpujie, "Wireless Sensor Network for Rainfall Measurement Using a Tipping Bucket Rain Gauge Mechanism," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2017, pp. 740-744.

[10] O.Manzombi, E. M. Dogo and N. I. Nwulu, "Design and Implementation of a Wireless Patient Health Monitoring System," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2019, pp. 1-6.

[11] Kestwal, M.C., Joshi, S., & Garia, L.S. (2014). Prediction of Rain Attenuation and Impact of Rain in Wave Propagation at Microwave Frequency for Tropical Region (Uttarakhand, India).

[12] Recommendation ITU-R P.838-3 (03/2005), “Specific attenuation model for rain for use in prediction methods”, P Series, Radiowave propagation.

[13] Recommendation ITU-R PN.837-1 (08/1994), “Characteristics of precipitation for propagation modelling”, P Series, Radiowave propagation.

[14] Recommendation ITU-R P.530-9 (02/2001), “Propagation data and prediction methods required for the design of terrestrial line-of-sight systems, P Series, Radiowave propagation.

[15] Pathak, T.B.; Maskey, M.L.; Dahlberg, J.A.; Kearns, F.; Bali, K.M.; Zaccaria, D. Climate Change Trends and Impacts on California Agriculture: A Detailed Review. Agronomy 2018, 8, 25.

[16] Rappaport, T. S., Murdock, J. N., & Gutierrez, F. (2011). State of the art in 60-GHz integrated circuits and systems for wireless communications. Proceedings of the IEEE, 99(8), 1390-1436. [5958173]. https://doi.org/10.1109/JPROC.2011.2143650.

[17] Ahmadi, Sassan. 5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards. N.p.: Academic, 2019. Print.

[18] Brouwer, C., Goffeau, A. and Heibloem, M. (1985) Irrigation Water Management: Training Manual No. 1-Introduction to Irrigation. FAO, Rome.

[19] Ancāns, Guntis & Bobrovs, Vjaceslavs & Ancans, Arnis & Kalibatiene, Diana. (2017). Spectrum Considerations for 5G Mobile Communication Systems. Procedia Computer Science. 104. 509-516. 10.1016/j.procs.2017.01.166.

Development of Wearable Heart Sound Collection Device

Ximing HUAI[1], Shota Notsu[2], Panote Siriaraya[4]
Noriaki Kuwahara[5]
Graduate School of Science and Technology
Kyoto Institute of Technology, Kyoto, Japan

Dongeun Choi[3]
Faculty of Informatics
The University of Fukuchiyama
Kyoto, Japan

Abstract—**In recent years, the mortality rate of cardiovascular diseases and the younger generation have attracted people's attention. At the same time, there is an increasing demand for devices that can monitor the physiological parameters of the heart. In this research, a wearable devices was designed and developed for heart sound collection. Microphones wrapped in urethane resin holders were directly fixed on the vest for heart sound collection. The device has received many positive reviews in terms of comfort. The cumulative contribution rate of the two common factors (material factor and clothing design factor) obtained through factor analysis was 75.371%, which was the main factor affecting the experience of using the device. Finally, the heart sounds of 11 healthy young people were collected and input into the completed convolutional neural network for detection, and an accuracy rate of 71.3% was obtained. Therefore, it can be concluded that the device improves the user experience and has a good effect on heart sound collection and detection.**

Keywords—Cardiovascular diseases; wearable devices; heart sound collection; convolutional neural network

I. Introduction

The number of people dying from cardiovascular disease (CVD) is steadily rising, including one-third of all deaths globally in 2019, according to a paper in the Journal of the American College of Cardiology that reviewed the total magnitude of CVD burden and trends over 30 years around the world [1]. With the current aging of the world, the mortality rate of cardiovascular diseases will continue to increase. At the same time, cardiovascular disease is no longer a patent for the elderly. Among the sick people, the ratio of middle-aged and young people is also rising, and cardiovascular diseases are getting younger. The main reason for this trend is the bad lifestyle habits of young people, such as long-term smoking, lack of sleep, unhealthy eating habits, lack of exercise, etc. [2-3].

As the rejuvenation of cardiovascular disease becomes more and more obvious, it has attracted widespread attention, and people's health awareness is gradually improving. At the same time, it means that the demand for medical equipment is constantly expanding, not only for medical equipment in hospitals, but also for equipment that can monitor human physiological parameters in daily life. This kind of equipment does not require the patient to stay in the hospital all the time, but it can monitor various physiological parameters of the human body in real time, reflect the health of the human body, and provide accurate basis for clinical diagnosis [4-7] for doctors to make timely diagnosis and treatment, to improve the efficiency of the hospital's diagnosis of the patient's condition. For elderly people and young people with great work pressure, getting 24 hours uninterrupted supervision in their daily life, can make them gain a sense of security to a large extent. The remote monitoring of physiological parameters has become a research hotspot in the field of biomedical engineering [8-9].

At present the type of physiological parameter monitoring equipment range on the market, divided by monitoring the way, there are wearable, portable, desktop detection equipment etc.; divided by function, there are heart sound and electrocardiogram (ECG) monitoring, defibrillation monitoring, sleep monitoring equipment etc. [10-11]. With the continuous advancement of technology and the increase of market demand, wearable monitoring equipment has become the future research direction. The most representative equipment at present is the wearable vest-type physiological parameter monitoring equipment Life Shirt developed by Vivo Metrics company in the United States. As shown in Fig. 1. The device can monitor different physiological parameters of the human body through multiple biosensors, such as ECG, blood pressure, heart rhythm, body temperature, etc. [12-14]. It has been used as a clinical medical instrument by many hospitals, but the data collected by the device cannot be transmitted in real time, so remote real-time monitoring cannot be realized.

In view of the high mortality rate of cardiovascular diseases, monitoring equipment specifically for heart sounds and ECG is also emerging in endlessly. Since heart sound auscultation is one of the most basic diagnostic methods for cardiovascular diseases, it is an important basis for the initial diagnosis [15-16], so the monitoring of heart sounds can reflect the human heart condition in a timely manner. This helps doctors find abnormalities in the heart as soon as possible so as to improve the diagnosis efficiency of the disease and enable patients to receive timely treatment. In the current research on heart sound monitoring equipment, it can be roughly divided into two categories: electronic stethoscope collecting heart sounds and microphone collecting heart sounds. Among them, the development of electronic stethoscope collection of heart sounds is relatively mature. For example, Beck C et al. [17] designed and developed a multi-mode physiological parameter collection wearable device including heart sound auscultation, which can be used with MATLAB software to obtain real-time data or store data in software; Tiwari, Hemant Kumar et al. [18] have designed an embedded stethoscope served as a platform for the computer aided diagnosis of cardiac sound for the detection of cardiac murmur, the device can display heart sounds on the TFT LCD display in real time, and stored on the micro SD card. However, these devices are inconvenient for daily carrying due to their large size. In terms of microphone collection of heart sounds, Kirchner J et al. [19] proposed a wearable system for long-term capture of chest sound, including four microphones for receiving physiological signals and one microphone for receiving ambient noise. The performance of the device was evaluated in different environments, and the use of combined microphones improved the noise cancellation; Aguilera-Astudillo C et al. [20] demonstrate a chest piece which consists of an electret microphone embedded into the drum of a 3D printed chest piece. And use the electronic dongle to amplify the microphone signal and reduce external noise, while the mobile phone can display the heart sound signal. From the above studies, we can find that the use of electronic stethoscopes to collect heart sounds is not convenient for daily wear due to the size of electronic stethoscopes, and the use of microphones to collect heart sounds did not investigate the comfort of the device and the wearing experience, so we believe that there is still more room for research in the wearing comfort of heart sound monitoring devices.

Fig. 1. Life Shirt Wearable Physiological Parameter Monitoring System.

In view of previous research, this paper is devoted to the design and development of an easy-to-wear heart sound collection device, to discuss and study the comfort of the device, and to analyze the collected heart sounds using convolutional neural networks (CNN). The organization structure of this article is as follows: Section II explains the design of the heart sound collection vest in detail; Section III introduces the heart sound collection experiment and heart sound data analysis; Section IV analyzes the results and discusses them; Section V draws conclusions and Section VI summarizes the shortcomings and makes suggestions for future work.

II. Design of the Heart Sound Collection Vest

This part introduces the design and development of the heart sound collection vest. First, a 3D printing mold was designed and printed according to the size of the microphone. After that, urethane resin was used to make a super soft microphone holder. Finally, according to the 4 positions of the heart sound auscultation, the microphone was fixed on the vest to make a wearable heart sound collection vest.

A. The Design of the Microphone Holder

1) 3D printing mold design: The microphone used in this design is the Uni-directional Electret Condenser Microphone produced by Primo Company, the model is EM297, as shown in Fig. 2, the diameter is 10mm, and the thickness is 4.5mm. According to the instructions of the microphone, the sound hole on the back of the microphone cannot be blocked in order to receive the sound. At the same time, in order to ensure that the microphone can accurately receive the sound, the microphone needs to be close to the skin, which requires the microphone to withstand a certain amount of pressure, so it is necessary to make a suitable holder to ensure that the microphone can receive normally and accurately.

According to [21], when the diameter of the holder is 17mm, and there is a 1.5mm thick silicone layer between the microphone and the skin, the microphone has the best sound reception. Therefore, according to the shape and size of the microphone, the 3D printing mold of the holder has shown in Fig. 3 was designed. The mold is a cylinder and adopts a fully-wrapped type except for the sound hole on the back. At the same time, a dedicated space is reserved for the microphone wire.

2) Development of soft skin-friendly holder: In view of the fact that the microphone holder needs to be in direct contact with the skin, the urethane resin with a hardness of 0 produced by EXSEAL company was selected to make the holder in consideration of skin affinity when selecting the material. This urethane resin has excellent softness, sufficient molding strength even if the hardness is 0, and this urethane resin has a softness like a human skin.

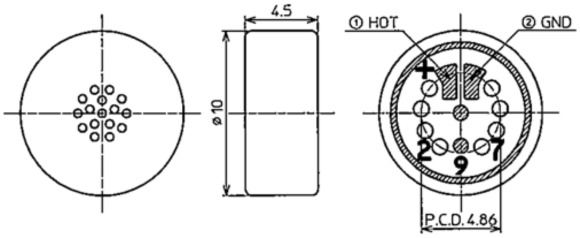


Fig. 2. Schematic Diagram of Microphone.

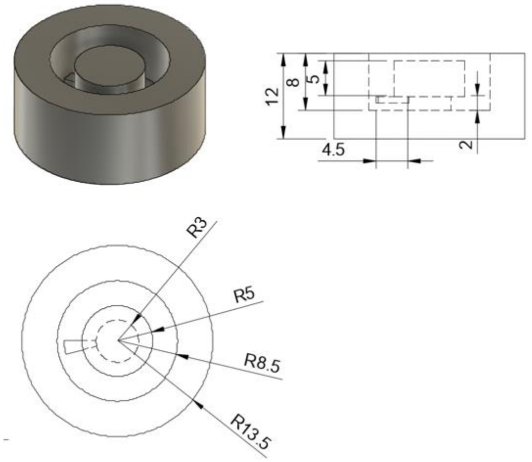


Fig. 3. Schematic Diagram of the 3D Printing Model of the Microphone Holder (unit: mm) (a) Holder Molding Model (b) Model Front view (c) Model Top View.

According to the product's instructions, first calculate the volume of the mold before preparing the gel, and spray the mold with a special release agent for urethane resin in advance to facilitate the release of the holder afterwards; then mix the two liquids (the weight ratio of main agent and curing agent is 3:1) and pour them into the 3D printing mold; finally, after 24 hours, take out the molded holder from the mold. Fig. 4 shows the process of manufacturing the microphone holder. The holder wraps the microphone well and protects the relatively fragile and breakable wires of the microphone. Fig. 5 is a schematic diagram after the microphone is installed in the holder.

B. The Design of the Heart Sound Collection Vest

1) Location of heart sound auscultation: According to [22], there are 4 commonly used positions for heart sound auscultation. As shown in Fig. 6, the mitral valve is located at the strongest point of the apical beat, and is normally located at the fifth intercostal space on the inner side of the left midclavicular line; the pulmonary valve is in the second intercostal space on the left edge of the sternum; the aortic valve is in the second intercostal space on the right edge of the sternum; the tricuspid valve is on the left edge of the lower end of the sternum, that is, the 4-5th intercostal space on the left edge of the sternum. The general auscultation sequence is the beginning of the mitral valve - the pulmonary valve - the aortic valve - the tricuspid valve.

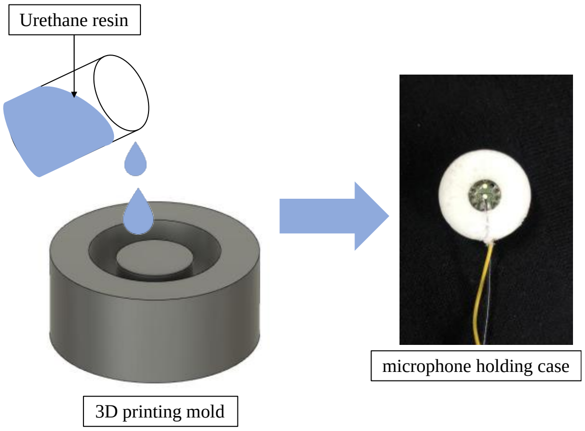


Fig. 4. Schematic Diagram of Making Microphone Holder.

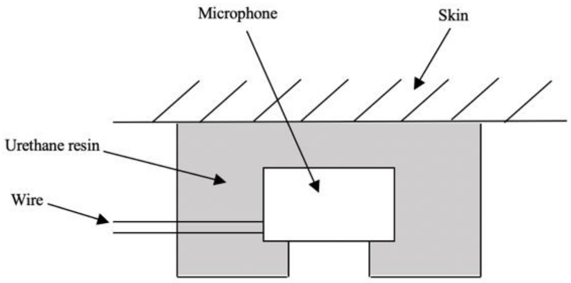


Fig. 5. Schematic Diagram of the Microphone Placed in the Holder.

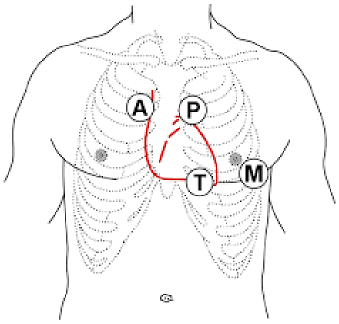
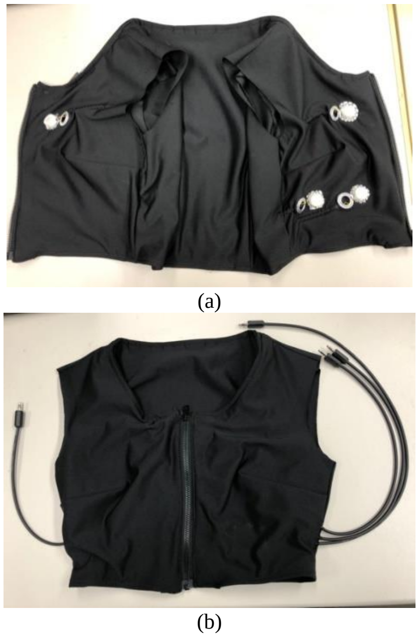


Fig. 6. Position of Heart Sound Auscultation. A = Aortic Valve; P = Pulmonic Valve; M = Mitral Valve; T = Tricuspid Valve.

2) Making of heart sound collection vest: A wearable vest for heart sound acquisition was proposed as the main object of this study based on the location of heart sound auscultation and taking into account the comfort, fit and appearance of the device.

A fabric made of Polyester 90% and Lycra 10% was used as the base fabric, and a suitable vest was made according to the m size of men's and women's clothing size. The vest was designed with two layers of fabric, which could facilitate the fixation of the microphone and make the vest neat in appearance. After that, according to the four positions of heart sound auscultation, the fixing sleeve with microphone is fixed on the corresponding position on the inside of the vest: firstly, the position of the microphone is determined, then it is wrapped with cloth and sewn on the vest, then holes are made at the suitable position nearby for the wire to pass through, and finally the wire is fixed on the vest with needle and thread and passed out from the side of the vest. As shown in Fig. 7, (a) is a diagram of the microphone device inside the vest, and (b) is a diagram of the exterior of the vest.



(a)

(b)

Fig. 7. Physical Diagram of the Heart Sound Acquisition Vest (a) Microphone Device on the Inside of the Vest (b) Diagram of the Outside of the Vest.

III. Experiments and Methods

This section describes the experiments used to evaluate the acquisition effectiveness of the designed heart sound collection vest and the comfort of wearing it.

A. *Heart Sound Collection Experiment*

A heart sound collection experiment was conducted on healthy young people to evaluate the collection effect and wearing comfort of the heart sound collection vest. 11 students (5 females and 6 males) participated in the experiment. Ages ranged from 21 to 29. Height, weight and chest circumference of the participants were measured before the experiment: height ranged from 1.61-1.7m, weight ranged from 54-56 kg and chest circumference ranged from 80-93 cm for females, and height ranged from 1.58-1.75m, weight ranged from 55-68 kg and chest circumference ranged from 81-98 cm for males.

The heart sound collection experiment was carried out in an indoor experimental environment with a temperature of 20 degrees Celsius, a humidity of 45%, and a surrounding environment with a decibel number of 15dB. The participants wore a vest of appropriate size, and the heart sounds were collected and stored in the following order of auscultation: mitral valve area - pulmonary valve area - aortic valve area - tricuspid valve area, with 3-5 heart sounds randomly collected for 30-180 seconds at each location, as shown in Fig. 8..

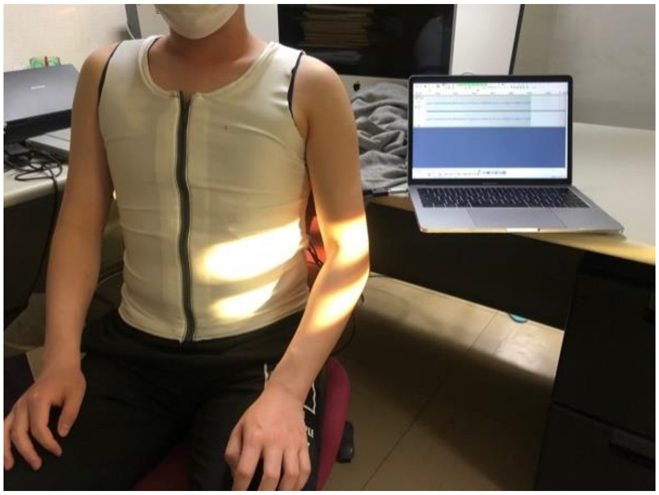


Fig. 8. Experimenter in Heart Sound Collection Experiment.

B. *Investigation of Device Comfort*

After the completion of the heart sound collection experiment, the experimental participants were asked to complete an assessment questionnaire for the heart sound collection vest. The questionnaire was set up with reference to the Semantic Difference [23], which is a measure of semantic differentiation. It was developed by the social psychologist Osgood et al. in the 1950s. Such scales consist of a series of bipolar adjective word pairs that are generally divided into seven equal rating scales. They have the quality of showing the semantic space in which any concept has meaning, and can be used accordingly to describe the underlying meaning of any concept and its associated problematic nature or property aspects. In the present study for the wearable device, considering the user-friendly characteristics that the device needs to have, seven contrasting adjectives and a 7-point rating scale from -3 to +3 were proposed in terms of material, prolonged wear, and appearance for the experimental participants to rate this heart sound collection vest. The seven adjectives were 1) tight/loose, 2). Strong/weak, 3) heavy/light, 4) hard/soft, 5) unattractive/nice, 6) inconvenient/convenient, 7) dislike/like. Each adjective group corresponds to an explanatory question: 1) Do you think the pressure exerted by this device on the skin is tight or loose? 2) Do you think the stimulation of this device on the skin is strong or weak? 3) Do you think this device is heavy? 4) Do you think the touch of this device is hard or soft? 5) Do you think the design of this device is good-looking? 6) Do you think this device is convenient to put on and take off? 7) Do you like to use this device for a long time?

IV. Results and Discussion

A. *Heart Sound Data Denoising*

There are many noises generated during heart sound collection, for example: ambient noise; vibration of the microphone caused by chest vibration and the zipper of the vest is located in the middle of the chest causing the microphone in the pulmonary valve area and aortic valve area to not fit well to the skin, generating a lot of noise, etc. Therefore, filtering and denoising are needed before heart sound data detection. In this study, an algorithm written in python is used to remove the noise. The steps of the algorithm are: 1) calculate Fast Fourier Transform (FFT) on the noisy audio fragment, 2) calculate statistics from the FFT of the noise, 3) calculate a threshold based on the statistics of the noise (and the desired sensitivity of the algorithm), 4) calculate FFT from the signal, 5) determine the mask by comparing the signal FFT with the threshold, 6) the mask is smoothed in frequency and time using a filter, 7) the mask is applied to the FFT of the signal and is inverted. After processing the heart tone data using this algorithm, the spectrograms of the heart tone data before and after denoising are obtained as shown in Fig. 9. The upper panel is before denoising and the lower panel is after denoising, and it can be clearly seen that most of the noise is removed.

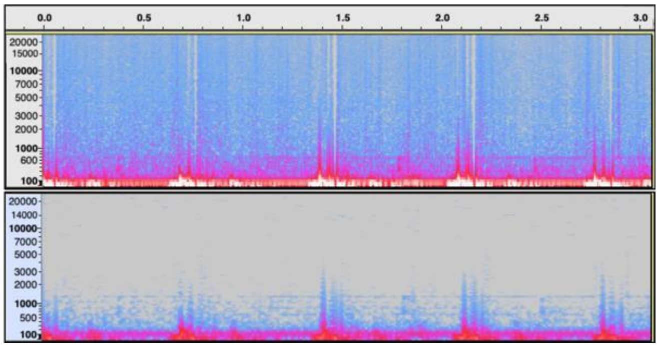


Fig. 9. Spectrograms of Heart Sounds before and after Denoising.

B. *Heart Sound Data Detection Results*

To test the denoising effect, the heart sound data before and after denoising were fed into the convolutional neural network (CNN) [24] that had been built to obtain the accuracy of heart sound data detection. Firstly, the heart sound data were divided in 5-second intervals, and then the sample heart sounds were converted into grayscale spectrograms by an automatic procedure using the librosa library, and finally fed into the CNN. This CNN consists of 5 layers, including 3 convolutional layers, 1 fully connected layer and 1 normalized exponential function softmax classification layer. The network parameters for each layer are as follows: for the first convolutional layer, we used 32 filters with a convolutional kernel size of 3 × 3 and a step size of 1 × 1, and the pooling (using the max-pooling method) size was 2 × 2 and a step size of 1 × 1. We used a modified linear unit (ReLU) activation function with a random deactivation (loss) rate of 0.1. For the second and third convolutional layers, we used 64, 128 filters. The convolution kernel size is 3 × 3 and the step size is 1 × 1. The pooling operation, activation function and dropout probability were the same as in the previous layer. For the fourth layer, we use a fully connected dense layer with an output size of 500, a ReLU activation function and a dropout function (probability 0.25). In the last classification layer, we used a Softmax classifier.

The detection of the convolutional neural network shows that the accuracy of the un-denoised heart sound data is 34.89% and the accuracy of the denoised heart sound data is 71.30%. In order to investigate the detection accuracy of the heart sound collection vest under the wearing of different gender groups, the detection results of the heart sound data were analyzed according to male and female as well as each person, as shown in Tables I, II, and III.

From Table I, it can be seen that although the accuracy of heart sound data for males before denoising was lower than that of heart sound data for females, after denoising, the accuracy of heart sound data for both males and females exceeded 70%, which indicates that the denoising procedure was effective and was able to remove most of the noise and improve the accuracy of heart sound detection.

Table II shows the comparison of the accuracy of heart sound data before and after denoising for males, from which it can be found that the accuracy of the heart sound data obtained in the state of collecting heart sounds while wearing the same vest varies due to the different height, weight and chest circumference of each experimental participant. The reason for this is that the tighter the vest fits to the skin, the clearer the heart sound data is and the higher the accuracy rate is after denoising. Therefore, the better the fit of the vest to the skin, the better the collection results.

Table III shows the comparison of the accuracy of heart sound data before and after denoising in women, from which it can be found that the heart sound data of experimental participants No. 1 and No. 2 obtained more than 91% accuracy after denoising, and the other three also obtained more than 64% accuracy, indicating that the heart sound acquisition vest can acquire heart sounds well. However, the detection accuracy of individual heart sound data varied widely due to the difference in body size of each individual and the fact that the vest for women was designed without the inclusion of a bra part.

C. Factor Analysis of Heart Sound Collection Vest

Fig. 10 visualizes the semantic difference (SD) evaluation scores of men and women on wearable heart sound collection devices. It can be seen from the figure that men and women have roughly the same evaluation of the device, but women have given more positive feedback. Among them, both men and women think that the device has weak skin irritation, not heavy, soft, convenient to wear, and suitable for long-term use. These positive feedbacks may be attributed to the following factors: 1) Polyester and lycra used in the fabric of the vest are soft and stretchable, suitable for personal wear; 2) The microphone holder is made of urethane resin with a hardness of 0, it has a softness like a human skin; 3) The vest is designed with a zipper to facilitate putting on and taking off. At the same time, women think that the device has little pressure on the skin and the appearance is attractive, but men think that the device has a certain amount of pressure on the skin and the appearance is not very attractive. This is because there are certain differences in the design of the male and female vests, and it is also the direction of improvement in the future.

TABLE I. COMPARISON OF THE ACCURACY OF HEART SOUND DATA DETECTION BEFORE AND AFTER DENOISING

	Accuracy of data without denoising	Accuracy of data after denoising
Male	27.36%	70.01%
Female	40.16%	72.20%

TABLE II. COMPARISON OF THE ACCURACY OF HEART SOUND DATA DETECTION BEFORE AND AFTER DENOISING FOR MALES

	Height (cm)	Weight (kg)	Chest circumference (cm)	Accuracy of data without denoising	Accuracy of denoised data
M-1	168	68	95	42.54%	85.39%
M-2	173	67.5	98	43.51%	83.76%
M-3	175	65	87	11.28%	72.16%
M-4	175	60	89	12.96%	70.15%
M-5	167	55	81	12.16%	55.32%
M-6	158	58	86	41.35%	54.31%

TABLE III. COMPARISON OF THE ACCURACY OF HEART SOUND DATA DETECTION BEFORE AND AFTER DENOISING FOR FEMALES

	Height (m)	Weight (kg)	Chest circumference (cm)	Accuracy of data without denoising	Accuracy of denoised data
F -1	162	55.6	90	24.20%	91.59%
F-2	170	55	80	44.95%	91.57%
F-3	163	55	92.5	31.18%	68.25%
F -4	161	54	85	51.95%	66.67%
F -5	169	56	82	64.73%	64.87%

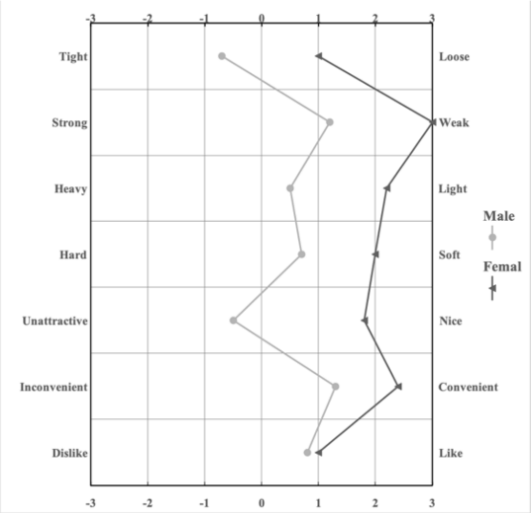


Fig. 10. Image Scale of SD Evaluation Score (Average Value) of Male and Female Wearable Heart Sound Acquisition Devices.

Factor analysis was conducted on the SD scores of 11 experimental participants on the wearable heart sound collection vest [25]. Factor analysis is the conversion of multiple measured variables into a small number of composite indicators (or latent variables), and it reflects an idea of dimensionality reduction. The variables with high correlation are clustered together through dimensionality reduction, thus reducing the number of variables that need to be analyzed while reducing the complexity of the problem analysis. Thus, the use of factor analysis provides insight into the potential factors affecting wearable heart sound collection devices.

Factor analysis was conducted using python. First, to confirm the suitability of the data for factor analysis, Kaiser-Meyer-Olkin (KMO) and Bartlett tests were performed with: KMO test is used to examine the bias correlation between variables, taking values between 0 and 1; the closer the KMO statistic is to 1, the stronger the bias correlation between variables and the better the factor analysis. Generally the statistic is above 0. 6 is adapted to do factor analysis [26]; and the smaller the value of P-Value indicates the more significant data differences. The results obtained through the test are shown in Table IV, the KMO of this experimental data = 0.617> 0.6,P-Value = 0.001, so it is suitable for factor analysis.

Then, the scree plot was drawn according to the number of factors and the corresponding eigenvalues of each factor, as shown in Fig. 11; Table V shows the total variance explained for each component of the factor analysis of the heart sound collection vest. From Fig. 11 and Table V, it can be obtained that the eigenvalues of the first 2 factors are greater than 1, while the variance contribution of these 2 common factors accounts for 75.371%, indicating that the extraction of these 2 common factors can explain the majority of the information of the original data. From Fig. 11 and Table V, it can be obtained that the eigenvalues of the first 2 factors are greater than 1, while the variance contribution of these 2 common factors accounts for 75.371%, indicating that the extraction of these 2 common factors can explain the majority of the information of the original data. Therefore, the number of common factors was determined as 2 to perform the required factor extraction.

TABLE IV. THE KMO AND BARLETT'S TEST OF THE FACTOR ANALYSIS OF THE WEARABLE DEVICES

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.617
Bartlett's Test of Sphericity	Approx. Chi-Square	41.853
	Sig.(P-Value)	0.001

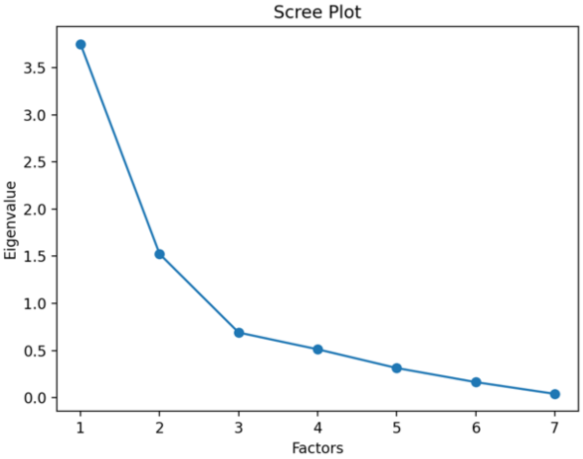


Fig. 11. Scree Plot of the Factor Analysis of the Heart Sound Collection Vest Evaluation.

TABLE V. TOTAL VARIANCE EXPLAINED OF EACH COMPONENT OF THE FACTOR ANALYSIS OF THE HEART SOUND COLLECTION VEST

	Initial Eigenvalues			**Rotation Sums of Square Loadings**		
Component	***Total***	***of Variance %***	***Cumulative %***	***Total***	***of Variance %***	***Cumulative %***
1	3,751	53.578	53.578	3.499	49.983	49.983
2	1.525	21.783	75.361	1.777	25.388	75.371
3	0.691	9.870	85.231			
4	0.513	7.328	92.559			
5	0.315	4.499	97.058			
6	0.165	2.357	99.415			
7	0.041	0.586	100.000			

TABLE VI. THE ROTATED COMPONENT MATRIX OF THE FACTOR ANALYSIS

	Factor1	Factor2
Heavy - Light	0.932	0.151
Unattractive - Nice	0.888	0.418
Strong - Week	0.811	0.011
Tight - Loose	0.795	0.271
Hard - Soft	0.733	-0.222
Inconvenient - Conveniente	0.053	0.864
Dislike - Like	0.102	0.843

Finally, the results shown in Table VI were obtained after rotating the factor loading matrix according to the method of great variance. Through Table VI, it can be found that common factor 1 has larger loadings on factors 1-5 (light pressure; weak skin irritation; lightweight; soft; good-looking), and common factor 2 has larger loadings on factors 6 and 7(easy to put on and take off, preferring long-term use), indicating that these 7 variables can be reduced to 2 male factors: material factor and garment design factor.

Compared with previous studies [17-20], this research designed and developed a complete wearable device instead of a single heart sound collector, and explored the wearing experience of the heart sound collection device, and the experimental participants gave a positive evaluation, all of whom found that the device is suitable for daily wear. In consistent with them the denoised heart sounds were clear, and additionally heart sound detection was added, using a convolutional neural network system to detect the pairs of collected heart sounds and obtain better results.

V. CONCLUSION

In this paper, a comfortable wearable heart sound acquisition device is presented. First, a urethane resin holder was fabricated according to the size and characteristics of the microphone. Then, suitable materials were selected and vests were made according to the physical characteristics of men and women respectively. Then, the microphone wrapped with holder was fixed on the corresponding position of the vest with reference to the heart sound auscultation position. Subsequently, 11 healthy young people (6 males and 5 females) were subjected to heart sound collection, and the collected heart sounds were input into the constructed CNN for testing to evaluate the performance of the device, and the highest accuracy rate of 85.39% for a single male, the highest accuracy rate of 91.59% for a single female and an average accuracy rate of 71.3% were obtained. Finally, the evaluation of the device by the experimental participants was analyzed and 2 common factors were extracted: material factor and garment design factor, indicating that these two factors can be focused on for future updates of the device. This study demonstrates that the wearable heart sound collection device is effective in heart sound collection and has been well evaluated in terms of product design.

VI. FUTURE WORK

The wearable heart sound collection device developed in this study can be used well for heart sound collection, but there are still many shortcomings. First, the design of the zipper in the middle of the chest causes a lot of noise during heartbeat collection, and the position of the zipper needs to be reconsidered and adjusted; second, the position of the microphone cannot be adjusted according to the size of each person and needs to be redesigned; third, a more skin-friendly material could be selected; and finally, the data transmission of the device should be updated to wireless transmission. It is hoped that in the future, a real-time heartbeat detection device can be made that can be worn daily for a long time.

ACKNOWLEDGMENT

This research was supported by JSPS KAKENHI Grant Number 19H04154. We would also like to thank the China Scholarship Council (CSC) for supporting the research of PhD student Huai Ximing, and to teachers and experiment participants who assisted in completing this research.

REFERENCES

[1] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., ... & GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group, “Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study,” Journal of the American College of Cardiology, vol. *76*(25), 2982-3021, 2020.

[2] W. Herrington, B. Lacey, P. Sherliker, J. Armitage and S. Lewington, “Epidemiology of atherosclerosis and the potential to reduce the global burden of atherothrombotic disease,” Circulation research, vol.118(4), pp.535-546, 2016.

[3] M. Nahrendorf and F. K. Swirski, “Lifestyle effects on hematopoiesis and atherosclerosis,” Circulation research, Vol.116(5), pp.884-894,2015.

[4] P. Pandian, K. Mohanavelu, K. Safeer, T. Kotresh, D. Shakunthala, P. Gopal, et al. “Smart Vest: Wearable multi-parameter remote physiological monitoring system,” Medical engineering & physics, vol.30(4), pp.466-477,2008.

[5] Z. Xu, Z. Fang, L. Du, Z. Zhao, X. Chen, D. Chen, et al. “A Wearable Multi-parameter Physiological System,” In Ubiquitous Information Technologies and Applications, pp. 643-648, Springer, Berlin, Heidelberg, 2014.

[6] C. Yu, Q. He, R. Li, J. TAN and L. YU, “Research of Multiple Physiological Parameters Monitoring System for Sub-healthy Groups,” Piezoelectrics and Acoustooptics, vol.35(1), pp.136-139, 2013.

[7] J. Welch, J. Moon and S. McCombie, “Early detection of the deteriorating patient: the case for a multi-parameter patient-worn monitor,” Biomedical instrumentation & technology, vol.46(s2), pp. 57-64, 2012.

[8] S. Dai and Y. Zhang, “A wireless physiological multi-parameter monitoring system based on mobile communication networks,” In 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), IEEE, pp. 473-478, 2006.

[9] M. Cavalleri, R. Morstabilini and G. Reni, “Integrating telemonitoring with clinical informationsystems: a case study,” Proceedings of the 2005 IEEE engineering in medicine and biology 27th annual conference. Shanghai: IEEE, pp.573 – 576, 2005.

[10] A. Pantelopoulos and N. G. Bourbakis, “A survey on wearable sensor-based systems for health monitoring and prognosis,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),vol.40(1),pp.1 – 12, 2010.

[11] A. Lmberis and A. Dittmar, “Advanced wearable health systems and applications - research and development efforts in the European Union,” IEEE Engineering in Medicine and Biology Magazine, vol.26(3), pp.29 – 33, 2007.

[12] P. Grossman, “The LifeShirt: a multi-function ambulatory system monitoring health, disease, and medical intervention in the real world,” Studies in Health Technology and Informatics, vol.108, pp.133 –141, 2004.

[13] F. H. Wilhelm, W. T. Roth and M. A. Sackner, “The LifeShirt: an advanced system for ambulatory measurement of respiratory and cardiac function,” Behavior Modification, vol.27(5), pp.671 – 691, 2003.

[14] HC. A. Hollier, A. R. Harmer, L. J. Maxwell, C. Menadue, G. N. Willson, D. A. Black, et al. “Validation of respiratory inductive plethysmography(LifeShirt) in obesity hypoventilation syndrome,” Respiratory Physiology & Neurobiology, vol.194, pp.15 – 22, 2014.

[15] D. Roy, J. Sargeant, J. Gray, B. Hoyt, M. Allen and M. Fleming, “Helping family physicians improve their cardiac auscultation skills with an interactive cd-rom,” Journal of Continuing Education in the Health Professions, vol.22, pp. 152-159, 2002.

[16] T.-h. CHEN, L.-q. HAN, H.-t. TANG and R.-j. ZHENG, “Research on Analysis Method and Application of Heart Sound Signals,” Journal ofBeijing Technology and Business University (Natural Science Edition), vol.27(2), pp. 35-39, 2009.

[17] C. Beck and J. Georgiou, "Wearable, multimodal, vitals acquisition unit for intelligent field triage." Healthcare technology letters, vol.3(3), pp.189-196, 2016.

[18] H. K. Tiwari and A. Harsola, "Development of embedded stethoscope for Heart Sound." 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE, pp.1547-1551, 2016.

[19] J. Kirchner, S. Souilem and G. Fischer, "Wearable system for measurement of thoracic sounds with a microphone array." 2017 IEEE SENSORS. IEEE, pp.1-3, 2017.

[20] C. Aguilera-Astudillo, M. Chavez-Campos, A. Gonzalez-Suarez and J. L. Garcia-Cordero, "A low-cost 3-D printed stethoscope connected to a smartphone." 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp.4365-4368, 2016.

[21] S. Shimizu, M. Otani and T. Hirahara, "Frequency characteristics of several non-audible murmur (NAM) microphones." Acoustical science and technology, vol.30(2), pp.139-142, 2009.

[22] B. Karnath and W. Thornton, "Auscultation of the heart." Hospital Physician, vol.38(9), pp.39-45, 2002.

[23] L. Feng, B. An, “Partial Label Learning by Semantic Difference Maximization,” IJCAI ,pp. 2294-2300, 2019.

[24] X. Huai, S. Kitada, D. Choi, P. Siriaraya, N. Kuwahara and T. Ashihara, “Heart sound recognition technology based on convolutional neural network,” Informatics for Health and Social Care, pp.1-13, 2021.

[25] T. A. Brown, “Confirmatory factor analysis for applied research,” Guilford publications, 2015.

[26] A. G. Yong and S. Pearce, "A beginner’s guide to factor analysis: Focusing on exploratory factor analysis," Tutorials in quantitative methods for psychology, vol. 9(2), pp. 79-94, 2013.

Matters of Neural Network Repository Designing for Analyzing and Predicting of Spatial Processes

Stanislav A. Yamashkin[1]
Anastasiya A. Kamaeva[4]
Institute of Institute of Electronic and Lighting Engineering
National Research Mordovia State University
Saransk, Russia

Anatoliy A. Yamashkin[2]
Geography Faculty
National Research Mordovia State University
Saransk, Russia

Ekaterina O. Yamashkina[3]
Institute of Information Technologies
MIREA — Russian Technological University
Moscow, Russia

***Abstract*—The article is devoted to solving the scientific problem of accumulating and systematizing models and machine learning algorithms by developing a repository of deep neural network models for analyzing and predicting of spatial processes in order to support the process of making managerial decisions in the field of ensuring conditions for sustainable development of regions. The issues of architecture development and software implementation of a repository of deep neural network models for spatial data analysis are considered, based on a new ontological model, which makes it possible to systematize models in terms of their application for solving design problems. An ontological model of a deep neural network repository for spatial data analysis is decomposed into the domain of deep machine learning models, problems being solved and data. Special attention is paid to the problems of storing data in the repository and the development of a subsystem for visualizing neural networks using a graph model. The authors have shown that for organizing a repository of deep neural network models, it is advisable to use a scientifically grounded set of database management systems integrated into a multi-model storage, characterizing the domains of using relational and NoSQL storages.**

Keywords—Repository; deep learning; artificial neural network; spatial data; visual programming

I. Introduction

A significant role in solving the problem of strengthening the connectivity of territories of countries and regions is played by the introduction of effective digital infrastructures of spatial data (SDI), aimed at operational diagnostics of natural-social-production systems (NSPS) and high-precision forecasting of the development of natural processes and phenomena [1]. The core of systems of this class is represented by methods and algorithms for machine analysis of spatial data, which allow solving a whole range of applied problems – anomaly detection, data classification, data fusion. The subject of analysis can be space imagery data, aerial photography, information arrays about natural, social and economic objects with a distributed geospatial organization [2]. The areas of application of the results of the analysis in the national economy are extremely wide - from increasing the efficiency of agriculture to assessing the consequences of natural processes [3, 4].

SDI as a system can be decomposed into subsystems for storing, analyzing and distributing geospatial data. The development of the federal spatial data infrastructure is necessary in order to effectively solve the problem of remote monitoring of mobile objects and geographically distributed resources in order to ensure information connectivity of countries and create conditions for sustainable development of the country [5]. At present, the functioning of spatial data infrastructures should be based on and application of new effective methods, approaches and algorithms for the analysis of spatio-temporal data, which can function both on the basis of classical hard and soft computations based on the complex application of fuzzy logic, neural network models, evolutionary modeling [6].

With the development of the scientific and engineering basis and computing power, the direction of deep machine learning has recently been strengthened, based on the systematic use of many levels of nonlinear information processing for the extraction and transformation of hierarchical features, analysis and classification of patterns. Despite the growing interest in the topic of deep machine learning, this problem area requires serious study in the direction of strengthening the project orientation of the process of using deep neural network models for analyzing large spatial data [7]. This statement is based on the assumption that the processes of designing complex neural networks, determining their hyperparameters, the format of input and output data, and integrating training data arrays should be determined by the features of real problems to be solved [8]. In other words, the use of deep neural network models for solving practical design problems can be effective if the conditions for the consolidation of accumulated knowledge are met, heuristics and rules are collected into a single system, for interaction with which convenient graphical and application program interfaces are organized. Quality documentation on the use and flexible customization of pretrained models is of key importance. The indicated provisions determine the feasibility of forming a repository of deep neural network models in the digital SDI system, which provides system access to the storage of neural network models, as well as tools for choosing the optimal solution in a specific problem area and disclosing the features of its use.

The purpose of this article is to discuss problematic issues devoted to solving the problem of machine learning models systematizing by developing a repository of deep neural network models for analyzing and predicting of spatial processes. The matters of architecture development and software implementation of deep neural network models repository are considered.

II. Related Works

The task of designing and training effective deep neural network models for analyzing large arrays of spatial data encounters many problematic moments that require finding solutions. This article is devoted to solving the scientific problem of accumulating and systematizing models and machine learning algorithms by developing a repository of deep neural network models for analyzing and predicting the development of spatial processes in order to support the process of making managerial decisions in the field of ensuring conditions for sustainable development of regions. The purpose of the study is to provide support for the process of making management decisions in the field of forming conditions for sustainable development of regions through the development and development of a repository of deep neural network models for the analysis of spatial data [9]. The solution to the problem of forming the architecture and software implementation of a repository of deep neural network models should be based on a new ontological model that defines a formalized description of the topologies of deep models, tasks to be solved, a set of analyzed data, learning algorithms, as well as relationships between these entities [10].

The development of information systems based on machine learning algorithms is very different from the development of traditional software applications based on specifications and repeatedly tested algorithms. In addition, the process of training machine learning models can be expensive and computationally intensive. This fact determines the feasibility of creating a tool that provides the ability to reuse pre-trained machine learning models for the development of applied applications. The use of pretrained machine learning models deployed as web services is one of the top technology trends. Popular frameworks such as Tensorflow [11] and PyTorch [12] provide high-level API support for using machine learning model components.

Since deploying and maintaining machine learning models is still a non-trivial task that requires in-depth knowledge of machine learning and systems administration, organizations are building repositories of machine learning models in order to simplify the use of models within their business processes. Other companies form repositories that can be accessed by third parties. The Wolfram Neural Net Repository (launched in June 2018) [13] and the AWS marketplace (launched in November 2018) [14] are useful examples. Each model repository has a different structure and uses different heuristics to group the models. For example, AWS labels each model using a set of criteria (including the analyzed data), and each model can meet several criteria. Machine learning model repositories promise to close the operational gap between data scientists and software developers. In the field of spatial data analysis, the task of developing a repository of neural network models has its own specifics.

III. Methodology and Methods of Research

Analysis of the strengths and weaknesses of the currently existing repositories of deep neural network models of general purpose (AWS Marketplace, Open Neural Network Exchange, Wolfram Neural Net Repository and others) [15] made it possible to form a list of specific problems, the solution of which will ensure the creation of an effective system suitable for solving specific practical oriented tasks in the field of spatial data analysis.

1) Designing an ontological model for the storage of deep neural networks.

2) Development of a storage scheme for models of deep machine analysis of spatial data in the form of a meta-language.

3) Implementation of the function of converting repository models into representations used by modern software systems for machine learning.

4) Development of graphical web interfaces for a repository of deep neural network models with navigation functions.

5) Creation of a user authentication subsystem in order to differentiate access rights to deep neural network models.

6) Development of interfaces for obtaining structured information about specific neural network models.

7) Development of a subsystem for visualizing deep machine learning models in the form of graph-schemes within the framework of adaptive web interfaces.

8) Development of an application programming interface based on the architectural style of REST, to provide a unified program exchange.

9) Updating the repository of deep neural network models in the process of solving practice-oriented tasks in the field of ensuring conditions for sustainable development of Russian regions.

10) Development of a recommendation system for the selection and configuration of deep neural network models.

Work on the design, development and implementation of a repository of deep neural network models involves obtaining results characterized by scientific novelty. So, the ontological model of the storage of deep neural networks should be distinguished by the achievement of a detailed and comprehensive systematization of deep machine learning models used to analyze large spatial data by classes of design problems to be solved, types of analyzed data, architecture, objective (numerical) and subjective (expert) performance metrics [16]. On the basis of the ontological model, a conceptual storage scheme for deep machine analysis models can be created in the form of a meta-language, which allows them to be converted into representations used by modern machine learning frameworks [17].

Graphical web interfaces of the repository of deep neural network models should be justified from the standpoint of system analysis of user experience, allowing the selection of a relevant machine learning model for solving specific problems of spatial data analysis, obtaining systematized information about the required deep neural network model. It is also supposed to justify and deploy a subsystem for visualizing deep learning models through web interfaces in the form of a graph diagram, with the function of online editing of the model topology within the capabilities of a thin web client, as well as organizing an application programming interface based on REST architectural style to provide unified programmatic interaction with the repository. Of particular importance is the design and development of a new recommender system for the selection and configuration of deep neural network models, which allows for a relevant search for an effective architectural solution and its fine tuning for solving design problems through a graphical web interface of the neural network repository [18]. The repository of deep neural network models acquires practical value if the database is updated with new models for the analysis of spatial data, including remote sensing data. Each deep neural network should be tested on test sites in order to determine estimates of its effectiveness in solving practical problems: classification of remote sensing data, forecasting the development of natural processes.

IV. Results and Discussion

A. Development of an Ontological Model of a Repository

The development of a deep neural network repository should be based on an ontology that provides a formalized description of entities (ANN topologies, learning processes and accuracy assessment), as well as the relationships between them [19]. The relevance of this provision is determined by the hypothesis that the process of using deep learning to solve design problems can be supported only if the acquired knowledge, heuristics and rules are collected in a system for which convenient means of interaction are organized.

An ontological model of a deep neural network repository for spatial data analysis can be decomposed into domains of deep machine learning models, tasks to be solved, and data (Fig. 1). This allows you to give a comprehensive definition of the formalized area of knowledge: each stored model must be compared with a set of specific tasks and data sets (tensor, raster, vector, attributive). The domain of machine learning models is defined by concepts and relationships that describe various topologies (branching and chain structures of layers of various types (including fully connected, convolutional, recurrent) with various activation and regularization functions), as well as learning methods and algorithms.

The deep model is characterized by the format of the data received for analysis and the type of the output signal, the loss function, the initialization algorithm, the learning strategy (with a teacher, without a teacher, with partial involvement of a teacher, with reinforcement). Each deep model is associated with a meta description containing its category, characteristics of the project-oriented tasks to be solved, and guidelines for applied use. The design solution must store pre-trained copies, ready for solving design problems after fine-tuning. The software implementation of the repository was carried out on the basis of the artifacts of the object-oriented design stage, including the determination of the use cases of the system, the construction of structural and behavioral UML diagrams. The development of graphical interfaces of the repository must be carried out using UI / UX design methodologies using a software stack of web technologies, which will allow you to use the repository from anywhere in the world connected to the Internet. The API for interacting with external systems is based on the REST architectural style.

Thus, the process of forming a repository of deep neural networks in a digital SDI system should be based on a project-oriented approach, based on which each stored deep ANN should be compared with the range of design tasks within which it can be used.

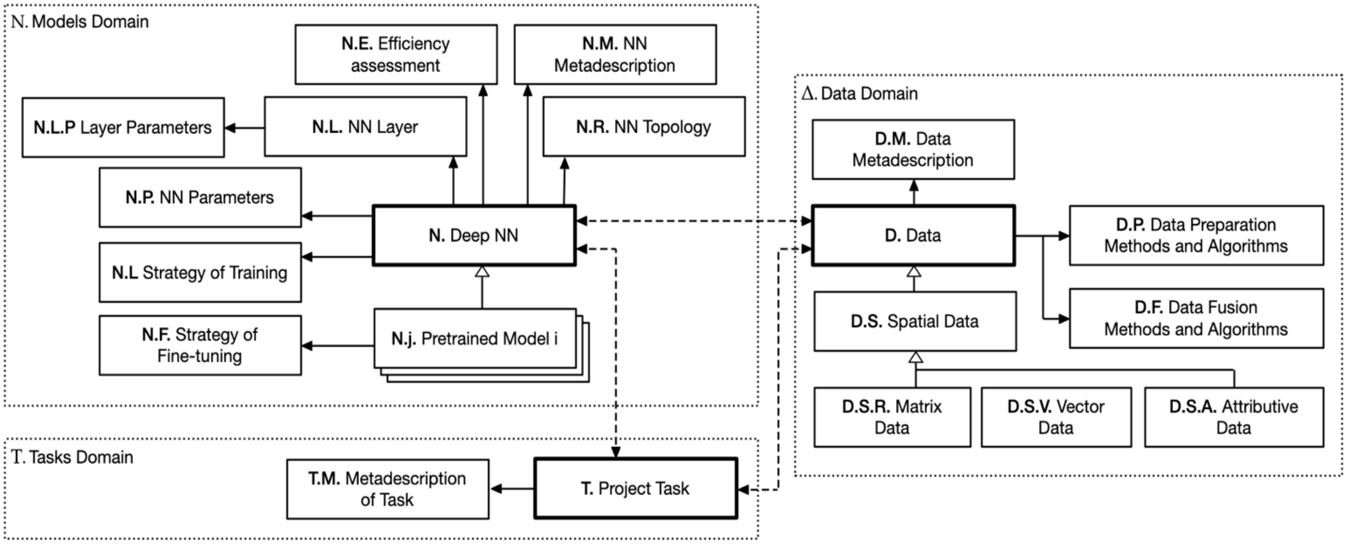


Fig. 1. Ontological Model of the Repository.

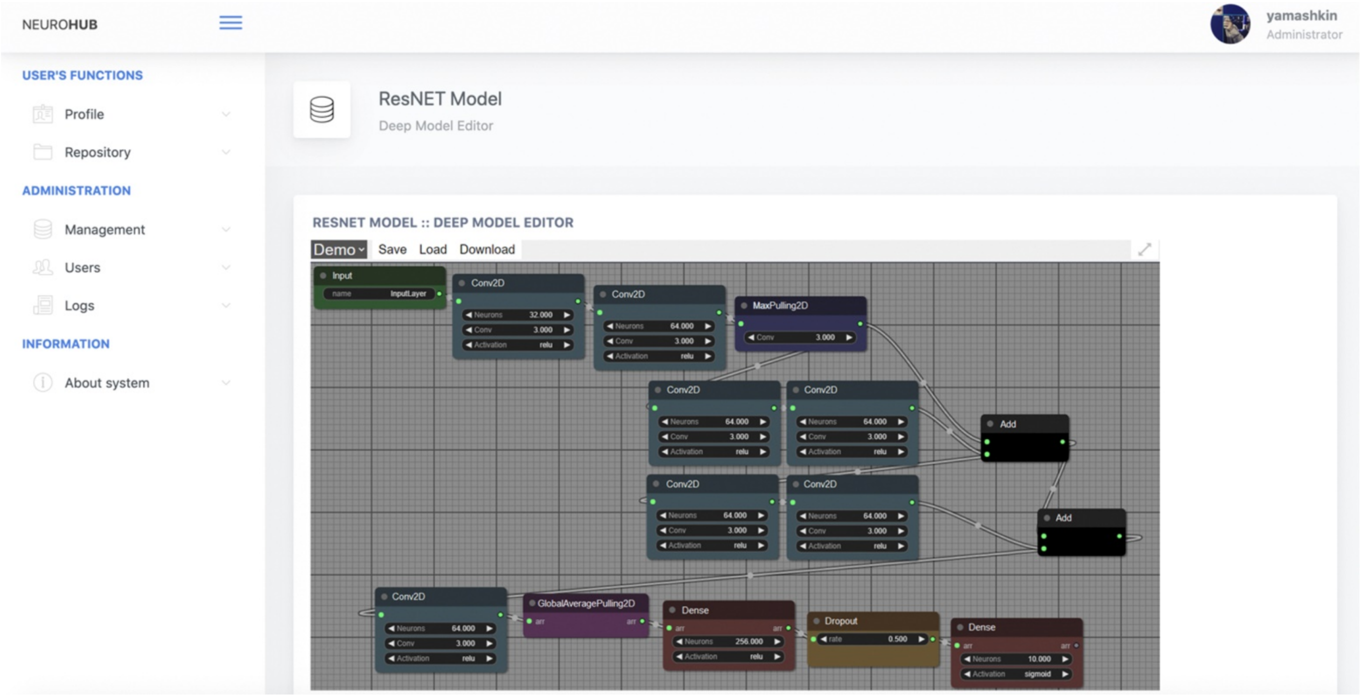
B. Designing the Storage Subsystem of Repository

To organize a data warehouse of a repository of neural networks, it is advisable to use a comprehensive science-based set of database management systems (DBMS). So, in relational storages, the organization of the data integration process is based on entities and relationships established between them. The highly organized structure and flexibility makes relational storage powerful and adaptable to different types of data domains, tasks, and models. To store spatial data, it is advisable to use a relational DBMS with support for spatial operations: PostgreSQL MySQL. Alternative advantages to the process of integrating spatial data can be provided by NoSQL storages, which should be classified into the following categories: a) resident databases - Redis (preferred for the development of data caching systems, buffers of high-speed exchange with the repository through software interfaces); b) document databases - MongoDB, RethinkDB (relevant for systematization of semi-structured data analyzed by neural networks, such as information for monitoring spatial processes, storage and registration of events); c) graph databases - Neo4j, JanusGraph (useful for storing information about the structure of neural network models); d) columnar databases - ClickHouse, Cassandra (represent an uncontested solution for organizing interactive analytical data processing modules (OLAP components) in a neural network repository); e) time series databases - InfluxDB, TimescaleDB (implemented to collect and manage spatial data that change over time, including indicators of the development of natural processes and transactions of the Internet of Things). Thus, for organizing a repository of deep neural network models, it is advisable to use a scientifically grounded set of database management systems integrated into a multi-model storage.

C. Development of Interfaces for Interacting with the Repository

The software implementation of the repository was carried out on the basis of artifacts of the object-oriented design stage, including the definition of the use cases of the system, the construction of structural and behavioral UML diagrams with the optimization of the object-oriented metrics of the system. The development of graphical interfaces of the repository was carried out using UI / UX design methodologies using a software stack of web technologies, which will allow you to use the repository from anywhere in the world connected to the Internet. The API is implemented using the REST architectural style. The graphical web interfaces of the deep neural network model repository should be equipped with navigation functions that allow the selection of a relevant machine learning model for solving specific problems of spatial data analysis from any device connected to the Internet (via a web browser). Subsystems for user authentication in the repository system have the task of differentiating the rights to select, use and edit models.

The developed web interfaces solve the problem of providing information about a specific neural network model (including a systematized description, architecture class, format of analyzed and output data, information about topology, subjective and objective performance indicators, recommendations for fine-tuning the model, examples of practice-oriented use). To improve the convenience of interacting with the repository, a subsystem for dynamic visualization of learning models was developed based on adaptive web interfaces in the form of dynamically aligned graphs, with an interactive ability to directly edit the architecture and topology of the neural network model through a thin web client (web browser) (Fig. 2).



Input: name InputLayer

Conv2D: Neurons 32.000; Conv 3.000; Activation relu

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

MaxPulling2D: Conv 3.000

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

Conv2D: Neurons 64.000; Conv 3.000; Activation relu

GlobalAveragePulling2D: arr arr

Dense: arr arr; Neurons 256.000; Activation relu

Dropout: rate 0.500

Dense: arr arr; Neurons 10.000; Activation sigmoid

Fig. 2. Graphical Interfaces for Editing a Deep Neural Network Model.

In the field of neural networks, there are a huge number of libraries, modules and patterns that can be implemented, combined and used for a wide range of tasks. For a visual display of the developed blocks with the aim of their effective use, you can turn to the visual design paradigm. Visual programming is the process of graphically representing a program using a standard set of graphical elements. With visual development, the number of forced control errors in the program is significantly reduced, therefore, the quality of the result obtained is increased.

When developing visual programming languages, drawing of graphs is used as the main approach, that is, in the form of a set of vertices (nodes) connected by arcs (edges). Unlike the textual form of notation, in which objects (symbols and words) form a sequence, and each object is associated only with the left and right "neighbors", the graph form allows you to visually depict more complex relationships, since in it each object can be connected to several others objects. In this sense, the text form is one-dimensional, while the graph form is two-dimensional. The ability to vary the geometric dimensions, shape and color of vertices, the appearance and thickness of arcs, change the relative position of the vertices without changing the topology of the graph significantly increases the expressive capabilities of the graph form of the program algorithm.

The creation of a visual (figurative) style of software development is the main motive for the development of graph-symbolic programming (GSP) technology. GSP technology is a technology for designing and coding algorithms and models based on a graphical way of representing programs, with the goal of fully or partially automating the design, coding and testing of software. This programming technology adheres to two fundamental principles: a) visual, graphical form of presentation of program algorithms and other components of their specifications; b) the principle of structured procedural programming. The implementation of this concept is excellent for solving the problem of visual programming of neural networks. Visual programming increases the clarity of the presented codes, significantly reduces the number of errors made at the design and coding stage of programs, and thereby speeds up the development process and increases the reliability of codes of developed programs. Together with the use of the block approach, visual programming methods will not only speed up development due to simple graphical interchangeability of logical blocks, but also provide simple portability of such programs. An application programming interface (API) based on the REST architectural pattern provides a unified interaction with the system for programmatic data exchange, including deep model export, editing, structured data retrieval, and other use cases.

The main directions for the further use of the expected results:

1) Implementation of intelligent systems for forecasting the development of natural and man-made emergency processes based on new technologies of integration, intellectual analysis and dissemination of large geospatial data into the activities of industrial enterprises and executive authorities of the region.

2) Provision of services for the use of a repository of deep neural network models according to the SaaS (Software as a Service) model with the possibility of flexible configuration of the provided solution.

3) Design, development and implementation of geoportal solutions aimed at creating information support for sustainable development of ecological, socio-economic systems of regions, effectively modified for specific corporate customers on the basis of a project-oriented approach. Development and development of socially-oriented geoportal solutions that ensure the effective dissemination of spatial data about nature, economy, social life, history and culture of the regions of the world.

4) Development and implementation of practice-oriented educational programs in the field of sustainable development of regional and global ecological-socio-economic systems with annual bilateral internships and practices to form competencies in the field of information support for sustainable development of world regions and global ecological-socio-economic systems.

The repository of neural networks gains practical value provided that the deep model storage is constantly updated. In the course of further work on the project under a grant from the President of the Russian Federation, it is planned to create and test new models of deep machine analysis, which will be tested in solving specific scientific and practical problems in the field of analyzing the state of natural geosystems, predicting the development and assessing the consequences of emergency situations.

V. Conclusions

The conducted research allows us to determine the reference points for the design, development and implementation of a repository of deep neural network models:

1) Ontological model of the repository, which determines the principles of systematization of deep models for the analysis of spatial data according to the classes of problems to be solved, the nature and dimension of the analyzed data, architecture and topology, properties of efficiency.

2) A formalized storage scheme for deep machine analysis models of spatial data in the form of a meta-language that allows them to be converted into representations used by modern machine learning software systems (Caffe, Torch, MXNet, TensorFlow, Keras).

3) Adaptive graphical web interfaces of the repository with navigation functions that allow you to select the relevant machine learning model for solving specific problems of spatial data analysis.

4) Subsystem for user authentication in the repository in order to restrict the rights to read, use and edit neural network models.

5) Adaptive web interfaces for obtaining systematized information about the required deep neural network model, including a systematized description, objective and subjective performance indicators, type of architecture and topological scheme (branching and chain structures of layers of various types: fully connected, convolutional, recurrent), practical recommendations for flexible setting hyper-parameters of the model, examples of applied use.

6) Subsystem for visualization of the neural network model of deep learning in the form of a graph diagram, with the possibility of interactive online editing of the topology and architecture of the model through a web browser, a thin client.

7) Recommended web system for the selection and configuration of deep neural network models, which allows to provide relevant search and fine-tuning for solving specific design problems in the field of spatial data analysis.

8) An application programming interface based on the REST architectural style, which allows for a unified interaction for programmatic data exchange with the repository (including import and export of deep models, obtaining information about them).

9) Database of deep neural network models repository, which includes pretrained models based on convolutional and recurrent layers, regularization, normalization and subdescritization modules with the possibility of their direct use for solving applied problems or further fine tuning.

The implementation of the project will make it possible to form a platform solution for solving urgent scientific problems of consolidation, storage, selection and effective use of deep models for solving project-oriented problems in the field of analyzing large arrays of spatial data. Integration of neural network models into the repository will allow starting the formation of a bank of intelligent algorithms designed to solve specific problems in the field of spatial data analysis to ensure sustainable development of regions, but also to solve the problem of interactive search for an effective model by forming a system of recommendations and developing expert tools that optimize the choice of algorithms.

Acknowledgment

This work was supported by Grant of the President of the Russian Federation under Project no. МК-199.2021.1.6.

References

[1] S. Yamashkin, M. Radovanovic, A. Yamashkin, and D. Vucovic, “Using Ensemble Systems to Study Natural Processes,” Journal of Hydroinformatics, vol. 20, no. 4, pp. 753–765, Apr. 2018.

[2] R. A. Schowengerdt, Remote sensing: models and methods for image processing, 3rd ed. Orlando, FL, USA: Academic Press, 2006, pp. 387–456.

[3] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8–36, Oct. 2017, DOI. 10.1109/MGRS.2017.2762307.

[4] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” IEEE Geoscience and Remote Sensing Magazine, vol. 4, no. 2, pp. 22–40, Jun. 2016, DOI. 10.1109/MGRS.2016.2540798.

[5] S. V. Kovshov, and S. Tingnting, “Application of Computer Modeling for the Accident Rate Assessment on Separate Sites of the Mohe–Daqing Oil Pipeline in Permafrost Conditions,” Transportation Infrastructure Geotechnology, vol. 7, pp. 605–617, Jul. 2020.

[6] C. Tao, H. Pan, Y. Li, and Z. Zou, “Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification,” IEEE Geoscience and Remote Sensing Let., vol. 12, no. 12, pp. 2438–2442, Dec. 2015, DOI. 10.1109/LGRS.2015.2482520.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436, May 2015, DOI. 10.1038/nature14539.

[8] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui “Deep learning-based classification methods for remote sensing images in urban built-up areas,” IEEE Access, no. 7, 36274-36284, Mar. 2019, DOI. 10.1109/ACCESS.2019.2903127.

[9] H. Wu and S. Prasad, “Convolutional recurrent neural networks for hyperspectral data classification,” Remote Sensing, vol. 9, no. 3, 298, Mar. 2017, DOI. 10.3390/rs9030298.

[10] W. Li, H. Fu, L. Yu, P. Gong, D. Feng, C. Li, and N. Clinton, “Stacked Autoencoder-based deep learning for remote-sensing image classification: a case study of African land-cover mapping,” International journal of remote sensing, vol. 37, no. 23, pp. 5632–5646, Dec. 2016, DOI. 10.1080/01431161.2016.1246775.

[11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, and X. Zheng, “Tensorflow: A system for large-scale machine learning,” 12th {USENIX} symposium on operating systems design and implementation, pp. 265-283, 2016.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” arXiv preprint arXiv:1912.01703, 2019.

[13] J. V. Alva, “Neural Network Framework,” Beginning Mathematica and Wolfram for Data Science, pp. 375-406, 2021.

[14] Amazon Web Services Marketplace – Machine Learning. Accessed: May. 2021. [Online]. Available: https://aws.amazon.com/marketplace/solutions/machinelearning.

[15] Wolfram Repository of Neural Network Models, Apr. 2021, [online] Available: http://resources.wolframcloud.com/NeuralNetRepository.

[16] D. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, Jan. 2011, DOI. 10.9735/2229-3981.

[17] E. Yamashkina, S. Kovalenko, and O. Platonova, “Development of repository of deep neural networks for the analysis of geospatial data,” IOP Conference Series: Materials Science and Engineering, vol. 1047, no. 1, 012124, Feb. 2021.

[18] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv: 1502.03167 [cs.LG], Mar. 2015.

[19] M. Xiu, Z. M. J. Jiang, B. Adams, “An Exploratory Study on Machine-Learning Model Stores,” arXiv preprint, 1905.10677, 2020.

Bluetooth-based WKNNPF and WKNNEKF Indoor Positioning Algorithm

Sokliep Pheng[1*]
School of Information and Communication
Guilin University of Electronic Technology
National and Local Joint Engineering Research Center of Satellite Navigation and Location Service
Guilin, China

Ji Li[2], Yanru Zhong[4]
Institute for Artificial Intelligence Interdisciplinary Research
Guilin University of Electronic Technology
Guilin, China

Luo Xiaonan[3]
Guangxi Key Laboratory of Intelligent Processing of Computer Image and Graphics
School of Computer Science and Information Security
Guilin University of Electronic Technology
Guilin, China

***Abstract*—Indoor Positioning System (IPS) in generally perform as a network of devices that always located the objects or people inside a building wirelessly. An IPS has direction relies nearby anchors and also can be entirely local to your smartphone. With the rapid growth and sharp increase in Indoor Positioning System (IPS) demand in the world, there are a lot of researchers trying to invent new algorithm to develop IPS. This paper proposed the Bluetooth-Base Indoor Positioning Algorithm. The RF characteristics such as RSSI and WLAN RSSI fingerprinting system normally formed by two phases, fist is offline phase and second is online phase. Fingerprinting system handling both off-line and online data and estimate the user's location. Our algorithm design is a collection of Weighted K-Nearest Neighbors (WKNN) and Filtering algorithms by KALMAN Filter. Finally, to avoid the problems of IPS and get a better accurate we proposed two algorithms: Weighted K-Nearest Neighbors Particle Filter (WKNNPF) and Weighted K-Nearest Neighbors Extended Kalman Filter (WKNNEKF) compare to KNN and WKNN result. After comparing we found that the result of WKNNPF and WKNNEKF is better result than KNN and WKNN. The Probability in 3M of WKNN is about 79%, WKNNEKF is about 89%, and WKNNPF is about 95.1%. Among one of the proposed algorithms WKNNPF is better than WKNNEKF on accuracy 1.7-2 meters with 42.2m/s response time.**

Keywords*—Indoor Positioning System (IPS); Bluetooth low energy; WLAN; RSSI; WKNNPF; WKNNEKF; KNN; WKNN

I. Introduction

In present, indoor positioning system became more interested and more advantages for the people in the world. Many applications that we have seen some papers introduced in field of m-commerce that based on the principle of the well-estimated location of the various customer as well as in wireless network sector. For example here, it is talking about advertising in large stores or guides in museums which using modern portable devices is possible, in case if we estimate the exact location of a mobile terminal in every single time. Moreover, an automated delivery developed procedure, which is the method based on the user location is also necessary in term of to provide quality service in the area of wireless network, especially in condition of overweight shipments [1].

Bluetooth certification which is developed and manufacturing by the Special Interest Group (SIG) has widely point the technology that suitable for future use in the home or any indoor environment [2][3]. Hence, the totally proposed ideas of this paper was notice about the challenge with all of the issues faced in location estimation plus with the general evaluation criteria which focus on a Bluetooth-based indoor positioning system as well. Although, we can realized that the Location estimation take place in a mobile terminal without any changes of the values in the geometry network zone. The system developed as usual and it's based on a well-known and well-publicized of the triangular method by using the theoretical of received signal strength of the surrounding environment of the Bluetooth access point we have.

Anyways, to obtain precise position estimation, we must determine the dependence between distance and the received signal strength by specific condition carefully. Especially in indoor areas, some relevant boundary conditions, like make use of equations reflection and the wall drying for free as well as propagation impossible. Then, the necessary distance is mainly focus and calculated as well as by an approximate estimate of the received signal strength indicator (RSSI) in simple ways [4][5]. As we known that, the Bluetooth devices of retrieving the actual received signal strength do not provide any interface as well as we want, so our research developed based on this positioning system focus on the point which using the RSSI values that provided as mentioned as in the standard technic to obtain a range estimate between the access point and the mobile terminal as well. Although, this research demonstrated the theory of mathematical system and introduced the mathematical approach by positioning system is based. Our steps start from the beginning with position estimating which mainly based on signal strength using the Least Square

[*] Corresponding Author

Estimation (LSE), and then we have converted the RSSI measurement to the distance. After that, the implementation results of the actual test are presented.

II. Related Work

A. The Position Estimation based on Signal Strength

In this part we proposed the related technic which is a part of achieving the result. As shown in below equation, a position estimation method based on signal strength by using LSE is introduced and explain in detail as well.

In the mobile network we assumed N≥2 as the number of a base station and the position of a base station k is defined

by: $\overrightarrow{p_k} = (x_k, y_k)^T, k \in 1..N.$

Hence, the distances $r_i(\vec{x}), r_j(\vec{x})$ between stations i, j and a point in the x-y zone, has given as well as by

$\vec{x} = (x, y)^T$:

$$\begin{cases} r_i(\vec{x}) = \sqrt{(x-x_i)^2 + (y-y_i)^2} \\ r_j(\overrightarrow{x}) = \sqrt{(x-x_j)^2 + (y-y_j)^2} \end{cases} \quad (1)$$

The solution of the system of equation (1) as shown in above section, it causes two possible intersections of the corresponding circles. So, to get a perfect solution to solve this problem, it's importantly to calculate and set the location of the mobile device which based on the distance between the terminal and at least three different base stations among them is included [6]-[9]. Other thing, the single distances also can be obtained as well as by measuring the functional correlation to the signal strength. If the value of the distance between both points, the base station N>2 and the mobile terminal is already known. Otherwise, the location estimate also can be effectively as well by calculated using the LSE method. In numerical operations, this method calculates that point correctly in the x-y field, where the position that provides the least squared sum of the distance to the boundary of all possible parts given by Equation (2) as describe as below.

$$\begin{cases} r_i = \sqrt{(x-x_i)^2 + (y-y_i)^2} \\ r_j = \sqrt{(x-x_j)^2 + (y-y_j)^2} \end{cases}, i, j \in 1..N \quad (2)$$

Boundaries are directly calculated from Equation (3)

$$r_i^2 - r_j^2 = (x-x_i)^2 + (y-y_i)^2 - (x-x_j)^2 - (y-y_i)^2 \quad (3)$$

The position estimation is defined by $\overrightarrow{m} = (\hat{x}, \hat{y})^T$.

$$H \cdot \overrightarrow{m} = C \quad (4)$$

With

$$H = \begin{bmatrix} h_{x(2,1)} & h_{y(2,1)} \\ h_{x(N,1)} & h_{y(N,1)} \\ h_{x(N,N-1)} & h_{y(N,N-1)} \end{bmatrix}_{\binom{N}{2}\times 2}, \overrightarrow{m} = \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}, C = \begin{bmatrix} C_{2,1} \\ C_{N,1} \\ C_{N,N-1} \end{bmatrix}_{\binom{N}{2}\times 1}$$

And $\begin{cases} h_{x(i,j)} = 2(x_j - x_i) \\ h_{y(i,j)} = 2(y_j - y_i) \\ C_{i,j} = r_i^2 - r_j^2 + x_j^2 - x_i^2 + y_j^2 - y_i^2 \end{cases}$

Hence, below is the formula of location estimation of the terminal:

$$\overrightarrow{m}^T = (H^T H)^{-1} H^T C \quad (5)$$

B. The Approximation of RSSI Measurements

This section talked about the way of converting the signal strength measurement into distances between shape of a sender and receiver in the free fields. Below is the equation (6) [10]:

$$s(\vec{x}) = c(d(\vec{x}))^{-\alpha} \quad (6)$$

Variations of signal propagation in indoor areas normally caused by dimming and reflection are strongly considered in easy way as well as by equation (6). Because there is no line of reflection and attention, this happened because those holding Bluetooth device obtain signal propagation in the empty which not suitable for indoor area. Hence, according to reason above we chose to estimate the correlation between signal strength and distance by focusing on on the measurement to improve our estimation as well. Bluetooth specification at that moment didn't release the possibility to extract signal directly. Consequently, after go through reason above we are using the RSSI values set by the Bluetooth protocol to connect the communication between the sender and the receiver in the network as well as we have shown below [11]-[14]. The RSSI value role as important paly that it is giving the distance between the received signal strength and the optimal receiver power rating, called the gold receiver power status. The definition of a golden power receiver has explained and shown in Fig. 1 below.

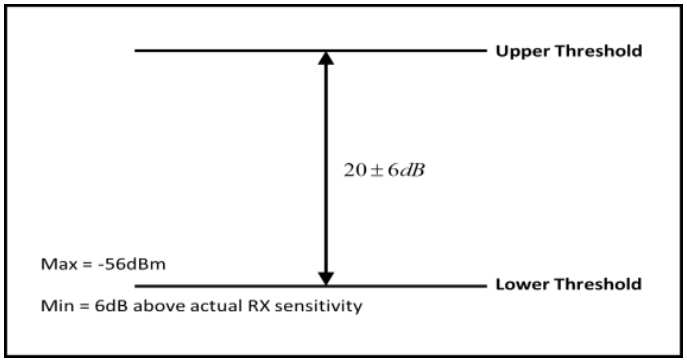


Fig. 1. Golden Receiver Power Rank of the RSSI.

Here, we explain the Fig. 1 above about the golden receiver power status about it working principle. It is determined by two levels which describe below [15][16]. First is the low level that noted by the 6dB to the actual the receiver and maximum of this value mentioned by -56dBm. Second is the high level is 20dB on the low one. It provided accuracy of the upper threshold and the noted with the value is about ±6dB. Hence, we can assume S assigned the received signal strength and the value of S is noted by:

$$S = RSSI + T_0, for RSSI > 0$$
$$S = RSSI - T_{u,} for RSSI < 0$$
$$T_0 = T_u + 20dB \qquad (7)$$

Where: T_0 : Upper Threshold

T_u : Lower Threshold

Normally, pursuant to the definition of a gold recipient rating determines the RSSI to distance conversion. Although, if the value of RSSI is within the range of the golden receiver defined by zero, it’s mean that, there is no special function can be estimated as well as we can [17]. Therefore, only measurements that provide result in a positive range of RSSI could be considered and granted by functional estimation. Then we can achieve the estimate result by the parameterization in suitable function we chose:

$$y = c \ln x + b,$$
$$y = c_0 + c_1 x + c_2 x^2$$
$$y = c_0 + c_1 x + c_2 x^2 + c_3 x^3 \qquad (8)$$

III. Methodology

Our research methodology used the Bluetooth positioning system model to provide a user interface as well as separated into three levels. The indoor area under consideration is covered by the x-y is the first stage of preparation phase, then RSSI function is covering to the environmental conditions which can estimated as well. Hence, the last part of our solution we used the location method to determine the mobile terminal location by dealing based on the RSSI estimation measurement. In this relation, we are mainly focus on the triangular method in term of mathematical calculations in above section by calculation RSSI values of three entry points. Below is the flowchart of the matching process which mentioned as Fig. 2, it is detail and overview of the operational analysis.

In addition to improve the fingerprint method (RSSI), our algorithm design is a collection of Weighted K-Nearest Neighbors (WKNN) and Filtering algorithms (KALMAN Filter). Other thing, to avoid the problems of IPS and get a better accurate we proposed two algorithms are WKNNEKF and WKNNPF.

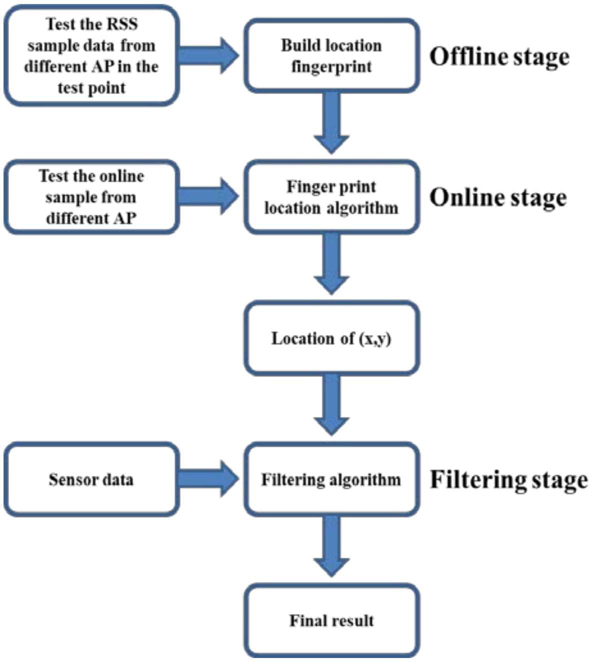

Test the RSS sample data from different AP in the test point
Build location fingerprint
Offline stage
Test the online sample from different AP
Finger print location algorithm
Online stage
Location of (x,y)
Sensor data
Filtering algorithm
Filtering stage
Final result

Fig. 2. Flowchart of the Matching Process.

The access points have been assigned and share in a laboratory room of school of information and communication engineering, Guilin University of Electronic Technology. That room is a computer room with a size of 12 x 21m which is shown in Fig. 3.

In this experiment we have done carefully. We were recorded several times of the distances of the randomly selected sections to archive a good approximation function between the RSSI and the single access point as well as we can. The average results among of all measurements which belonging to a segment and access point may generate the RSSI reference value per segment. In our scenario, we can realize from Fig. 2 and 3 and we can explain that the RSSI values of all access points are best equipped with the polynomial function of the order 3. The positioning algorithm has been implemented in MATLAB.

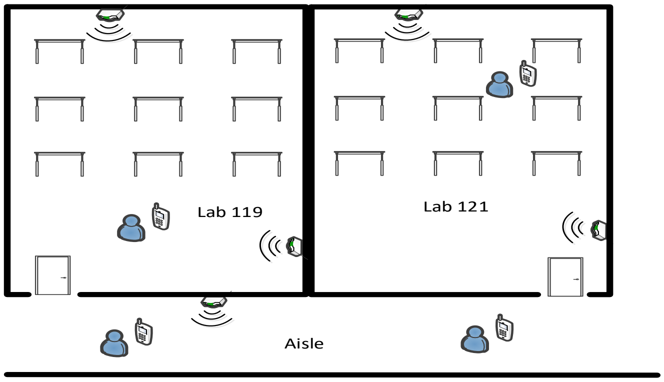


Fig. 3. Indoor Environments Experimental.

IV. Results and Analysis

According to the experiment methodology, we separated results with 3 different algorithms are WKNN Algorithm in Fig. 4, WKNNEKF Algorithm in Fig. 5, WKNNPF Algorithm in Fig. 6.

After we got the results in each algorithm of Fig. 4, 5, 6. We found that the accurate point of WKNN algorithm is bigger than WKNNEKF and WKNNPF algorithm, and the accurate point of WKNNEKF algorithm is smaller than WKNN and bigger than WKNNPF. And the last one is the accurate point of WKNNPF algorithm is smaller than WKNNEKF and much smaller than WKNN so far. Moreover, we have compared the cumulative probability of the positioning error to find the best performance of each algorithm in Fig. 7. Those algorithms mentioned in three difference colors. Red stand for WKNN algorithm, green stand for WKNNEKF algorithm and blue stand for WKNNPF.

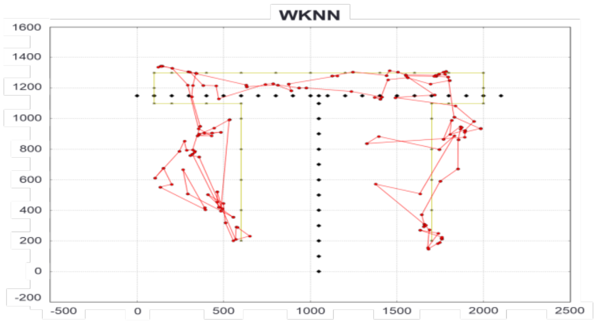


Fig. 4. WKNN Algorithm Result.

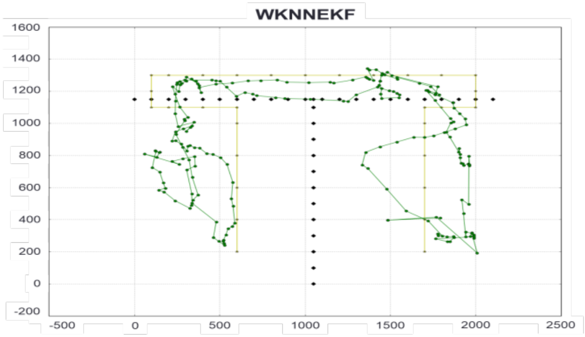


Fig. 5. WKNNEKF Algorithm Result.

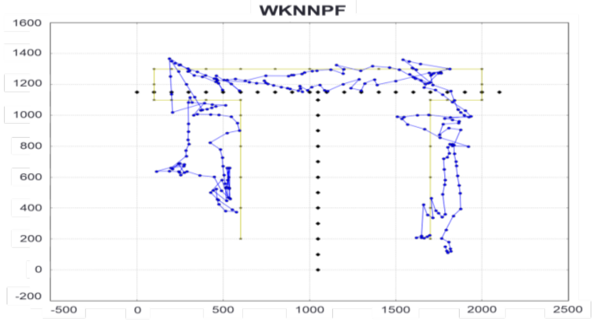


Fig. 6. WKNNPF Algorithm Result.

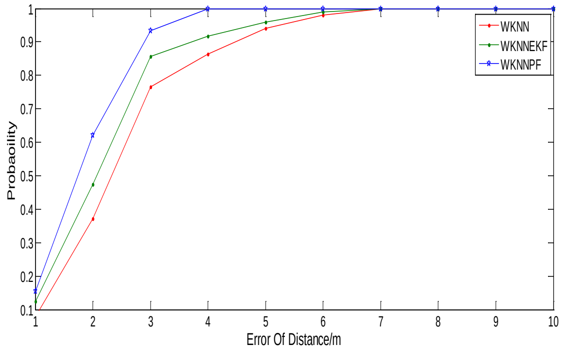


Fig. 7. Errored of Distance/m.

TABLE I. The Probability in 3M

Algorithm	*Mean derivation*	*Minimum derivation*	*Maximum derivation*	*The probability in 3M*
WKNN	3.2m	1.2m	3.8m	79.0%
WKNNEKF	2.5m	0.92m	3.0m	89.0%
WKNNPF	2.0m	0.22m	3.0m	95.1%

The result in Fig. 7 and Table I shows that the probability in 3M of WKNN is about 79.0%, the probability in 3M of WKNNEKF is about 89.0% which is better than WKNN about 10%, and the probability in 3M of WKNNPF is about 95.1% which is better than WKNN about 16.1% and WKNNEKF about 6.1%. It's mean that the proposed algorithm WKNNEKF is better than WKNN and the proposed algorithm WKNNPF is better than WKNNEKF. After we found that WKNNPF algorithm is the best algorithm among the proposed algorithm in this paper, we try to find the algorithm response time of WKNNPF algorithm performance. Then we found that WKNNPF response time is 42.2m/s as shown in Fig. 8 below.

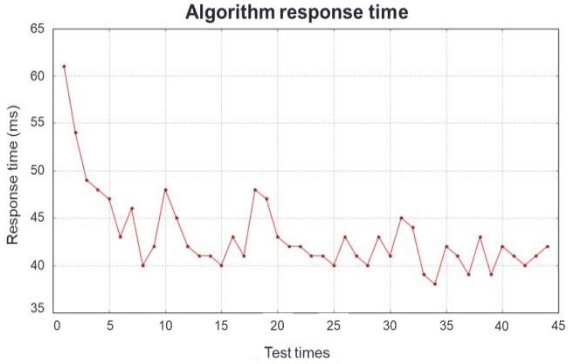


Fig. 8. WKNNPF Performance.

V. Conclusions

In this paper, we carefully experimented with both theory and practice to apply indoor location algorithms. As we all know, there are many previous documents describing IPS with different ways and solutions. Most of them focus on wireless sensors to develop their new algorithms. In a similar way, this paper proposes to study of the Bluetooth-based Indoor Positioning Algorithm. Based on RF characteristics such as RSSI and WLAN RSSI fingerprinting system normally consists of two phases, offline phase and online phase. Fingerprinting system handling both off-line and online data and estimate the user's location. Our algorithm design is a collection of Weighted K-Nearest Neighbors (WKNN) and Filtering algorithms by KALMAN Filter. Finally, to avoid the problems of IPS and get a better accurate we proposed two algorithms: WKNNPF and WKNNEKF which get the better result than KNN and WKNN, and one of the proposed algorithms WKNNPF is better than WKNNEKF on accuracy 1.7-2 meters with 42.2m/s response time.

Acknowledgment

First of all, I would like to thanks to supporter including resources and experiment place. This research can't complete without their support. This work was fully supported by three departments, first is school of information and Communication Engineering of Guilin University of Electronic Technology offered funds and place for experiment and second is the National Natural Science Foundation of China (Nos. 61562016, 61702129, 61772149, and 61320106008) provided software and technic, and third is Guangxi Key Laboratory of Intelligent Processing of Computer Images and Graphics (No. GIIP201703) offered funds and the ideas of algorithm.

References

[1] Specification of the Bluetooth Core System 1.1, www.bluetooth.org.

[2] Nathan J. Muller: "Bluetooth", 1. edition, mitp, Bonn, 2001.

[3] J. B. Andersen, T.S. Rappaport, and S.Yoshida: Propagation messurements and models for wireless communications channels., IEEE Communications Magazine, 33(1):42-49, 1995.

[4] Paramvir Bahl, Venkata N. Padmanabhan: .RADAR: An In-Building RF-based User Location and Tracking System., Proceedings of IEEE INFOCOM 2000, Israel , March 2000.

[5] Thomas Fritsch, Kurt Tutschku, Kenji Leibnitz: Field Strength Predication by Ray-Tracing for Adaptive Base Station Positioning in Mobile Communication Networks, 2nd ITG Conference on Mobile Communication, Sep. 1995.

[6] Yongguang Chen, Hisashi Kobayashi: Wall Map Aided Indoor Geolocation Based on Signal Strength, Communications, 2002. ICC 2002. IEEE International Conference on, Volume: 1, 2002, Page(s): 436 .439.

[7] M. Hellebrandt, R. Mathar; Location Trackings of Mobiles in Cellular Radio Networks., IEEE Transactions and Vehicular Technology 48, Nr. 5, Sept.: 1558-1562.

[8] Rizaldi, Bahri & Pambudi, Doni & Bariyah, Taufiqotul. (2020). IMPLEMENTATION OF BLUETOOTH LOW ENERGY TECHNOLOGY AND TRILATERATION METHOD FOR INDOOR ROUTE SEARCH. JUTI: Jurnal Ilmiah Teknologi Informasi. 18. 57. 10.12962/j24068535.v18i2.a897.

[9] P. Najera, J. Lopez, and R. Roman, ”Real-time location and inpatient care systems based on passive RFID”, Journal of Network and Computer Applications, Volume 34, Issue 3, May 2011, pp. 980989.

[10] L. Pei, et al., ”Inquiry-based Bluetooth Indoor Positioning via RSSI Probability Distributions”, Second International Conference on Advances in Satellite and Space Communications, 2010, pp. 151-156.

[11] L. Pei, et al., ”Using Inquiry-based Bluetooth RSSI Probability Distributions for Indoor Positioning”, Journal of Global Positioning System, Volume 9, Issue 2, 2010, pp. 122-130.

[12] F. Subhan, H. Hasbullah, A. Rozyyev, and S. T. Bakhsh, ”Indoor Positioning in Bluetooth Networks using Fingerprinting and Lateration approach”, International Conference on Information Science and Applications, April 2011, pp. 1-9.

[13] M. Irani, B. Rousso, and S. Peleg, ”Detecting and tracking multiple moving objects using temporal integration”, Computer Vision ECCV 1992, Lecture Notes in Computer Science, Volume 588, 1992, pp. 282-287.

[14] Rodríguez-Damián, María & Vila, Xose & Rodríguez-Liñares, Leandro. (2019). Accuracy of Bluetooth based Indoor Positioning using different Pattern Recognition Techniques. Journal of Computer Science and Technology. 19. e01. 10.24215/16666038.19.e01.

[15] Wang, Yan & Madson, Ryan & Rajamani, Rajesh. (2015). Nonlinear observer design for a magnetic position estimation technique. 6986-6991. 10.1109/CDC.2015.7403320.

[16] Lee, Haemin & Jung, Chang-Sik & Kim, Ki-Wan. (2020). A Position Estimation Technique for Motion Compensation of Synthetic Aperture Radar. The Journal of Korean Institute of Information Technology. 18. 65-75. 10.14801/jkiit.2020.18.5.65.

[17] Attivissimo, F. & Di Nisio, Attilio & Lanzolla, A. & Ragolia, Mattia. (2020). Analysis of position estimation techniques in a surgical EM Tracking System. IEEE Sensors Journal. PP. 1-1. 10.1109/JSEN.2020.3042647.

Exploring Machine Learning Techniques for Coronary Heart Disease Prediction

Hisham Khdair[1], Naga M Dasari[2]
International Institute of Business and Information Technology,
Federation University Associate
Adelaide, Australia

***Abstract*—Coronary Heart Disease (CHD) is one of the leading causes of death nowadays. Prediction of the disease at an early stage is crucial for many health care providers to protect their patients and save lives and costly hospitalization resources. The use of machine learning in the prediction of serious disease events using routine medical records has been successful in recent years. In this paper, a comparative analysis of different machine learning techniques that can accurately predict the occurrence of CHD events from clinical data was performed. Four machine learning classifiers, namely Logistic Regression, Support Vector Machine (SVM), K- Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP) Neural Networks were identified and applied to a dataset of 462 medical instances and 9 features as well as the class feature from the South African Heart Disease data retrieved from the KEEL repository. The dataset consists of 302 records of healthy patients and 160 records of patients who suffer from CHD. In order to handle the imbalanced classification problem, the K-means algorithm along with Synthetic Minority Oversampling TEchnique (SMOTE) was used in this study. The empirical results of applying the four machine learning classifiers on the oversampled dataset have been very promising. The results reported using different evaluation metrics showed that SVM has achieved the highest overall prediction performance.**

Keywords*—*Coronary heart disease; machine learning; prediction; classification

I. Introduction

Heart disease refers to a wide range of conditions that affect the structure and function of the heart. CHD is one of the most common types of heart disease, and it is one of the leading causes of death around the world. CHD occurs when plaque builds up in the walls of the coronary arteries, it restricts blood flow to the heart muscle, and will eventually result in a heart attack. According to the Australian Institute of Health and Welfare (AIHW), CHD was Australia's leading cause of death in 2018, accounting for 17,500 deaths. This accounts for 11% of all deaths in Australia and 42% of all cardiovascular deaths [1].

Traditional risk factors for CHD are thought to be High-LDL cholesterol, low-HDL cholesterol, high blood pressure, diabetes mellitus, smoking, a family history of CHD, age, obesity, and an unhealthy lifestyle [2]. The estimated cost of CHD in 2015–16 in Australia was more than $2.2 billion. Private hospital services and public hospital admitted patient services accounted for a minimum cost of $813 million and $693 million, respectively. The burden on the Pharmaceutical Benefits Scheme (an Australian Government subsidy on medicine) for CHD was estimated to be around $218 million [1].

CHD can however be effectively managed with a change in lifestyle and adopting healthy habits, and hence save the high cost of medical treatment and hospitalization if early detected. With early detection of CHD, patients can have a range of treatments advised by doctors to reduce the risk of future heart problems and relieve or manage symptoms. In this context, electronic health records (EHRs, also called medical records (EMRs)) can be considered a useful resource of information to help medical practitioners in the detection or the prediction of CHD [3-6].

Advances in machine learning and artificial intelligence have motivated many scientists to use such technologies in the early detection of high-risk diseases such as heart diseases, diabetes, various types of cancer [7-9]. Machine learning applied to EHR can be a useful tool for predicting the CHD event with heart disease symptoms [10-12] as well as exploring the most significant clinical features and risk factors that may lead to heart attack and deaths. Clinicians and physicians can take advantage of machine learning for clinical feature ranking and unveil hidden and non-obvious correlations and relationships between patients' data. Several supervised machine learning classifiers were used for this purpose and have achieved success in this regard such as logistic regression, SVM, deep learning, KNN, decision tree [3, 13-15].

However, most of the machine learning models designed for the prediction of CHD have achieved modest accuracy [16], More recent models show some improvements but only in the prediction accuracy though [17, 18]. Moreover, the predicting variables of these models have limited interpretability [5, 12]. Even though scientists have identified a large number of predictors and indicators, there is still no consensus on such clinical features and their roles in affecting the occurrence of CHD [2, 19].

In this paper, we study a dataset of 462 medical records obtained from South African Heart Disease. The dataset is a quantitative sample of males in a heart-disease high-risk region of the Western Cape in South Africa-KEEL[20]. The objectives of this study are:

- To investigate machine learning techniques that achieve a high prediction performance in predicting CHD.
- To identify the most effective machine learning models that achieve the best prediction performance on the given dataset.
- And, to identify the best features that help in achieving

the best performance on the given dataset.

In this study, the machine learning classifiers that have been identified and utilized are Logistic Regression, SVM classifier, KNN, and MLP Neural Network. They were identified based on the literature as well as their performance on the given dataset and the suitability of the nature of the available data. The structure of the paper is as follows, we first discuss the related work in section 2, then we discuss the methodology in section 3, where we describe in detail the dataset, the exploratory data analysis, and the feature selection methods. The experimental framework is presented in section 4, results, discussion, and conclusion are discussed in sections 5 and 6 respectively.

II. Related Work

Several researchers have performed studies on routine clinical data (or EHR) obtained from primary health care centers or family practices to predict the occurrence of heart disease [3, 21-24, 46]. Electronic health records have been used or in combination with several machine learning algorithms to predict CHD [25, 26]. Machine learning algorithms have proved to be efficient techniques in predicting heart diseases [3, 18, 27, 47].

In a study performed on 378,256 instances of patient data obtained from UK family practices, the authors in [3] have used the machine learning algorithms logistic regression, random forest, gradient boosting machine, and neural networks. The authors have established that the algorithms have improved the prediction of heart disease, CHD. Improvements in accuracy according to AUC c-statistic are random forest +1.7%, logistic regression +3.2%, gradient boosting +3.3%, neural networks +3.6% when compared to baseline American Association of Cardiology (ACA) and American Heart Association method. In another study, the researchers in an experimental analysis [22] have applied several machine learning algorithms; Decision Tree, Naïve Bayes, K-nearest neighbors, SVM, Multi-Layer perceptron, radial basis function, and Single Conjunctive Rule Learner individually and in combination on the Cleveland dataset [28] which is available at University of California Irvine (UCI) machine learning repository. The authors have compared the algorithms using Precision, Recall, F-Measure, ROC, and accuracy. Support vector machine has provided the best results in the experiment with 84.15% accuracy and 0.897 F-measure. They have applied bagging, boosting and stacking methods to improve the results.

In recent work, the Cleveland and Statlog [29] datasets are further experimented with, by another set of researchers with impressive results [17]. Statlog dataset contains 270 samples with 150 absence of heart disease and 120 presence of it. Cleveland database contains 303 instances with 164 without heart disease and 139 positive cases. The algorithms used in this work were support vector machine, Logistic Regression, Naïve Bayes, deep neural networks, random forest, decision tree, and k-nearest neighbor. Their experiment results have shown that deep neural networks work better for Statlog database whereas SVM works better for the Cleveland database. However, the accuracy in both cases is very high, a fraction above 97%, which is signficantly high when compared to any other study. While the reported accuracy was very high at 97%, the other metrics such as precision, recall and specificity were not investigated which are important to measure the efficiency of a machine learning algorithm.

In experimenting with ensemble machine learning algorithms the authors in [27] have used 4 different datasets obtained from Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH) to predict CHD. The datasets contain 303, 294, 200, and 123 instances, respectively. All the patient instances were formatted uniformly with 76 attributes each out of which only 29 were used due to missing values. Adaptive boosting algorithm has been used for training and prediction. The experimental results produced accuracy and F-score for the different datasets in the order CCF – 80.14, 0.76; HIC- 89.12, 0.83; LBMC-77.78, 0.87; and SUH-96.72, 0.98.

In an experiment on deep learning, Baccouche et al. [30] have worked on heart disease data consisting of 900 samples with 149 attributes each, out of which 16% are related CHD instances. The data was obtained from Medica Norte Hospital, a Mexican hospital in Mexico. The authors have proposed an ensemble neural network framework with Bidirectional Long-Short Term Memory (BiLSTM) or Bidirectional Gated Recurrent Unit (BiGRU) with a CNN model with an accuracy rate of 91%.

Working on the dataset we are working on, Gonsalves et al. [16] performed experimental analysis using Decision Tree, Naïve Bayes, and support vector machine algorithms on WEKA tool. The accuracies obtained for all the three algorithms are above 70% with Naïve Bayes showing the highest with 71.5%. They have attributed the low accuracy to the small size of the dataset and the class imbalance problem in the dataset.

Based on the literature it is noticed that machine learning techniques such SVM, KNN, MLP Neural Networks, decision tree and boosting algorithms are widely used for predicting coronoary heart disease.

III. Methodology

We present the dataset we used for the experiment, exploratory data analysis, and feature selection methods used in this section.

A. Dataset

The dataset for this study has been retrieved from South African Heart Disease [20], which is a subset of a wider dataset. It has a total of 462 medical observations (instances) and 10 features, 9 as independent clinical features, and 1 is the target variable, a labeled binary class as 0 or 1, i.e., CHD event has been detected for the medical observations as positive or negative. The data is for a group of men from a high-risk area for heart disease in South Africa.

Each high-risk patient was monitored in the dataset and the features retrieved were as follows: systolic blood pressure (Sbp), cumulative tobacco in kg (Tobacco), bad cholesterol also known as low-density lipoprotein cholesterol (Ldl), adiposity, family history of heart disease (Famhist), type-A behavior (TypeA), Obesity, current alcohol consumption (Alcohol),

TABLE I. Description of Features in the Dataset

Feature	Explanation	Type and Range	Null Values
Systolic Blood Pressure	Blood pressure measure against the artery walls as the heart beats	Numerical [101, 218]	no
Tobacco	Accumulative tobacco in the body in (kg)	Numerical [0.0, 31.2]	no
LDL Choles-terol	low-density lipoprotein, also called bad cholesterol	Numerical [0.98, 15.33]	no
Adiposity	Adiposity is a measure of percentage of body fat	Numerical [6.74, 42.49]	no
Family History	Family history of heart disease	Binary [0, 1]	no
Type A Behavior	Type A behavior and personality	Numerical [13, 78]	no
Obesity	Weight-to-height ration measure (body mass index, bmi)	Numerical [0.0, 147.19]	no
Alcohol	Current alcohol consumption	Numerical [15, 64]	no
Age	Age of the patient	Numerical [15, 64]	no
CHD Event target	If Coronary heart disease was detected	Binary 0, 1	no

age at onset (Age), and coronary heart disease (Chd) (yes=1 or no=0).

B. Data pre-processing

The original dataset was in .dat format, we have converted it to .csv, and we edited the name of the columns to be more expressive. We have encoded the existing categorical text values in the original dataset into numerical values to be able to be fitted into machine learning models. The description of the features is shown in Table I.

C. Exploratory Data Analysis

The statistical quantitative characteristics of the dataset for numerical features are described in Table II. It can be noticed that the measurements for LDL Cholesterol, Obesity, and also Type A Behavior has slight differences in the mean value for patients with positive CHD event and negative event. The visualization of the counts of observations of the Family History binary class with respect to the negative CHD events and positive CHD events, as well as frequency of the target class CHD Event in the dataset, are shown in Fig. 1 and Fig.2 respectively. Out of 302 subjects without heart disease, 206 of them do not have CHD in the family history whereas 96 have the family history. For positive cases, 64 of them do not have CHD in the family history and 96 have CHD in the family history.

Fig. 3 presents the distribution of each feature's data based on CHD events with the minimum value, first quartile (Q1), median, third quartile (Q3), and the maximum value. The classes in many features seem overlapping, and several features record many outliers in the dataset. The distribution of the data is also skewed.

TABLE II. Statistical Characteristics of the Dataset

Feature	Full Sample		Full Sample		Full Sample	
	mean	std	mean	std	mean	std
Systolic Blood Pressure	138.33	20.5	143.74	23.68	135.46	17.98
Tobacco	3.64	4.59	5.52	5.57	2.63	3.61
LDL Choles-terol	4.74	2.07	5.49*	2.23	4.34*	1.87
Adiposity	25.41	9.82	54.49*	10.25	52.37*	9.52
Type A Behavior	53.1	9.82	54.49*	10.25	52.37*	9.52
Obesity	26.04	4.21	26.62*	4.39	25.74*	4.09
Alcohol	17.04	24.48	19.15	26.18	15.93	23.5
Age	42.82	14.61	50.29	10.65	38.85	14.88

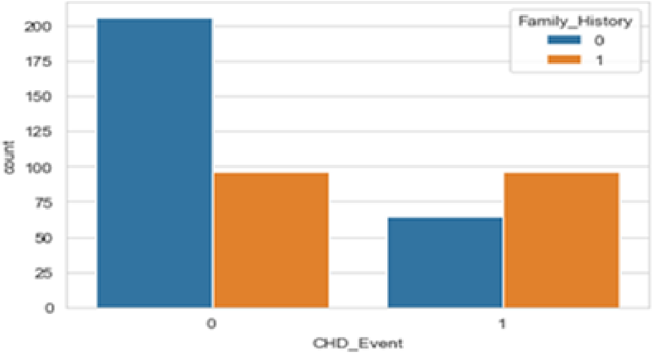


Fig. 1. Family History vs CHD_Event.

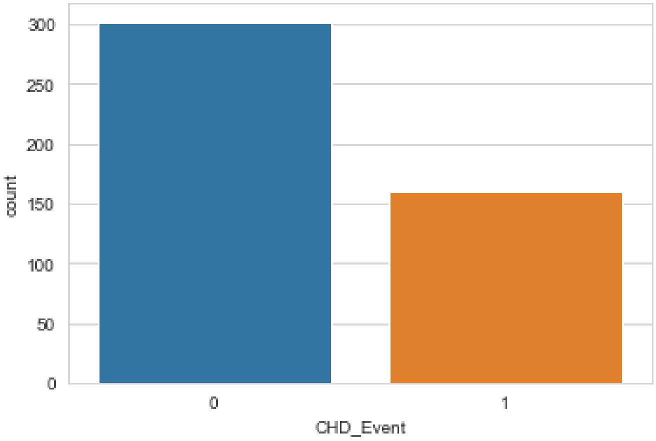


Fig. 2. Frequency of Positive and Negative CHD Events in the Dataset.

D. Feature Selection

The dataset includes many of the widely known risk factors or features that cause CHD, but we aim to rank which features are the most relevant to the target in predicting CHD and which features are the least relevant. This allows it to be further analyzed and interpreted by experts in the domain and could be used as the basis for gathering more or different data.

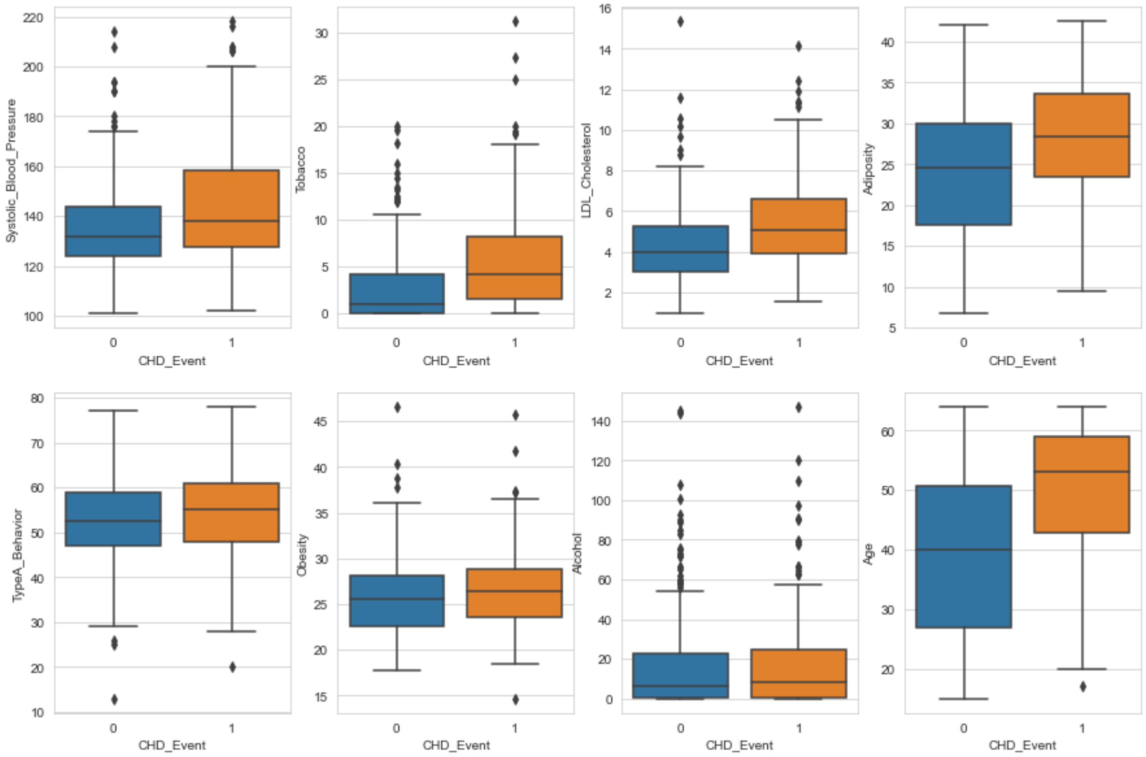


Fig. 3. Boxplot Representing the Distribution of Data Features with respect to CHD_Event.

1) ANOVA method: ANOVA stands for änalysis of variance änd it is a parametric statistical hypothesis test that determines whether the means of two or more samples of data (usually three or more) come from the same distribution or not [31]. An F-statistic, also known as an F-test, is a class of statistical tests that use a statistical test like ANOVA to measure the ratio between variance values. An ANOVA f-test is a type of F-statistic that uses the ANOVA method [32].

We used an implementation of the ANOVA f-test function from the scikit-learn machine learning library, which suits our classification problem task. Table III shows the scores of ANOVA f-test ranking of the features, i.e., the scores calculated for each input feature and the target variable (CHD Event) in descending order, the higher the score, the more important the feature.

2) Feature Importance: Statistical methods calculate the score of the feature ranking with relation to the features and the target variable, however, the importance and ranking of the features might be different when working together to predict the target variable. However, using machine learning methods for feature ranking provides insight into prediction models and which features are the most important and least important to the models when making a prediction.

Another method we used to compute a set of feature importance scores for our dataset is the permutation feature importance. The concept of Permutation Feature Importance was first introduced by Breiman [33] and applied to a random forest model. Permutation Feature Importance works by randomly changing the values of each feature column, one column at a time. It then evaluates the model.

TABLE III. ANOVA F-TEST RANKING

Feature	Score
Age	74.330
Tobacco	45.400
Family History	36.861
LDL Cholesteroal	34.197
Adiposity	31.756
Systolic Blood Pressure	17.674
Type A Behavior	4.948
Obesity	4.655
Alcohol	1.806

We have used the permutation_importance function from scikit-learn library with Random Forest as the fit model. We chose accuracy as the standard metric to measure performance in this context because this a classification problem. The ranking of the features is shown in Fig. 4.

The different methods of feature ranking showed that several features are most common as the most important features such as Age, Tobacco, LDL Cholesterol, Systolic Blood Pressure, Adiposity, and Family History. However, in our machine learning modeling experiments we used almost all the features and dropped Obesity as it has a high correlation with Adiposity, we used Pearson's correlation to calculate the

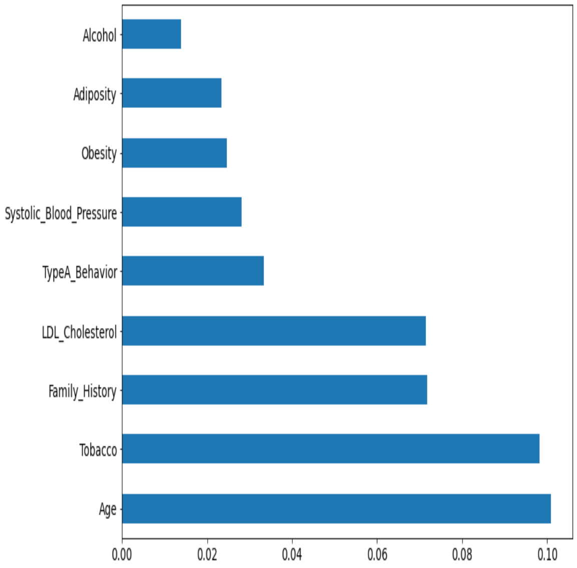


Fig. 4. Feature Importance using Random Forest.

score of the correlation between features and to drop the most correlated ones.

3) Pearson’s Correlation Coefficient: Pearson’s Correlation method is used for finding the feature correlation to remove the redundant features. As shown in Fig. 5, the correlation is represented as a number between -1 and 1, which indicates the extent to which two variables are related.

A correlation coefficient higher than 0.7 is considered strong and therefore one of the features can be dropped because this will affect the prediction accuracy. Given this, the obesity feature was dropped from the training dataset.

IV. Experimental Framework

In this study, we used scikit-learn Python library to conduct the experiments, the selected models, namely, Logistic Regression, SVM, KNN, and MLP Neural Network were applied on the dataset described in the last section, with 462 samples, 8 predictors (dropping Obesity feature) and 1 target variable.

Ten-fold stratified cross-validations were used for model training and testing. The stratified folds were used in these experiments because the dataset is imbalanced with evident imbalanced class distributions, as discussed earlier.

The machine learning techniques utilized for the prediction of CHD are set up as follows:

A. Logistic Regression

The logistic regression technique uses the logistic function [34] to model a binary dependent variable. The technique is capable of solving linear separable classes as well as complex problems. We have used the GridSearchCV function from scikit-learn library to find the optimal parameters, and the logistic regression was configured with ‘lpfgs’ solver, ‘l2’ penalty, and we set up ‘C’ to 0.25.

B. SVC

Based on the Support Vector Machine algorithm [35], this technique separates data points that belong to different classes with a decision boundary (hyperplane). The main parameter here is the kernel, it maps the observations into some feature space. With the help of GrisdSearchCV function, the kernel was set up to ’rbf’, ‘C’ to 10 and we configured ‘gamma’ to auto.

C. KNN

KNN does not try to build an internal model, the computations are not done until the classification time. KNN stores instances of the training data in the features space and the class of an instance are determined based on the distance measure from its neighbors, Therefore, the most important parameter is the number of neighbors to be considered, here we set it up to 17. We used the elbow method [36] to calculate the optimal number of neighbors. And we set up ‘minkowski’ as the metric for the distance measure.

D. MLP Neural Networks

MLP is a neural network that consists of more than two layers with a number of neurons in each layer. We set up 3 layers with 50, 20, and 10 neurons consecutively. The activation function was set up to ‘tanh’ and the learning rate to ‘0.01’.

E. Classification of Evaluation Metrics

Accuracy, Precision (Positive predictive value), Recall (Sensitivity or True Positive rate), F1 score, in addition to Specificity (True Negative rate) were mainly used to evaluate the performance of the prediction models.

To calculate these, the confusion matrix is used to describe the performance of each predicted negative and positive class, as in Fig. 6.

Where:

- TN: is the total number of patients who correctly identified that they have no CHD.
- FN: is the total number of patients incorrectly identified that they have no CHD.
- TP: is the total number of patients correctly identified that they have CHD.
- FP: is the total number of patients incorrectly identified that they have CHD.

In imbalanced datasets precision, recall, and F1 score are often more important measures than accuracy. In this problem, even the accurate prediction of the CHD patients matters the most, i.e., high precision or high recall [37]. However, there is always a precision/recall trade-off, and in CHD prediction, high recall might be even preferred over high precision.

The aforementioned metrics can be calculated from the confusion matrix as follows:

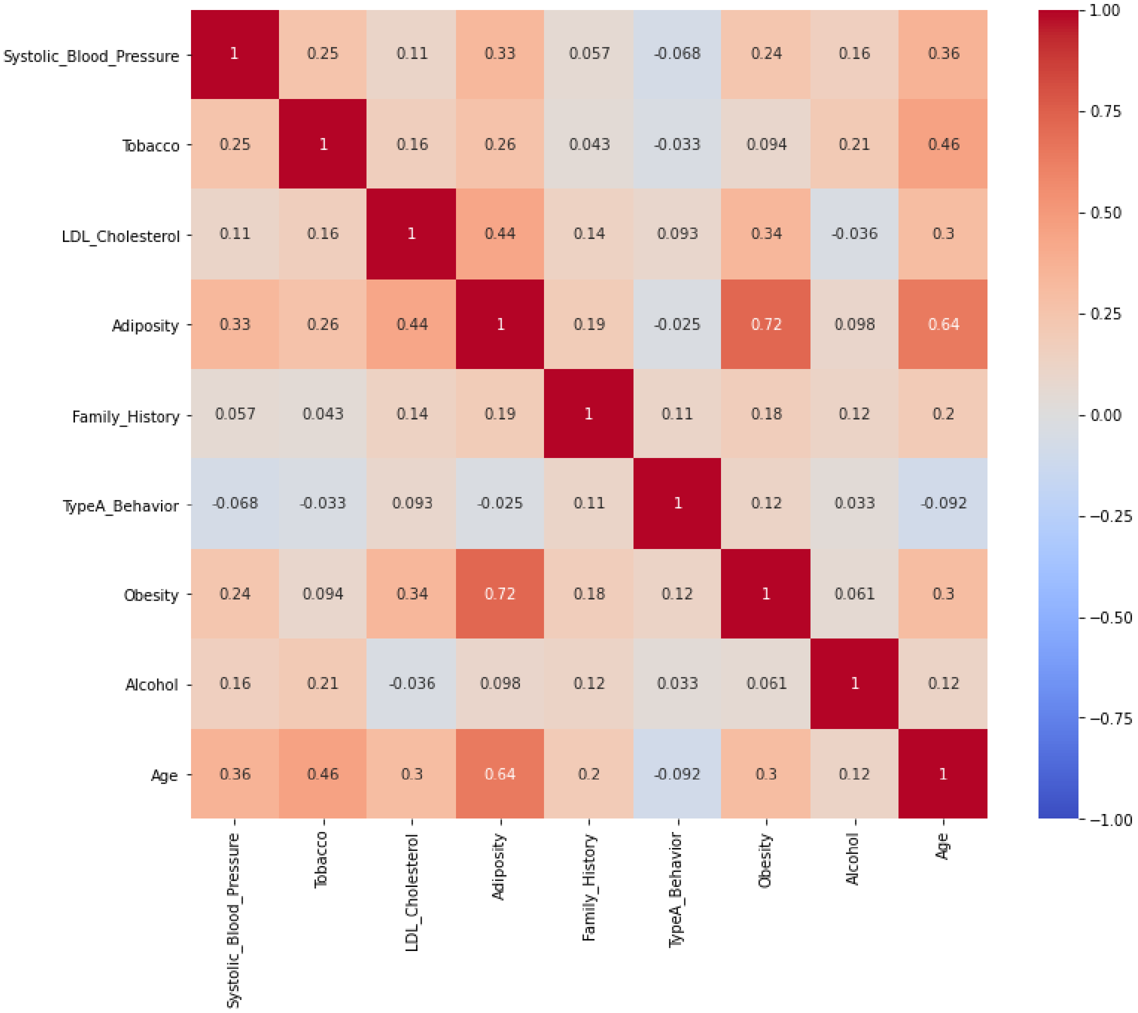


Fig. 5. Pearson's Correlation Matrix.

True			
0	TN	FP	
1	FN	TP	
	0	1	Predicted

Fig. 6. Confusion Matrix.

- Precision (Positive predictive value)

$$TP = \frac{TP}{TP + FP} \quad (1)$$

- Recall (Sensitivity or True positive rate)

$$recall = \frac{TP}{TP + FN} \quad (2)$$

- Specificity (True negative rate)

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

- F-score

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (4)$$

- Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

F. Data Oversampling

In general, many medical datasets show signs of imbalanced class distribution which greatly hampers the detection of rare events, as most classification methods implicitly assume an equal occurrence of classes [38, 39]. The dataset of this study is a very small size with a total of 462 instances, distributed into 302 negative CHD instances and 160 positive CHD instances.

In an imbalanced class distribution problem, the sample size is critical in evaluating the classification model, and the high error rate caused by the imbalanced class distribution decreases as the size of the training dataset increases [39]. Furthermore, in the dataset at hand, many variables are not linearly correlated, not linearly separable, and are complexly

TABLE IV. PREDICTION RESULTS - MEAN OF 10 FOLD CROSS-VALIDATION

Classifier	Accuracy	F1 Score	Precision	Recall	Specificity
SVM	**0.738***	0.550	0.679	0.463	0.884
MLP Neural Network	0.734	0.553	0.661	0.475	0.871
KNN	0.732	0.504	**0.7***	0.394	**0.911***
Logistic Regression	0.727	**0.563***	0.633	**0.506***	0.844

overlapping, as discussed earlier. The authors in [40] have stated that not the imbalanced distribution of classes is the main problem in the classification with imbalanced data classification, but many characteristics, among them “the presence of small disjuncts, the lack of density in the training data, the overlapping between classes, the identification of noisy data”.

Several techniques have been introduced and used for handling the imbalanced datasets and improving the prediction [41-43]. In this study we have used Synthetic Minority Oversampling TEchnique (SMOTE) for short, this technique was first described by [41]. In particular, we have used the K-means SMOTE method [44, 45].

1) K-means SMOTE: SMOTE with the K-means method improves classification by producing minority class samples in safe areas of the input space. The method reduces noise while effectively addressing imbalances within and within samples. We used the KMeans SMOTE class from the imbalanced-learn Python library.

V. RESULTS

The mean accuracy results of applying the 10-fold stratified cross-validation on the dataset were obtained show that SVM slightly outperformed MLP neural network classifier, KNN, and Logistic Regression. The results were 73.8%, 73.4%, 73.2% and 72.7% respectively. However, as discussed before, accuracy alone is not the main concern here because this is an imbalanced dataset, the distribution of the labeled target class is unequal. The mean scores of applying 10-fold stratified cross-validation with F1 score, Precision, Recall, Specificity of the 4 classifiers are summarized in Table IV. The results show improvements in the accuracy compared to previous research results on the same dataset.

A. Results on the Oversampled Dataset

The dataset now has 604 samples with 302 instances with negative CHD events and 302 instances with positive CHD events. The mean scores of applying the classifiers 10-fold stratified cross-validation on the dataset after oversampling are summarized in Table 5. The classifiers are ordered based on the accuracy we also calculated the Matthews Correlation Coefficient (MCC) score, to highlight which classifier achieved good results in all 4 categories of the confusion matrix.

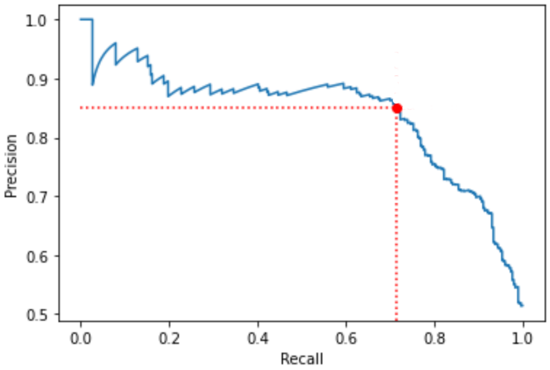


Fig. 7. Precision vs Recall in SVM.

The results show major improvement on the recall score with an average of 32% for all classifiers. From 10 fold cross-validation, MLP neural network has recorded an average 80.3% of Recall score, while KNN has achieved an average 80% of Precision score and an average 85.8% of Specificity score. The overall accuracy for all classifiers has also been improved, SVM has achieved the highest overall accuracy and good results in all scores combined.

Fig. 7 shows the plot of the precision vs recall for SVM classifier, as it can be noticed the precision has dropped at around 78% of recall, so we can even create an SVM model with let’s say over 85% of Precision score with over 71% of recall by tunning the threshold of precision.

VI. DISCUSSION AND CONCLUSION

Our results show that, given sufficient data and proper selected clinical features, machine learning techniques are capable of predicting the occurrence of CHD events with high accuracy. The application of the four machine learning techniques SVM, KNN, MLP neural networks, and logistic regression using the South African Heart Disease dataset with the selected features reported roughly as high as 74% accuracy. While this shows a noticeable improvement in the prediction performance compared to previous researches on the same data, the main issue in this study was to resolve the imbalanced classification problem in the dataset and achieve even higher scores in Precision and Recall in particular, in addition to improving the overall prediction accuracy. Such a problem in the dataset was tackled by applying K-means SMOTE oversampling techniques, and as a result, the prediction performance of all prediction models has significantly enhanced, with an average improvement of 32% on the Recall score and an average improvement of 11% on the Precision score.

Among the four prediction techniques applied on the oversampled dataset in this study, SVM has obtained the best results in all the four confusion matrix categories, marginally followed by KNN, MLP neural network, and logistic regression respectively. However, from the usability standpoint, one might choose to use KNN as a prediction model for this problem, since KNN has obtained an 80% Precision score and around 86% Specificity score. Whereas MLP neural network has reported an 80% Recall score. Recent trends in prediction and classification are going toward using a combination of

TABLE V. Prediction Results on the Oversampled Dataset- Mean of 10 Fold Cross Validation

Classifier	MCC	Accuracy	F1 Score	Precision	Recall	Specificity
SVM	**+0.561**	**0.781***	0.780	0.784	0.776	0.785
MLP Neural Network	+0.549	0.774	0.781	0.760	**0.803***	0.745
KNN	+ 0.553	0.776	0.767	**0.8***	0.737	**0.858***
Logistic Regression	+0.538	0.769	0.775	0.755	0.795	0.742

prediction techniques for more accurate and more reliable outcomes. That is, it is a good idea in practice to use the SVM, KNN, MLP neural network classification models together for predicting the positive and negative CHD cases, as they strengthen and complement each other.

Our feature selection techniques have showed and confirmed that clinical features and risk factors such as, Tobacco, LDL Cholesterol, Systolic Blood Pressure, Adiposity, and Family History are among the most important features that help in the early detection and the prediction of the presence of CHD events from medical records. Medical practitioners can take advantage of the exploratory data analysis conducted on the dataset to show correlations and relationships between patients' data.

The success of machine learning relies heavily on the richness of the data representing the phenomenon under consideration. Even though the selected dataset has the most widely known features and risk factors for predicting CHD, with a rather rich set of features, more data and more variables can potentially help improve the prediction results. If additional external datasets with the same features from different regions had been available, we would have used it as a validation of our findings.

As future work, we are planning to apply our machine learning approach on other datasets of cardiovascular diseases, cancer, and infectious diseases. We are also preparing to deploy the models as a web service and integrate it in a web application to allow medical practitioners assess its usability in the real world.

References

[1] AIHW. Coronary Heart Disease. 2020 [cited 2021 20 Feb]; Available from: https://www.aihw.gov.au/reports/australias-health/coronary-heart-disease

[2] Hajar, R., *Risk factors for coronary artery disease: historical perspectives.*, Heart views: the official journal of the Gulf Heart Association, 2017. 18(3): p. 109

[3] S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, *”Can machine-learning improve cardiovascular risk prediction using routine clinical data?”*, PloS one, 2017. 12(4): p. e0174944.

[4] P.K. Sahoo, S.K. Mohapatra, and S.L. Wu, *”Analyzing healthcare big data with prediction for future health condition”*, IEEE Access, 2016. 4: p. 9786-9799.

[5] D. Chicco, and G. Jurman, *”Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone”*, BMC medical informatics and decision making, 2020. 20(1): p. 16.

[6] S. Blecker, S.D. Katz, L.I. Howrwitz, and G. Kuperman, *”Comparison of approaches for heart failure case identification from electronic health record data”*, JAMA cardiology, 2016. 1(9): p. 1014-1020.

[7] M. Fatima, M. Pasha, *”Survey of machine learning algorithms for disease diagnostic”*, Journal of Intelligent Learning Systems and Applications, 2017. 9(01): p. 1.

[8] R. Sujatha, and A. Nithya, *”A Survey of Health Care Prediction Using Data Mining”*, International Journal of Innovative Research in Science, Engineering and Technology, 2016. 5(8): p. 14538.

[9] D. Kinge, and S.K. Gaikwad, *”Survey on data mining techniques for disease prediction”*, International Research Journal of Engineering and Technology (IRJET), 2018. 5(01): p. 630-636.

[10] K.G. Dinesh, K. Arumugaraj, K.D. Santosh, and V Mareeswari, *”Prediction of cardiovascular disease using machine learning algorithms”*, in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). 2018. IEEE.

[11] J.J. Beunza, E. Puertas, E. Garcia-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M.F. Landecho *”Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)”*, Journal of biomedical informatics, 2019. 97: p. 103257.

[12] Panicker, S.,*”Use of Machine Learning Techniques in Healthcare: A Brief Review of Cardiovascular Disease Classification”*. 2020.

[13] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, N.S. Naik, *”Prediction of coronary heart disease using supervised machine learning algorithms”*, TENCON 2019-2019 IEEE Region 10 Conference (TENCON). 2019. IEEE.

[14] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H. Rudd and M. van der Schaar, *”Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants”*, PloS one, 2019. 14(5): p. e0213653.

[15] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P.M. Kebria, F. Khozemeh, et al., *”Machine learning-based coronary artery disease diagnosis: A comprehensive review”*, Computers in biology and medicine, 2019. 111: p. 103346.

[16] A.H. Gonsalves, F. Thabtah, R.M.A Mohammad, G. Singh, *”Prediction of coronary heart disease using machine learning: an experimental analysis”*, in Proceedings of the 2019 3rd International Conference on Deep Learning Technologies. 2019.

[17] S.I. Ayon, M.M. Islam, and M.R. Hossain, *”Coronary artery heart disease prediction: a comparative study of computational intelligence techniques”*, IETE Journal of Research, 2020: p. 1-20.

[18] K.H. Miao, and J.H. Miao, *”Coronary heart disease diagnosis using deep neural networks”*, Int. J. Adv. Comput. Sci. Appl., 2018. 9(10): p. 1-8.

[19] G.D. Flora, and M.K. Nayak, *”A brief review of cardiovascular diseases, associated risk factors and current treatment regimes”*, Current pharmaceutical design, 2019. 25(38): p. 4063-4084.

[20] KEEL Data set. *”South African heart dataset”*, [cited 2021 18, February]; Available from: https://sci2s.ugr.es/keel/dataset.php?cod=184.

[21] R. Nakanishi, D. Dey, F. Commandeur, P. Slomka, J. Betancur, H. Gransar, C. Dailing, K. Osawa, D. Berman, and M. Budoff, *”Machine learning in predicting coronary heart disease and cardiovascular disease events: results from the multi-ethnic study of atherosclerosis (mesa)”*, Journal of the American College of Cardiology, 2018. 71(11S): p. A1483-A1483.

[22] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, *”A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease”*, 2017 IEEE symposium on computers and communications (ISCC). 2017. IEEE.

[23] S. Safdar, S. Zafar, N Zafar, N.F. Khan, *”Machine learning based decision support systems (DSS) for heart disease diagnosis: a review”*, Artificial Intelligence Review, 2018. 50(4): p. 597-623.

[24] K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T Dudley, and P.P. Sengupta, *”Machine learning in cardiovascular medicine: are we there yet?”*, Heart, 2018. 104(14): p. 1156-1164.

[25] C. Sowmiya, P. Sumitra. *”Analytical study of heart disease diagnosis using classification techniques”*, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). 2017. IEEE.

[26] E.B. Maini, B. Venkateswarlu, and A. Gupta. *”Applying machine learning algorithms to develop a universal cardiovascular disease prediction system”*, International Conference on Intelligent Data Communication Technologies and Internet of Things. 2018. Springer.

[27] K.H. Miao, J.H. Miao, and G.J. Miao, *”Diagnosing coronary heart disease using ensemble machine learning”*, Int J Adv Comput Sci Appl (IJACSA), 2016.

[28] *Heart Disease Data Set*, [cited 2021 20, March]; Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[29] *Statlog (Heart) Data Set*, [cited 2021 20, March]; Available from: http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

[30] A. Baccouche, B. Garcia-Zapirain, C Castillo Olea, and A. Elmaghraby, *”Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico”*. Information, 2020. 11(4): p. 207.

[31] B.G. Tabachnick, and L.S. Fidell, *”Experimental designs using ANOVA”*, 2007: Thomson/Brooks/Cole Belmont, CA.

[32] A. Bathke, *”The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data”*, Journal of Statistical Planning and Inference, 2004. 126(2): p. 413-422.

[33] L. Breiman, *”Random forests”*, Machine learning, 2001. 45(1): p. 5-32.

[34] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *”Logistic regression”*, 2002: Springer.

[35] T. Joachims, *Svmlight: Support vector machine*, http://svmlight. joachims. org/, University of Dortmund, 1999. 19(4).

[36] P. Dangeti, *”Statistics for machine learning”*, 2017: Packt Publishing Ltd.

[37] A. Géron, *”Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems”*, 2019: O’Reilly Media.

[38] A. Ali, S.M. Shamsuddin, and A.L. Ralescu, *”Classification with class imbalance problem”*, Int. J. Advance Soft Compu. Appl, 2013. 5(3).

[39] M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, and G.D. Tourassi, ”Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance”, Neural networks, 2008. 21(2-3): p. 427-436.

[40] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, *”An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”*, Information sciences, 2013. 250: p. 113-141.

[41] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, *”SMOTE: synthetic minority over-sampling technique”*, Journal of artificial intelligence research, 2002. 16: p. 321-357.

[42] W. Elazmeh, N. Japkowicz, and S. Matwin, *”Evaluating misclassifications in imbalanced data”*, in European Conference on Machine Learning. 2006. Springer.

[43] M.A. Maloof, *”Learning when data sets are imbalanced and when costs are unequal and unknown”*, ICML-2003 workshop on learning from imbalanced data sets II. 2003.

[44] G. Douzas, F. Bacao, and F. Last, *”Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE”*, Information Sciences, 2018. 465: p. 1-20.

[45] F. Last, G. Douzas, and F. Bacao, *”Oversampling for Imbalanced Learning Based on K-Means and SMOTE”*,.arXiv preprint arXiv:1711.00837, 2017.

[46] N. Kausar, A. Abdullah, B.B. Samir S., Palaniappan, B.S. AlGhamdi, and N. Dey, ”Ensemble cluster algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease”, Journal of Medical Imaging and Health Informatics, vol.6, number 1, February 2016, p.78-87(10).

[47] N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, and N. Dey,*”Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients.”*, Applications of intelligent optimization in biology and medicine, pp. 217-231. Springer, Cham, 2016.

A Survey of Specification-based Intrusion Detection Techniques for Cyber-Physical Systems

Livinus Obiora Nweke
Department of Information Security and Communication Technology,
Norwegain University of Science and Technology (NTNU),
Gjøvik, Norway

***Abstract*—Cyber-physical systems (CPS) integrate computation and communication capabilities to monitor and control physical systems. Even though this integration improves the performance of the overall system and facilitates the application of CPS in several domains, it also introduces security challenges. Over the years, intrusion detection systems (IDS) have been deployed as one of the security controls for addressing these security challenges. Traditionally, there are three main approaches to IDS, namely: anomaly detection, misuse detection and specification-based detection. However, due to the unique attributes of CPS, the traditional IDS need to be modified or completely replaced before it can be deployed for CPS. In this paper, we present a survey of specification-based intrusion detection techniques for CPS. We classify the existing specification-based intrusion detection techniques in the literature according to the following attributes: specification source, specification extraction, specification modelling, detection mechanism, detector placement and validation strategy. We also discuss the details of each attribute and describe our observations, concerns and future research directions. We argue that reducing the efforts and time needed to extract the system specification of specification-based intrusion detection techniques for CPS and verifying the correctness of the extracted system specification are open issues that must be addressed in the future.**

***Keywords*—Cyber-physical systems; intrusion detection systems; specification-based intrusion detection mechanism; security**

I. Introduction

The recent years have witnessed an increasing growth in the development and deployment of different types of cyber-physical systems (CPS). CPS have shaped every aspect of our lives as their applications span through several domains including electrical power grids, water and wastewater management, oil and gas sector, traffic systems and many other domains. Considering the nature of CPS, security incidents could lead to physical harm to people, destruction of property or environmental disasters. For this reason, the secured operation of CPS is a major concern for all stakeholders.

According to Gartner analysts [1], CPS security incidents are expected to rise in the coming years due to a lack of security focus and spending that are aligned to CPS. They also observe that the liability for CPS security incidents will not only affect the corporate entity but will also lead to a personal liability for 75% of CEOs by 2024. This is a wakeup call for all those charged with the responsibility for the secured operation of CPS and for greater attention to the development and deployment of appropriate security controls for CPS.

One of the security controls for CPS involves the use of intrusion detection systems (IDS). Traditionally, there are three main approaches to IDS, namely: anomaly detection, misuse detection and specification-based detection. However, due to the unique attributes of CPS, the traditional IDS need to be modified or completely replaced before it can be deployed for CPS. A discussion of the techniques and challenges on the use of IDS in CPS have been provided by Han et al. in [2] . Our interest in this paper is to survey the use of specification-based intrusion detection techniques for CPS.

Some works in the literature have conducted surveys related to the use of IDS for CPS [3], [4], [5]. Mitchell and Chen [3] presents a survey of IDS design principles and techniques for CPS. They categorize the existing CPS IDS techniques in the literature, describe their advantages and disadvantages and suggest future research areas. Zarpelão et al. [4] also conducted a survey of IDS in Internet of Things (IoT). They classify the IDS proposed in the literature according to the following attributes: detection method, IDS placement strategy, security threat and validation strategy. A much recent survey related to the use of IDS for CPS is presented by Wu et al. [5]. They conducted a survey of the proposed IDS designs for in-vehicle networks. However, to the best of our knowledge, none of the existing surveys have considered the use of specification-based IDS for CPS.

In this paper, we present a survey of specification-based intrusion detection techniques for CPS. In particular, we classify the existing specification-based intrusion detection techniques in the literature according to the following attributes: specification source, specification extraction, specification modelling, detection mechanism, detector placement and validation strategy. We also discuss the details of each attribute and describe our observations, concerns and future research directions. We argue that reducing the efforts and time needed to extract system specification of specification-based intrusion techniques for CPS and verifying the correctness of the extracted specification are open issues that must be addressed in the future.

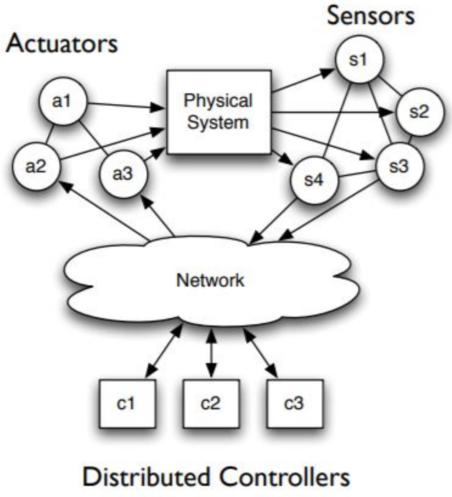
The rest of this paper is organised as follows. Section II presents a discussion on CPS and specification-based intrusion detection, which provides an understanding for describing the suitability of specification-based intrusion detection techniques for CPS. Section III provides a survey of specification-based intrusion detection techniques for CPS and the proposed taxonomy. Section IV describes our observations, concerns and future research directions, which is one of the most relevant contributions of this work. Section V concludes the paper.

II. Background

A. Cyber-Physical Systems

CPS facilitate the integration of computation and communication capabilities to monitor and control physical systems. This enables the accomplishment of time-sensitive functions with different degrees of interaction with the environment, including human interaction [6]. As a result of this, CPS are called time-sensitive systems which makes timing a central theme in their design and implementation. CPS are also referred to as safety-critical systems because the failure of the system due to faults or other external influences, could endanger the lives of humans operating the physical systems, those embedded with the CPS (medical devices) or those within the radius of their operation (nuclear plants). The application of CPS span through several domains and they include modern vehicles, medical devices, industrial systems, etc., all with different standards, requirements, communication technologies, and time constraints.

The general architecture of CPS is depicted in Figure 1. CPS as showed in the diagram, typically have a physical system that is being monitored and controlled, a set of sensors that report the state of the physical system, a set of actuators that are used by the controllers to maintain the system in the desired state, and a set of controllers (or a controller) that monitors and controls the physical system using the sensors and actuators, and via a communication channel [7]. The interaction between these components of CPS is known to be vulnerable to cyber attacks. For example, a power station located north of the city of Kiev, Ukraine, suffered a cyber attack which blacked out a portion of the Ukrainian capital equivalent to a fifth of its total power capacity [8]. This calls for increased efforts towards addressing the security issues of CPS.



[7]

Fig. 1. The General Architecture of CPS.

There are several security challenges in the operation of CPS. These challenges can be attributed to the unique features of CPS which makes the traditional security solutions ineffective in addressing the security challenges of CPS. For example, CPS have time constraints because the physical processes are generally time-aware and deadline sensitive [9]. Also, the complexities in the analysis and design of security solutions for CPS are further exacerbated by the need to understand and address the upstream and downstream dependencies of the component systems [6]. Therefore, the current information technology (IT) security controls would have to be modified significantly or to be completely replaced because they are unable to address the security challenges of CPS.

One of the security solutions for CPS involves the use of intrusion detection systems (IDS). There are three main approaches to IDS, namely: anomaly detection, which relies on comparing current behaviour with the pre-established normal behaviour to detect an intrusion; misuse detection, which use intrusion signatures to detect an intrusion; and specification-based detection, which depends on the monitoring of the specified system behaviour to detect an intrusion [10]. A review of the existing intrusion detection techniques for CPS has been provided by Mitchell and Chen in [3] and Han et al. in [2] discuss the techniques and challenges of intrusion detection in CPS. We are interested in the use of specification-based IDS for CPS in this paper. The following subsections provide an in-depth discussion on specification-based IDS and its suitability for CPS, so as to motivate our survey of the existing specification-based intrusion detection techniques for CPS.

B. Specification-based Intrusion Detection

The notion of specification-based intrusion detection was first introduced by Ko et al. in [10]. It leverages the specification of a system, which describes the expected behaviour of the system. Any deviation of the system operations from the defined correct behaviour is flagged as a security violation. In general, the specification-based intrusion detection process involves the use of a specification source to extract the expected behaviour of a system, which in turn is modelled. A detection mechanism is then applied to the modelled specification for monitoring the system behaviour for any deviation. Figure 2 provides a diagrammatic illustration of the specification-based intrusion detection process.

Specification-based intrusion detection has shown to be a better approach to IDS than anomaly detection and misuse detection [11]. Even though anomaly detection is able to detect novel attacks, it suffers from a high rate of false alarm because unseen legitimate system behaviours are classified as anomalies. Misuse detection, on the other hand, does not generate false alarms but it is unable to detect novel attacks. Hence, specification-based appears to be the mean between misuse detection and anomaly detection because it combines the advantages of both approaches. Its false positive rate is similar to misuse detection as it does not generate false alarms when unusual system behaviours are discovered. Similar to anomaly detection, specification-based intrusion detection is able to detect novel attacks because it detects attacks as deviations from the defined correct system behaviours.

The use of specification-based intrusion detection spans through several domains. Initially, it was intended for execution

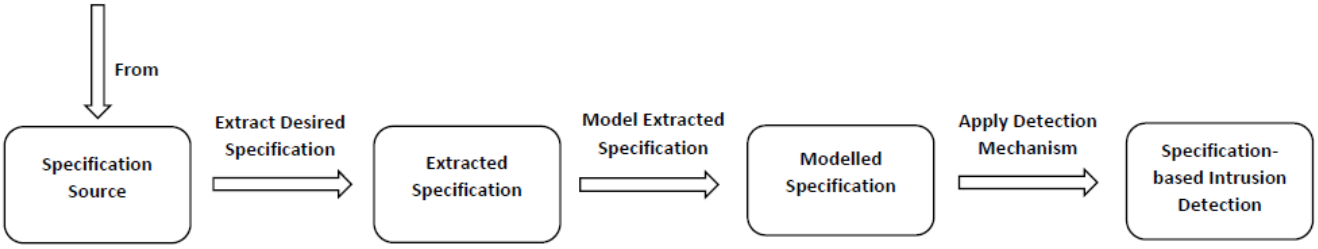

From
Specification Source
Extract Desired Specification
Extracted Specification
Model Extracted Specification
Modelled Specification
Apply Detection Mechanism
Specification-based Intrusion Detection

Fig. 2. Specification-based IDS Process.

monitoring of security-critical programs in distributed system [10]. However, it has been applied to routing protocols such AODV [12], [13], [14] or OSLR [15], DNP3 protocol [16], [17], [18] Voice over IP [19], [20], [21] and other areas of CPS as discussed in section III. A practical experience in the use of specification-based intrusion is presented by Uppuluri and Sekar in [11]. In this work, the experiments conducted show that specification-based intrusion detection is able to detect 80% of the attacks without knowledge about the attacks or the attacker behaviour. They observe that the combination of specification-based intrusion detection with some misuse specification increases the detection ability to 100% with 0% false positive rates.

Specification-based intrusion detection can also be combined with anomaly detection. This is the method adopted by Sekar et al. in [22] and Stakhanova et al. in [23]. Sekar et al. [22] use state-machine specification in combination with information about statistics that need to be maintained to detect anomalies. They evaluate effectiveness of the approach is using 1999 Lincoln Labs intrusion detection evaluation data and the results show that the proposed intrusion detection approach detects all of the probing and denial of service attacks with a low rate false alarm. In the case of Stakhanova et al. [23] the proposed technique facilitates the automatic development of the normal and abnormal behaviour specification in a form of variable-length pattern classified using anomaly-based method. They assess the proposed technique via simulations using publicly available synthetic data and the results show that the approach can detect unknown anomalous behaviour and known anomalous behaviour with a low rate false alarm.

C. Suitability of Specification-based Intrusion Detection for Cyber-Physical Systems

There are several characteristics of CPS that makes specification-based intrusion detection the most suitable type of intrusion detection approach for CPS. One of these characteristics is the laws of physics that govern the physical systems in CPS. The IDS deployed in CPS are expected to monitor physical processes for intrusion. These physical processes are governed by the laws of physics, which makes certain behaviours of the physical systems more likely to be seen than others [3]. Thus, specification-based intrusion detection technique can be used to define these behaviours and to monitor the physical systems for any deviation from these expected behaviours.

Another feature of CPS environment is that activities are usually automated and time driven in a closed-loop settings [3]. This provides some regularity and predictability in the CPS environment which can be used for monitoring. It is different from the IT environment where activities are user triggered and users' behaviours can be very unpredictable. Consequently, the regularity and predictability of CPS environment can be exploited by specification-based intrusion detection to define the correct behaviours of the system, which is subsequently used to monitor behaviours outside the defined behaviours.

Moreover, the protocols deployed in CPS are well-known and widely used which makes it easy to extract the correct behaviour of the system. As a result of this, it is attractive to use specification-based intrusion technique for CPS. This is because the protocol specifications which are readily available can be used as specification source to extract the expected behaviour of the system. Also, specification modelling and detection mechanism can then be employed to complete the specification-based intrusion detection process.

III. Specification-based Intrusion Detection Techniques for Cyber-Physical Systems

In this section, we present a survey of specification-based intrusion detection techniques for CPS. We observe that a common feature of the specification-based intrusion detection techniques for CPS is as follows: a set of properties, which indicates the correct system behaviour is sourced, extracted and modelled; and then, a detection mechanism is used to monitor for any deviation from the defined system specification. Using this understanding, we classify the existing literature according to the following attributes: specification source, specification extraction, specification modelling, detection mechanism, detector placement and validation strategy. The proposed taxonomy of specification-based intrusion detection techniques for CPS is depicted in Figure 3 and Table I summarizes the existing works on specification-based intrusion detection techniques for CPS.

A. Specification Source

Specification source refers to how the set of properties that indicates the correct system behaviour is obtained. There are three major specification sources of specification-based intrusion detection techniques for CPS, namely: protocol specification, reference model and observed behaviour.

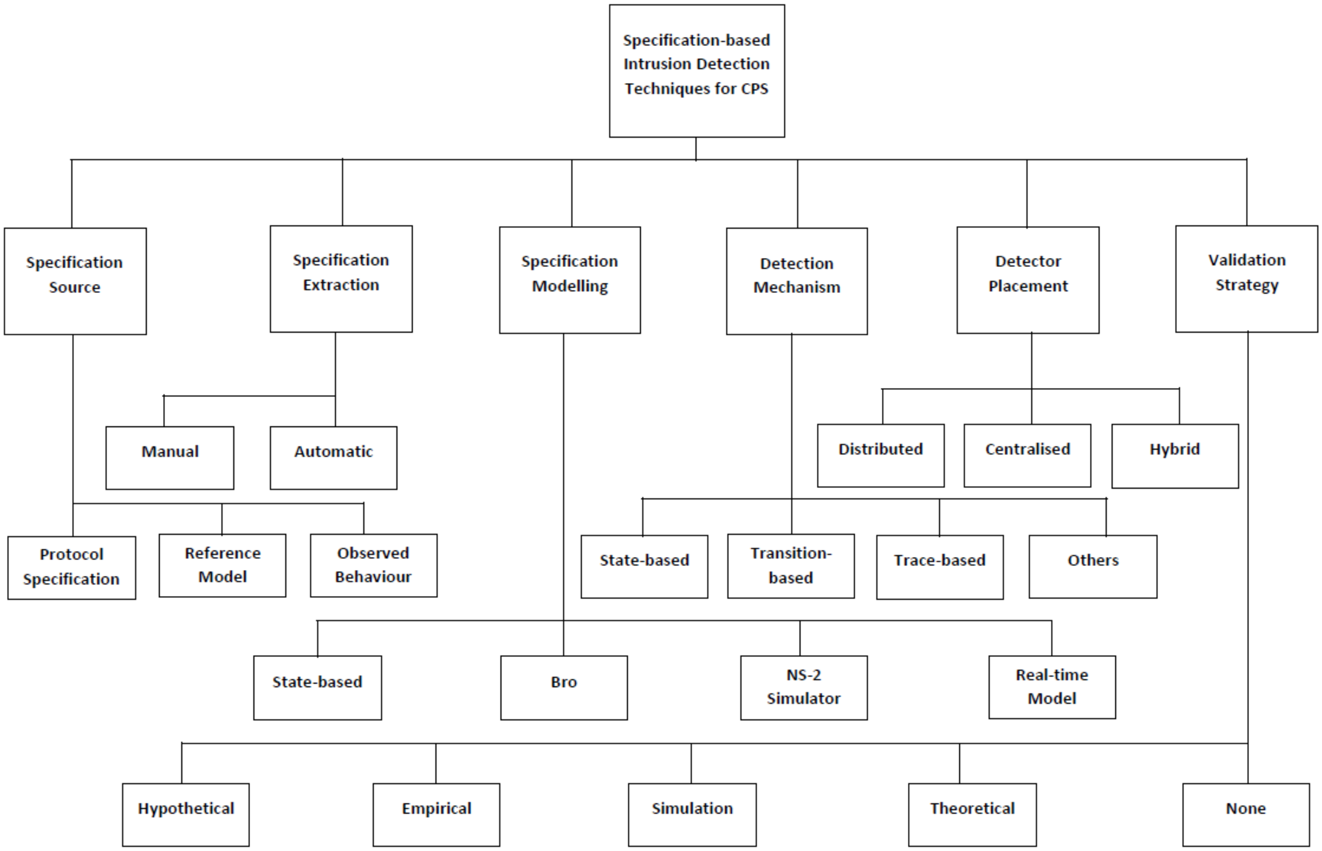


Fig. 3. Taxonomy of Specification-based Intrusion Detection Techniques for CPS.

1) Protocol Specification: Protocol specification is a formal document that defines the expected behaviour of a system. Given the well-defined behaviour of CPS, several protocol specifications have been deployed as the specification source to describe its correct behaviour in many studies. For example, Tseng et al. in [12] utilise the Ad hoc distance Vector (AODV) routing protocol specification as specification source. Other works that have employed AODV routing protocol specification as specification source include Hansson et al. in [13] and Hassan et al. in [14].

Gil et al. in [24] have used IEEE 802.11 protocol and the extensible authentication protocol (EAP) specifications as specification source to define the desired behaviour of wireless local area network. Song et al. in [25] combine both informal protocol specification and other documents from dynamic registration and configuration protocol as specification source. The spanning tree protocol (STP) specification has been leveraged as specification source by Jieke et al. in [26] to describe the expected behaviour for carrier Ethernet network infrastructure. And Tseng et al. in [15] use H.323 protocol specification as specification source.

McParland et al. in [18] deploy protocol guidelines for both ModBus TCP and DNP3 as specification source. They abstract the specific details away of the protocols to focus on the physics models of the system. Unlike them, Lin et al. in [16] and [17] employ only DNP3 as specification source to extract the normal behaviour of the system. Further, Berthier and Sanders in [27] use C12.22 standard protocol specification as specification source to ensure that all violations of the specified security policy of the system will be captured.

The controller area network (CAN) protocol specification has been utilised by Olufowobi et al. in [28] as specification source. Larson et al. in [29] employ the CAN protocol version 2 and the CANOpen application layer draft standard 3.01 as specification source to extract the expected behaviour of electronic control unit of an in-vehicle network. Also, Esquivel-Vargas et al. in [30] exploit the Building Automation and Control Networks (BACnet) protocol as specification source to depict the normal behaviour of each device in the BACnet network.

2) Reference Model: The reference model of the system under consideration has also been employed in several studies as specification source, to obtain the correct system behaviour [31], [32], [33], [34]. Mitchell and Chen in [31] employ the reference model of a modern electrical grid CPS embedding physical components as specification source. They also use the reference model of unmanned air vehicles and reference model of medical CPS in [32] and [33] respectively as specification sources. Also, Sharma et al. in [34] utilize the reference model of unmanned air vehicles CPS as specification source.

3) Observed Behaviour: Observed behaviour of the system under consideration is another method that can be employed as specification source, to define the correct behaviour of a system. It involves the monitoring of a system during its

TABLE I. Summary of the Existing Works on Specification-based Intrusion Detection Techniques for CPS

Reference	Specification Source	Specification Extraction	Specification Modelling	Detection Mechanism	Detector Placement	Validation Strategy
[12]	Protocol Specification	Manual	State-based	State-based	Distributed	Hypothetical
[13]	Protocol Specification	Manual	Stated-based	Other Methods	Distributed	Simulation
[14]	Protocol Specification	Manual	NS-2 Simulator	Other Methods	Distributed	Simulation
[15]	Protocol Specification	Manual	State-based	State-based	Distributed	Simulation
[16]	Protocol Specification	Manual	Bro	State-based	Centralised	Simulation
[17]	Protocol Specification	Manual	Bro	State-based	Centralised	Simulation
[18]	Protocol Specification	Manual	Bro	Transition-based	Centralised	Simulation
[24]	Protocol Specification	Manual	State-based	Transition-based	Centralised	Simulation
[25]	Protocol Specification	Manual	State-based	Trace-based	Centralised	Theoretical
[26]	Protocol Specification	Manual	State-based	Other Methods	Distributed	None
[27]	Protocol Specification	Manual	State-based	State-based	Centralised	Empirical
[28]	Protocol Specification	Automatic	Real-time model	Trace-based	Centralised	Simulation
[29]	Protocol Specification	Manual	Not Specified	Other Methods	Hybrid	Hypothetical
[30]	Protocol Specification	Automatic	Bro	Not Specified	Centralised	Simulation
[31]	Reference Model	Manual	State-based	State-based	Distributed	Simulation
[32]	Reference Model	Manual	State-based	State-based	Distributed	Simulation
[33]	Reference Model	Manual	State-based	State-based	Distributed	Simulation
[34]	Reference Model	Automatic	State-based	State-based	Distributed	Simulation
[35]	Observed Behaviour	Manual	State-based	Transition-based	Centralised	Simulation
[36]	Observed behaviour	Manual	ISML	State-based	Centralised	Simulation

normal operation and then using the knowledge obtained as specification source, to specify the correct behaviour of the system. For example, Pan et al. in [35] use time-synchronized data from synchrophasor and observable events from audit logs as specification source to define the correct behaviour for the cyber-physical environment in electric power system. Also, the specification source utilized by Carcono et al. in [36] is based on monitoring the evolution of the target system states.

B. *Specification Extraction*

Specification extraction is the method that can be deployed to extract the correct behaviour of the system using the specification source. This can either be accomplished manually or automatically.

1) Manual: Most of the specification extraction methods adopt a manual approach for the extraction of the correct system behaviour from the specification source [18], [27], [16], [12], [17], [24], [25], [26], [13], [14], [15], [35], [36], [33], [32], [31]. This method has been shown to be an expensive and very tedious process [27]. As a result of this limitation, there have been attempts in the past few years towards the automatic extraction of the correct system behaviour from specification sources.

2) Automatic: Efforts have been made in recent years to extract the correct system behaviour from the specification source automatically [30], [34], [28]. Esquivel-Vargas et al. in [30] made the first attempt to extract specification automatically. In this work, they implement automated specification extraction in two steps: a subset of the devices capabilities is observed in the network traffic; and based on this observation, an algorithm is used to extract all the devices capabilities from the specification source. Automated specification extraction has also been employed by Sharma et al. in [34] to derive the behaviour rules of IoT device using the operational profile as specification source. And most recently, Olufowobi et al. in [28] exploit real-time schedulability analysis of messages to automate specification extraction.

C. *Specification Modelling*

Specification modelling describes the modelling approach that is adopted to model the specification extracted from a specification source, to describe the correct system behaviour. This subsection presents the different methods that are currently being used for specification modelling.

1) Specification Modelling Using State-based Approach: State-based approach is the most common method for specification modelling. There are several variants of the state-based approach currently in use. The standard state machine has been used by Mitchell et al. in [31], [32], [33] for specification modelling where the extracted specification is transformed into state machines. Jeike et al. in [26] employ state machine for specification modelling. In this work, the states of the machine are states of the protocol, and the state transitions are caused by the receptions of BPDUs or expiration of timeouts. Standard state machines have also been deployed by Berthier and Sanders in [27] to capture the expected system behaviour. And Sharma et al. in [34] have converted the extracted specification into state machines for specification modelling.

Another variant of the state-based approach that has been adopted for specification modelling is the finite state machine. Tseng et al. in [12], [15] use finite state machine for specification modelling. They specify the correct AODV routing protocol behaviour using the finite state machine in [12] and describe the valid routing behaviour of a network node based on Optimised Link State Routing (OLSR) protocol in [15]. The extended finite state machine has been applied by Hansson et al. in [13] and by Song et al. in [25] for specification modelling. Gill et al. in [24] utilise state transition model to describe the extracted specification and state transition patterns are employed by Pan et al. in [35] for specification modelling. Also, a sector specific state modelling language referred to as Industrial State Modelling Language (ISML) has been employed in [36] for specification modelling.

2) Specification Modelling Using Bro: Another tool that can be used for specification modelling is the open source Bro network security monitor (now known as Zeek) [37]. McParland et al. in [18] use Bro scripts for specification modelling but the specific details of the communication protocols and technologies are removed to concentrate on the physics models of the devices being investigated. Similarly, the security specifications of DNP3 protocol have been modelled as a parser and integrated into Bro by Lin et al. in [16], [17]. And Esquivel-

Vargas et al. in [30] use Bro for specification modelling of the automatically extracted correct system behaviour.

3) Specification Modelling Using NS-2 Simulator: NS-2 stands for Network Simulator Version 2 and it is an open-source event-driven simulator for modelling the dynamic nature of communication networks [38]. Hassan et al. in [14] employ NS-2 simulator for specification modelling. In this work, the extracted specification for the runtime behaviour of AODV protocol is implemented using the NS-2 simulator, which allows the detection of any violations from the correct system behaviour.

4) Specification Modelling Using Real-time Model: A recent work by Olufowobi et al. in [28] has proposed the use of a real-time model for specification modelling. In this work, CAN traces that describe the normal behaviour of the network is used to extract real-time parameters as the features which represent the desired specification. Then, the real-time model of the CAN is deployed to specify the expected system behaviour and to flag the violations of the model as indications of a compromised network.

D. Detection Mechanism

Detection mechanism refers to the method that can be adopted to ascertain if there is any deviation from the expected system behaviour. Such deviations are flagged as malicious and since it only relies on the defined specification, this approach is able to detect zero-day attacks. We classify the detection mechanism based on the taxonomy suggested in [39] and the recent developments in the field.

1) State-based Detection Mechanism: Most of the detection mechanism deployed by specification-based intrusion detection techniques for CPS are based on states. The desired state of the system is defined using the specification source that have been extracted and modelled. The goal of the detection mechanism is to detect any deviation from the desired state. For example, Tseng et al. in [12], [15] use a finite state machine for detecting incorrect route request and route reply messages of the AODV routing protocol. They employ predefined finite state machine constraints in [12] which are based on the sourced, extracted, and modelled correct specification; any deviation from these constraints are flagged as malicious. And in [15] the detection mechanism involves checking whether the network node violates the constraints based on the finite state machine.

Mitchell et al. have also used state-based method as detection mechanism in [31], [32], [33]. In these works, they transform the behaviour rules into state machines, which are then used to monitor the system for deviations from the specified system behaviour. Similarly, Berthier and Sanders in [27] use a state machine module to keep track of the state of each device for which traffic is capture, to ensure that stateful constraints are not violated. And Sharma et al. in [34] transforms behaviour rules into a C-language state machine labelled with safe and unsafe states; against which normal and malicious behaviours of the IoT device can be statistically described.

The DNP3 analyser used by Lin et al. in [17], [16] as detection mechanism is based on states. They observed that the DNP3 analyser can maintain states from the parsed network packets and using this states, the incoming packets can be corrected and analysed to ensure there are no violations. Also, Carcano et al. in [36] propose the concept of critical state analysis and state proximity as detection mechanism. They argue that the critical states of a CPS are well documented and that by monitoring the evolution of the physical process states and keeping track of when the CPS enter into a critical state, it is possible to detect attack patterns (known or unknown) likely to drive the CPS into a critical state.

2) Transition-based Detection Mechanism: The detection of malicious behaviour can also be accomplished by monitoring the transition between states. McParland et al. in [18] describes operational protocols using pre- and post- conditions of physical state transitions and any transition that does not lead to a good state is flagged as a potential failure or attack on sensors or actuators. The detection mechanism presented by Gill et al. in [24] used as state transition modelling. In this work, the detection mechanism is achieved by monitoring any anomalous transition in the observed state transition model. And temporal-state transitions are used by Pan et al. in [35] as the detection mechanism. The method adopted in this work involves the monitoring of transition from state to state to detect patterns that are likely to interrupt the protection scheme.

3) Trace-based Mechanism: Trace-based method is another detection mechanism that can be used to monitor deviations from the specified system behaviour. For example, Song et al. in [25] define a set of valid network traces that indicates all finite traces of a network accepted by the specification. They then employ these traces as detection mechanism by monitoring for any trace violating the specification. Olufowobi et al. in [28] use CAN traces as detection mechanism. The CAN traces used in this work depict the normal behaviour of the network, and the detection mechanism involves checking to see if the CAN traces conform with the specification.

4) Other Methods: There are other methods that can be used as detection mechanism which is neither state-based, transition-based, nor trace-based. One of such method presented by Jieke and Redol in [26] and Hansson in [13] combines the attributes of state-based method and transition-based method. In this works, the detection mechanism depends not only on the state of the system but also on the transition between states. Another method is the detection mechanism that has been described by Hassan et al. in [14] which involves identifying misuses to routing messages based on the derived specification. Also, Larson et al. in [29] employ a detection mechanism that checks protocol violations by monitoring the ECU object directory for illegal modifications.

E. Detector Placement

In the use of specification-based intrusion detection techniques for CPS, the detector placement strategies can be distributed, centralised, or hybrid (a combination of both the distributed and the centralised detector placement). This type of classification has been used by Zarpelão et al. in [4] to describe the possible placement strategies for intrusion detection systems in Internet of Things. Hence, this subsection presents the three types of detector placement methods that can be used in specification-based intrusion detection techniques for CPS.

1) Distributed Detector Placement: The use of distributed detector placement is the most desired placement strategy of specification-based intrusion detection for CPS. This is because of the distributed nature of CPS and the need for an IDS which allows every device to be monitored by other devices and ensures there is no single point of failure. For this reason, the use of distributed detector placement in specification-based intrusion detection for CPS has been proposed in [33], [12], [26], [15], [31], [32], [26], [13], [14], [15], [34].

2) Centralised Detector Placement: Centralised detector placement refers to the detector placement where the detector is located at a centralised component, for example, a dedicated host or a network router. This is the strategy employed by most of the survey works [18], [27], [16], [30], [28], [35], [17], [24], [25], [36]. Even though the use of centralised detector placement creates a single point of failure, the ease of its implementation is responsible for its prevalence.

3) Hybrid Detector Placement: Hybrid detector placement is an approach that attempts to combine the benefits of distributed detector placement and centralised detector placement. This approach has been deployed by Larson et al. in [29]. They observe that placing a detector in the network device would make the use of specification-based intrusion detection impossible in CAN environment because it cannot ascertain if the source of the message is allowed to transmit, or if the destination is allowed to receive. As of result of this, they combine distributed detector placement and centralised detector placement to remedy the limitation of centralised detector placement.

F. Validation Strategy

This subsection aims to present the validation strategy that have been employed in the use of specification-based intrusion detection techniques for CPS. Validation is the process of ascertaining if the developed model behaves with acceptable accuracy according to the objectives of the study [40]. To classify the existing validation strategy in the use of specification-based intrusion techniques for CPS, we use the classification of validation methods proposed by Verendel in [41] namely: hypothetical, empirical, simulation, theoretical, and none.

1) Hypothetical: Here, hypothetical examples are used for the validation of the proposed techniques. This is the approach that is adopted by Larson et al. in [29] and Tseng et al. in [12]. Larson et al. [29] use hypothetical example where assumption about the capability of an attacker is made. They apply this to a conceptual network model connecting two networks through a common Gateway to validate their proposed specification-based intrusion detection technique. Similarly, Tseng et al. [12] employ a hypothetical example of how the network monitors trace AODV packets based on the AODV scenario they described to validate their proposed specification-based intrusion detection method.

2) Empirical: Empirical methods have also been used as validation strategy of specification-based intrusion detection for CPS. Berthier and Sanders [27] utilise empirical evaluation and observe that the objectives of such verification are two folds: verifying that the implementation is correct, and measuring the performance of the implementation under various conditions.

3) Simulation: Simulation is the most popular validation strategy used by the existing literature surveyed in this paper [18], [33], [16], [30], [28], [34], [35], [17], [24], [13], [14], [15], [36], [31], [32]. For example, McParland et al. [18] in the validation of their proposed approach use a collection of Modbus master and salve simulation tools and DNP3 simulation tools. Mitchell and Chen in [31], [32], [33] use Monte Carlo simulation for the validation of their proposed techniques. Lin et al. [16] employ a test-bed to simulate SCADA-specific attack scenarios in the bid to validate their proposed method. And Carcano et al. [36] simulated a prototype of the approach they described as validation strategy.

Moreover, the validation strategy using simulation may require the development of a specialised tool. This is the approach adopted by Hansson et al. in [13]. They develop a simulation environment in C++ called Aquarius, which is then deployed for the validation of their proposed technique. Well-known tools have also been deployed as validation strategy. For example, Hassan et al. [14] use NS-2 network simulator, Tseng et al. [15] use GloMoSim simulation platform and Gill et al. [24] use a custom Snort-Wireless preprocessor.

Pan et al. [35] implemented a test-bed to simulate an electric transmission system which they used to validate the specification-based intrusion detection framework proposed in their work. Esquivel-Vargas [30] also simulate a prototype which is implemented using third-party software tools and custom scripts to validate their proposed approach. The validation of the method proposed by Sharma [34] is accomplished using UAV-CPS simulated in MATLAB. And the validation strategy employed by Olufowobi [28] involves the simulation of their proposed method with real CAN logs collected from two passenger cars and on an open-source CAN dataset collected from real-world scenarios.

4) Theoretical: Theoretical methods as validation strategies involve the use of formal or precise theoretical arguments to support the obtained results. This method has been used by Song et al. in [25] to validate specification-based intrusion detection technique. They utilise ACL2 theorem prover [42] and the enforcement of security requirements is defined and proved as theorems in ACL2.

5) None: None refers to the papers where no validation methods are deployed. Among the works we surveyed in this paper, only the work by Jieke and Pan [26] falls in this category.

IV. Observations, Concerns and Future Research Directions

We observe from this study that specification-based intrusion detection technique has been applied in several domains of CPS. For example, it has been employed by McParland [18] for monitoring security of networked control systems. Mitchell and Chen [33] have proposed the use of specification-based intrusion detection for safety critical medical cyber physical systems. Also, specification-based intrusion detection has been proposed for monitoring in-vehicle networks by Larson et al.

[29] and Oluwofobi et al. [28]. Bertheir and Sanders [27] propose the use of specification-based intrusion detection to monitor traffic at the edge of an advanced metering infrastructure. Other applications of specification-based intrusion detection techniques for CPS include SCADA systems [16], [17], [36], building automation systems [30], mobile ad hoc networks [12], [25], [13], [14], [15], IoT devices [34], power system [31], [35], unmanned air vehicles [32] and wireless local area networks [24].

There are several protocols that have been deployed in the operation of CPS. For this reason, it is natural to see that many of the existing literature of specification-based intrusion detection techniques for CPS involve the monitoring of protocols used in CPS. For instance, CAN protocol which is used for in-vehicle networks has been studied by Larson et al. in [29] and Oluwofobi et al. in [28]. Also, specification-based intrusion detection has been proposed for monitoring DNP3 protocol [16], [17], [18]. Other protocols that have been considered by the surveyed papers include C12.22 standard protocol [27], BACnet protocol [30], IEEE 801.11 protocol [24], spanning tree protocol [26], OLSR protocol [15], dynamic auto-configurations protocol [25], and AODV protocol [12], [13], [14].

We also note that only the works by Berthier and Sanders [27], Sharma et al. [34] and Song et al. [25] employed formal modelling for the verification of the specified behaviour. The use of formal modelling is an important aspect of specification-based intrusion detection technique as it enables the verification of the extracted specification. Since specification-based intrusion detection techniques depend on the specified system behaviours, it is imperative that these behaviours represents the correct behaviour of the system and formal modelling provides a tool for such verification. Unfortunately, only a few of the existing literature on specification-based intrusion detection for CPS attempted to verify the derived system specification.

Moreover, we notice that the traditional IDS performance metrics have been used in some of the existing works on specification-based intrusion detection techniques for the evaluation of their proposed technique. Performance metrics are used to measure the performance of IDS. For example, false positive rate has been used in [27], [24], [13], [36], [15] and the combination of false positive rate and false negative rate have been used in [33], [30], [28], [31], [34], [32]. In addition, Lin et al. in [17], [16] use throughput as the performance metric.

One of the biggest concerns in the use of specification-based intrusion detection techniques for CPS is the efforts and time required for specification extraction. As we have already observed, most of the existing works in the use of specification-based intrusion detection techniques for CPS employed the manual approach for specification extraction. This method is prone to errors and could jeopardize the intrusion detection ability of the specification-based IDS. Although efforts have been made by Esquivel-Vargas et al. in [30], Olufowobi et al. in [28], and Sharma et al. in [34] to address the problem through the use of automatic specification extraction, it still remains an open research issue.

Another concern when deploying specification-based intrusion detection techniques for CPS is verifying the correctness of the extracted specification. The ability of a specification-based IDS to detect anomalous behaviour depends on how correct the extracted specification represents the normal system behaviour. One of the ways to verify the correctness of the extracted specification is through the use of formal modelling. Out of the 20 works we surveyed, only Berthier and Sanders [27], Sharma et al. [34] and Song et al. [25] have deployed formal modelling for the verification of the correctness of the extracted specification. Thus, future research works need to consider the best methods for verifying the correctness of the extracted specification so as to encourage the practical application of specification-based IDS for CPS.

V. Conclusion

In this paper, we presented a survey of specification-based intrusion detection techniques for CPS. We selected 20 papers in the literature that proposed the use of specification-based intrusion detection mechanism for CPS. These papers were published between 2005 and 2020. We proposed a taxonomy to classify these papers, which is based on the following attributes: specification source, specification extraction, specification modelling, detection mechanism, detector placement and validation strategy. We observed that to fully realize the potentials of specification-based intrusion detection techniques for CPS, more work needs to be done in the future to reducing the efforts and time required to extract the system specification and to verifying the correctness of the extracted system specification.

References

[1] Gartner, “Gartner predicts 75cyber-physical security incidents by 2024,” Sep. 2020. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2020-09-01-gartner-predicts-75–of-ceos-will-be-personally-liabl

[2] S. Han, M. Xie, H.-H. Chen, and Y. Ling, “Intrusion detection in cyber-physical systems: Techniques and challenges,” *IEEE Systems Journal*, vol. 8, no. 4, pp. 1052–1062, dec 2014.

[3] R. Mitchell and I.-R. Chen, “A survey of intrusion detection techniques for cyber-physical systems,” *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–29, apr 2014.

[4] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, “A survey of intrusion detection in internet of things,” *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, apr 2017.

[5] W. Wu, R. Li, G. Xie, J. An, Y. Bai, J. Zhou, and K. Li, “A survey of intrusion detection for in-vehicle networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 919–933, mar 2020.

[6] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, “Framework for cyber-physical systems: volume 1, overview,” National Institute of Standards and Technology, Tech. Rep., jun 2017.

[7] A. A. Cardenas, S. Amin, and S. Sastry, “Secure control: Towards survivable cyber-physical systems,” in *2008 The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, jun 2008.

[8] A. Greenberg, “’crash override’: The malware that took down a power grid,” 2017. [Online]. Available: https://www.wired.com/story/crash-override-malware/

[9] A. Shrivastava, P. Derler, Y.-S. L. Baboud, K. Stanton, M. Khayatian, H. A. Andrade, M. Weiss, J. Eidson, and S. Chandhoke, “Time in cyber-physical systems,” in *Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*. ACM, oct 2016.

[10] C. Ko, M. Ruschitzka, and K. Levitt, “Execution monitoring of security-critical programs in distributed systems: a specification-based approach,” in *Proceedings. 1997 IEEE Symposium on Security and Privacy (Cat. No.97CB36097)*. IEEE Comput. Soc. Press.

[11] P. Uppuluri and R. Sekar, “Experiences with specification-based intrusion detection,” in *Recent Advances in Intrusion Detection, 4th International Symposium, RAID 2001 Davis, CA, USA, October 10-12, 2001, Proceedings*, ser. Lecture Notes in Computer Science, W. Lee, L. Mé, and A. Wespi, Eds., vol. 2212. Springer, 2001, pp. 172–189.

[12] C. Tseng, P. Balasubramanyam, C. Ko, R. Limprasittiporn, J. Rowe, and K. N. Levitt, “A specification-based intrusion detection system for AODV,” in *Proceedings of the 1st ACM Workshop on Security of ad hoc and Sensor Networks, SASN 2003, Fairfax, Virginia, USA, 2003*, S. Setia and V. Swarup, Eds. ACM, 2003, pp. 125–134.

[13] E. Hansson, J. Grönkvist, K. Persson, and D. Nordquist, “Specification-based intrusion detection combined with cryptography methods for mobile ad hoc networks,” *Command and Control Systems Technical Report*, 2005.

[14] H. M. Hassan, M. Mahmoud, and S. El-Kassas, “Securing the AODV protocol using specification-based intrusion detection,” in *Q2SWinet’06 - Proceedings of the Second ACM Workshop on Q2S and Security for Wireless and Mobile Networks, Terromolinos, Spain, October 2, 2006*, A. Boukerche, H. Chen, and M. S. M. A. Notare, Eds. ACM, 2006, pp. 33–36.

[15] C. H. Tseng, T. Song, P. Balasubramanyam, C. Ko, and K. N. Levitt, “A specification-based intrusion detection model for OLSR,” in *Recent Advances in Intrusion Detection, 8th International Symposium, RAID 2005, Seattle, WA, USA, September 7-9, 2005, Revised Papers*, ser. Lecture Notes in Computer Science, A. Valdes and D. Zamboni, Eds., vol. 3858. Springer, 2005, pp. 330–350.

[16] Z. K. Hui Lin, Adam Slagell and R. K. Iyer, “Using a specification-based intrusion detection system to extend the dnp3 protocol with security functionalities,” *Coordinated Science Laboratory, University of Illinois at Urbana-Champaign*, 2012. [Online]. Available: http://hdl.handle.net/2142/90434

[17] H. Lin, A. Slagell, C. D. Martino, Z. Kalbarczyk, and R. K. Iyer, “Adapting bro into SCADA,” in *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop on - CSIIRW '13*. ACM Press, 2013.

[18] C. McParland, S. Peisert, and A. Scaglione, “Monitoring security of networked control systems: It's the physics,” *IEEE Security & Privacy*, vol. 12, no. 6, pp. 32–39, nov 2014.

[19] P. Truong, D. Nieh, and M. Moh, “Specification-based intrusion detection for h.323-based voice over IP,” in *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005.* IEEE, 2005.

[20] H. Sengar, D. Wijesekera, H. Wang, and S. Jajodia, “VoIP intrusion detection through interacting protocol state machines,” in *International Conference on Dependable Systems and Networks (DSN'06)*. IEEE, 2006.

[21] T. PHIT and K. ABE, “A protocol specification-based intrusion detection system for VoIP and its evaluation,” *IEICE Transactions on Communications*, vol. E91-B, no. 12, pp. 3956–3965, dec 2008.

[22] R. Sekar, A. Gupta, J. Frullo, T. Shanbhag, A. Tiwari, H. Yang, and S. Zhou, “Specification-based anomaly detection,” in *Proceedings of the 9th ACM conference on Computer and communications security - CCS '02*. ACM Press, 2002.

[23] N. Stakhanova, S. Basu, and J. Wong, “On the symbiosis of specification-based and anomaly-based detection,” *Computers & Security*, vol. 29, no. 2, pp. 253–268, mar 2010.

[24] R. Gill, J. Smith, and A. Clark, “Specification-based intrusion detection in wlans.” IEEE, 2006, pp. 141–152.

[25] T. Song, C. Ko, C. H. Tseng, P. Balasubramanyam, A. Chaudhary, and K. N. Levitt, “Formal reasoning about a specification-based intrusion detection for dynamic auto-configuration protocols in ad hoc networks,” in *Formal Aspects in Security and Trust*. Springer Berlin Heidelberg, 2006, pp. 16–33.

[26] P. Jieke, J. Redol, and M. Correia, “SPECIFICATION-BASED INTRUSION DETECTION SYSTEM FOR CARRIER ETHERNET,” in *Proceedings of the Third International Conference on Web Information Systems and Technologies*. SciTePress - Science and and Technology Publications, 2007.

[27] R. Berthier and W. H. Sanders, “Specification-based intrusion detection for advanced metering infrastructures,” in *2011 IEEE 17th Pacific Rim International Symposium on Dependable Computing*. IEEE, 2011, pp. 184–193.

[28] H. Olufowobi, C. Young, J. Zambreno, and G. Bloom, “Saiducant: Specification-based automotive intrusion detection using controller area network (can) timing,” *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 1484–1494, 2020.

[29] U. E. Larson, D. K. Nilsson, and E. Jonsson, “An approach to specification-based attack detection for in-vehicle networks,” in *IEEE Intelligent Vehicles Symposium*. IEEE, 2008, pp. 220–225.

[30] H. Esquivel-Vargas, M. Caselli, and A. Peter, “Automatic deployment of specification-based intrusion detection in the bacnet protocol,” in *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and PrivaCy, Dallas, TX, USA, November 3, 2017*, B. M. Thuraisingham, R. B. Bobba, and A. Rashid, Eds. ACM, 2017, pp. 25–36.

[31] R. Mitchell and I.-R. Chen, “Behavior-rule based intrusion detection systems for safety critical smart grid applications,” *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1254–1263, sep 2013.

[32] ——, “Adaptive intrusion detection of malicious unmanned air vehicles using behavior rule specifications,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 593–604, may 2014.

[33] ——, “Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems,” *IEEE Transactions on Dependable and Secure Computing*, vol. 12, pp. 16–30, 2015.

[34] V. Sharma, I. You, K. Yim, I.-R. Chen, and J.-H. Cho, “BRIoT: Behavior rule specification-based misbehavior detection for IoT-embedded cyber-physical systems,” *IEEE Access*, vol. 7, pp. 118 556–118 580, 2019.

[35] S. Pan, T. H. Morris, and U. Adhikari, “A specification-based intrusion detection framework for cyber-physical environment in electric power system,” *Int. J. Netw. Secur.*, vol. 17, no. 2, pp. 174–188, 2015. [Online]. Available: http://ijns.femto.com.tw/contents/ijns-v17-n2/ijns-2015-v17-n2-p174-188.pdf

[36] A. Carcano, A. Coletta, M. Guglielmi, M. Masera, I. N. Fovino, and A. Trombetta, “A multidimensional critical state analysis for detecting intrusions in SCADA systems,” *IEEE Transactions on Industrial Informatics*, vol. 7, no. 2, pp. 179–186, may 2011.

[37] Zeek (formerly Bro), “An open source network security monitoring tool,” 2020. [Online]. Available: https://www.zeek.org/

[38] T. Issariyakul and E. Hossain, *Introduction to Network Simulator NS2*. Springer US, 2009.

[39] H. Debar, M. Dacier, and A. Wespi, “A revised taxonomy for intrusion-detection systems,” *Ann. des Télécommunications*, vol. 55, no. 7-8, pp. 361–378, 2000.

[40] R. G. Sargent, “Verification, validation, and accreditation: verification, validation, and accreditation of simulation models,” in *Proceedings of the 32nd conference on Winter simulation, WSC 2000, Wyndham Palace Resort & Spa, Orlando, FL, USA, December 10-13, 2000*, P. A. Fishwick, K. Kang, J. A. Joines, and R. R. Barton, Eds. WSC, 2000, pp. 50–59.

[41] V. Verendel, “Quantified security is a weak hypothesis,” in *Proceedings of the 2009 workshop on New security paradigms workshop - NSPW '09*. ACM Press, 2009.

[42] P. C. Dillinger, P. Manolios, D. Vroon, and J. S. Moore, “ACL2s: “the ACL2 sedan”,” *Electronic Notes in Theoretical Computer Science*, vol. 174, no. 2, pp. 3–18, may 2007.

Exploring Factors Associated with Subjective Health of Older-Old using ReLU Deep Neural Network and Machine Learning

Haewon Byeon
Department of Medical Big Data, College of AI Convergence
Inje University, Gimhae 50834, Republic of Korea

***Abstract*—Resolving the health issues of the elderly has emerged as an important task in the current society. This study developed models that could predict the subjective health of the older-old based on gradient boosting machine (GBM), naive Bayes model, classification and regression trees (CART), deep neural network, and random forest by using the health survey data of the elderly and compared their prediction performance (i.e., accuracy, sensitivity, specificity) the models. This study analyzed 851 older-old people (≥75 years old) who resided in the community. This study compared the accuracy, sensitivity, and specificity of the developed models to evaluate their prediction performance. This study conducted 5-fold cross-validation to validate the developed models. The results of this study showed that the deep neural network with an accuracy of 0.75, a sensitivity of 0.73, and a specificity of 0.81 was the model with the best prediction performance. The normalized importance of variables derived from deep neural network analysis showed that depression, subjective stress recognition, the number of accompanying chronic diseases, subjective oral conditions, and the number of days walking more than 30 minutes were major predictors for the subjective health of the older-old. Further studies are needed to identify factors associated with the subjective health of the older-old with considering the age-period-cohort effects.**

Keywords—Gradient boosting machine; classification and regression trees; Naive Bayes model; deep learning; subjective health

I. Introduction

The elderly population is increasing rapidly in South Korea. As of 2018, South Korea has entered the aging society because the elderly population (≥65 years old) exceeded 14% of the total population [1,2]. Moreover, it is expected that South Korea will become a super-aged society, where the elderly population is at least 20% of the total population [1,2]. If this trend continues, it is estimated that the elderly population will be 16.16 million people, 3.7 folds of that in 2005 [1]. In particular, as the average life expectancy is extended, the older-old (75 years or older) is also expected to increase by 8.6 times compared to 2005 [1].

The increased elderly population due to aging caused health problems such as an increase in geriatric diseases and social problems such as an increase in the burden of medical expenses [3,4]. Resolving the health issues of the elderly has emerged as an important task in the current society [5]. In particular, as the baby boomer generation is rapidly turning into an elderly population, South Korean society needs to prepare measures for faces more serious and complex elderly issues to be resolved [6,7,8]. Consequently, the subjective health of the elderly has been studied extensively [6,7,8]. However, these previous studies have limitations. First, most previous studies [9, 10] that identified factors affecting the subjective health status of the elderly usually treated the elderly, ⩾65 years old, as a homogeneous group and developed a model for predicting a subjective health condition without categorizing them into different age groups. Second, it is difficult to generalize and extrapolate the results of previous studies [11] to the whole elderly population in South Korea because they mostly targeted a small elderly group in a specific area. Third, previous studies [12] targeting the elderly in South Korea mainly used tools that were developed in South Korea to measure subjective health or quality of life, and as a result, it is difficult to compare the results of these studies with those published in other countries that do not speak Korean. In particular, as researchers began to recognize issues associated with defining the elderly, ⩾65 years old, as a homogenous group [13], some studies [14,15] subdivided the elderly into the young-old (<75 years old) and the older-old (⩾75 years old).

As South Korea entered the aging society in 2017, it is necessary to explicitly identify the subjective health-related factors of the older-old whose daily living ability has declined. Previous studies [16,17,18,19,20,21,22] evaluated factors related to the subjective health of the older-old and reported that the number of accompanying chronic diseases, depression, socioeconomic status, gender, and social network with friends and relatives as the predictors of subjective health. However, since these previous studies used regression analysis to develop a prediction model, it was efficient to identify individual risk factors but the model is limited in identifying multiple risk factors such as lifestyle and socioeconomic level [23]. In addition, the regression model assumes independence, normality, and homoscedasticity [24]. If the model is developed using data that violate the normality assumption, it may produce biased results [24].

Recently, machine learning has been widely used in various fields as a way to solve the limitations of this regression model. Machine learning can accurately analyze even data that somewhat violate the normality assumption and nonlinear data in the estimation process, which are advantages of machine learning [25]. This study developed models that could predict the subjective health of the older-old based on gradient boosting machine (GBM), naive Bayes model, classification and regression trees (CART), deep neural network, and random forest by using the health survey data of the elderly and compared their prediction performance (i.e., accuracy, sensitivity, specificity) the models.

II. Methods and Materials

A. Study Subjects

This study analyzed the raw data of the 2016 Seoul Panel Study Data (SPS-data). The SPS-data was carried out from June 1 to August 31, 2016, for understanding the status of the welfare vulnerable class living in Seoul and estimating the welfare level of citizens. The population was the households living in Seoul at the time of the 2005 Population and Housing Census, a complete enumeration census. The sample was extracted from the census data using the randomized stratified sampling method for 25 districts in Seoul. This study excluded those who were imprisoned, were admitted to a nursing hospital, or moved to a silver town, and foreigners. As for the survey method, the interviewer visited the target household and conducted a computer-aided personal interview. This study analyzed 851 older-old people who resided in the community and completed the survey.

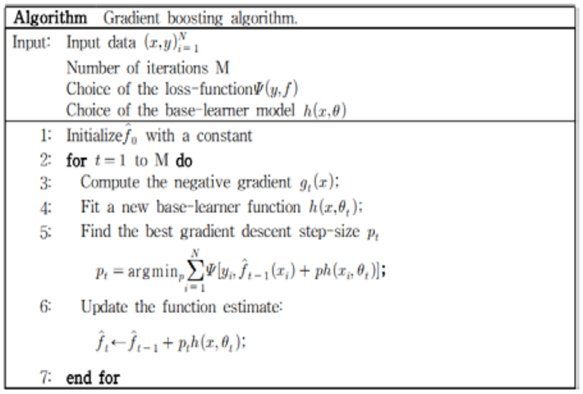
B. Measurements and Definitions of Variables

Subjective health status, outcome variable, was measured on a 5-point Likert scale (very good, good, moderate, bad, and very bad). Explanatory variables included age, sex, smoking (smoker or non-smoker), drinking (once a week or less or more than twice a week), economic activity (yes or no), mean monthly household income (less than KRW 1.5 million, KRW 1.5-3 million, or KRW 3 million or more), educational level (elementary school graduation and below, middle school graduation, high school graduation, or college graduation or above), social activities in the past month (yes or no), living with a spouse (living together, bereavement/separation, or single), disease/accident/addition experience in the past two weeks (yes or no), subjective stress recognition (yes or no), number of accompanying chronic diseases, body mass index, number of days walking for 30 minutes or more (1 or less than day per week or 2 days per week or more), depression, subjective oral conditions, self-perceived diet, frequency of meeting with neighbors (once a month or less or more than twice a month), and meeting frequency with relatives (once a month or less or more than twice a month).

The number of chronic diseases (e.g., diabetes, hypertension, and hyperlipidemia) was classified into none, 1, 2, and 3 or more. The body mass index (BMI) was classified as "underweight" when it was less than 18.5, "normal" when it was 18.5 or more and less than 25, "overweight" when it was 25 or more and less than 30, and "obesity" when it was 30 or more. Subjective stress was defined as a "high-stress group" when the respondents said they felt a lot of stress and a "low-stress group" when they answered that they felt a little stress. Depressive symptoms were classified as "depressed" when feeling depressed for more than two weeks in a row and "not depressed" when not feeling depressed for two weeks or more in a row. Subjective oral conditions were defined as "good", "normal", and "bad". The self-perceived dietary life was classified into "sufficient diet life" when the respondents answered "ate enough quantity of food and a variety of foods in the last year" or "ate enough quantity of food and not various foods" and "insufficient diet life" when they answered "insufficient food from time to time due to economic difficulties" or "frequently insufficient food due to economic difficulties".

C. Development of Prediction Models: GBM

GBM is a machine learning algorithm that creates a strong learner by combining weak learners of a decision tree by using an ensemble technique [26]. This model generalizes the model by generating a model for each step like other boosting methods and optimizing a loss function that can be arbitrarily differentiated. A model is created, even if the accuracy is low, and the errors of the generated model are compensated by the following model. Through this process, the accuracy of the current model becomes better than that of the previous model: a more accurate model (or a stronger learner). The basic principle is to increase accuracy by repeating this process. GBM's algorithm is presented in Fig. 1.



```
Algorithm  Gradient boosting algorithm.
Input:  Input data $(x,y)_{i=1}^{N}$
        Number of iterations M
        Choice of the loss-function$\Psi(y,f)$
        Choice of the base-learner model $h(x,\theta)$
Initialize$\hat{f}_0$ with a constant
for $t=1$ to M do
    Compute the negative gradient $g_t(x)$;
    Fit a new base-learner function $h(x,\theta_t)$;
    Find the best gradient descent step-size $p_t$
        $p_t = \arg\min_p \sum_{i=1}^{N} \Psi[y_i, \hat{f}_{t-1}(x_i) + ph(x_i,\theta_t)]$;
    Update the function estimate:
        $\hat{f}_t \leftarrow \hat{f}_{t-1} + p_t h(x,\theta_t)$;
end for
```

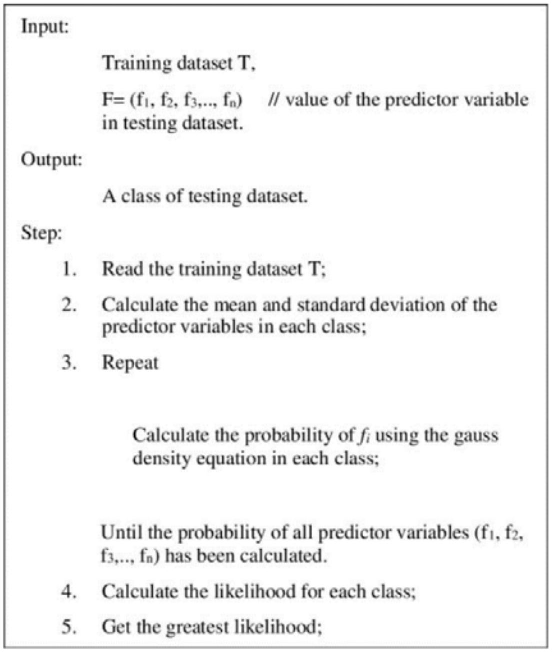
Fig. 1. Algorithm of GBM [26].

D. Naïve Bayes Classification

Naïve Bayes classification is a method of classifying observations into different groups using Bayes theory (Fig. 2)[27]. Bayes theory refers to a way of calculating the posterior probability by using an observation when there is a prior probability.

E. CART

CART is one of the analysis algorithms of the statistical decision classification model, which measures impurity using the Gini Index [28]. It is an algorithm based on a binary split, which generates only two child nodes from a parent node [28]. The advantages of CART are that it is easy to interpret the generated rules and can analyze both continuous and categorical variables. For continuous variables, a separation rule is created in the form of "$X \leqslant C$?" or "$X \geqslant C$?" [28]. For categorical variables, a rule of binary separation is created in the form of "$X \in \{A, B\}$" [28]. In the model of this study, the criterion for separating and merging decision rules for CART was set to 0.05. The number of child nodes was limited to 100, the number of parent nodes was limited to 200, and the number of branch branches was limited to 5.



```
Input:
        Training dataset T,
        F= (f1, f2, f3,.., fn)    // value of the predictor variable
        in testing dataset.
Output:
        A class of testing dataset.
Step:
    1.  Read the training dataset T;
    2.  Calculate the mean and standard deviation of the
        predictor variables in each class;
    3.  Repeat

            Calculate the probability of $f_i$ using the gauss
            density equation in each class;

        Until the probability of all predictor variables (f1, f2,
        f3,.., fn) has been calculated.
    4.  Calculate the likelihood for each class;
    5.  Get the greatest likelihood;
```

Fig. 2. Concept of Naïve Bayes Classification [27].

F. Deep Neural Network

Deep neural network is an algorithm made up of an input layer composed of independent variables, an output layer composed of dependent variables, and two or more hidden layers between the input and output layers [29]. Each layer has independent nodes and a node is connected to other nodes in other layers by weighted neurons (connecting lines) (Fig. 3).

This study used H2O Deep Learning among various deep learning types. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation.

In this study, the number of hidden layers was set to 2, nodes of each layer were set to 10 (total of 20), and epochs (number of repetitions) were set to 5. In this study, a model was developed using the Rectifier Linear Unit (ReLU) designated as a default as the active function of deep learning (Fig. 4).

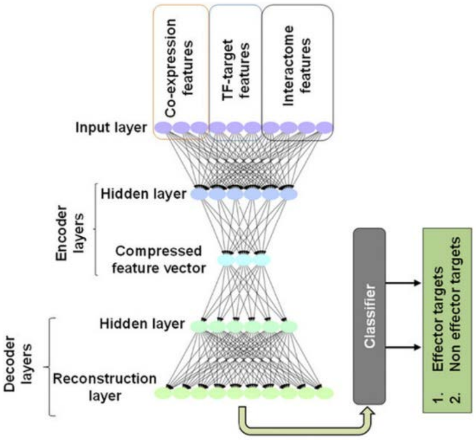


Fig. 3. The Concept of Deep Neural Network [30].

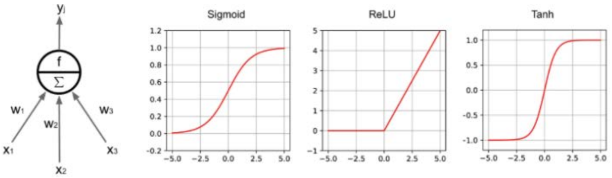


Fig. 4. The Widely used Nonlinear Activation Function of Deep Neural Networks [31].

G. Random Forest

Random Forest is an ensemble method that learns an independent decision tree for each sample after generating a large number of random samples from training data using bootstrap (conduct random restoration sampling of the same sample size from a given data) and decides a final model by synthesizing the results (Fig. 5).

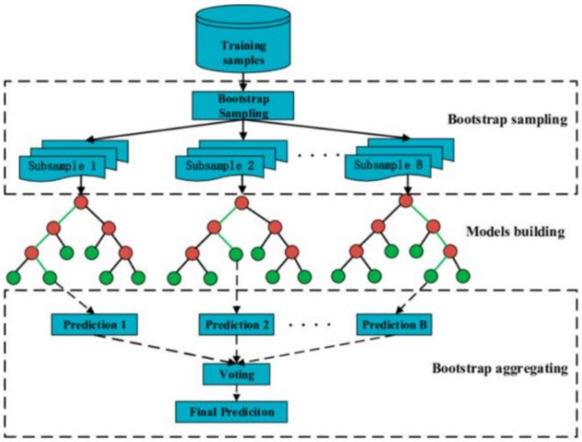


Fig. 5. The Concept of the Random Forest Model [32].

H. Validation of the Prediction Models

This study developed models for predicting the subjective health of the older-old using GBM, naive Bayes model, cart, deep neural network, and random forest. This study compared the accuracy, sensitivity, and specificity of the developed models to evaluate their prediction performance. This study conducted 5-fold cross-validation to validate the developed models.

In the analysis stage, models containing randomness, such as a random forest, were developed while fixing the seed number to "9876543". This study defined the model with the highest accuracy as the best prediction model (best prediction performance) after comparing prediction performance. If two models had the same accuracy, a model with a higher sensitivity was defined as a better prediction model. All analyzes were performed using R version 3.6.2 (Foundation for Statistical Computing, Vienna, Austria).

III. Results

The accuracy of five models (GBM, naive Bayes model, cart, deep neural network, and random forest) for predicting the subjective health of the older-old is presented in Fig. 6. The results of this study showed that the deep neural network with an accuracy of 0.75, a sensitivity of 0.73, and a specificity of 0.81 was the model with the best prediction performance.

The normalized importance of variables derived from deep neural network analysis showed that depression, subjective stress recognition, the number of accompanying chronic diseases, subjective oral conditions, and the number of days walking more than 30 minutes were major predictors for the subjective health of the older-old (Fig. 7). Among them, depression had the highest importance (Fig. 7).

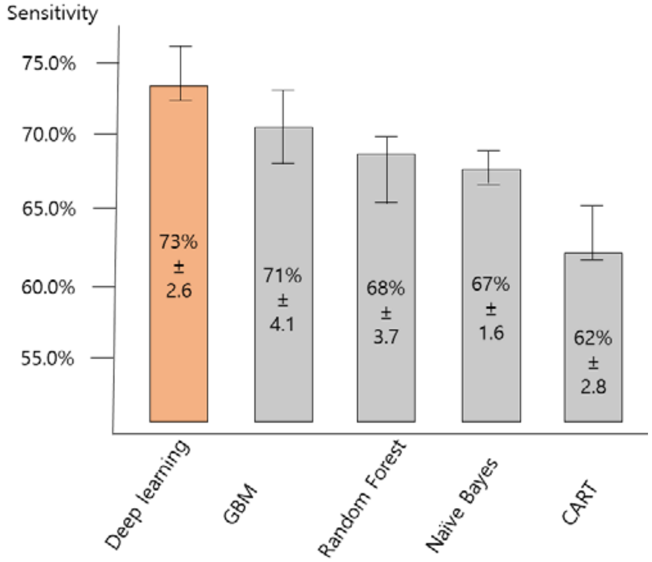


Fig. 6. Comparing the Sensitivity of a Machine Learning Model and a Deep Learning Model for Predicting the Subjective Health of the Older-Old.

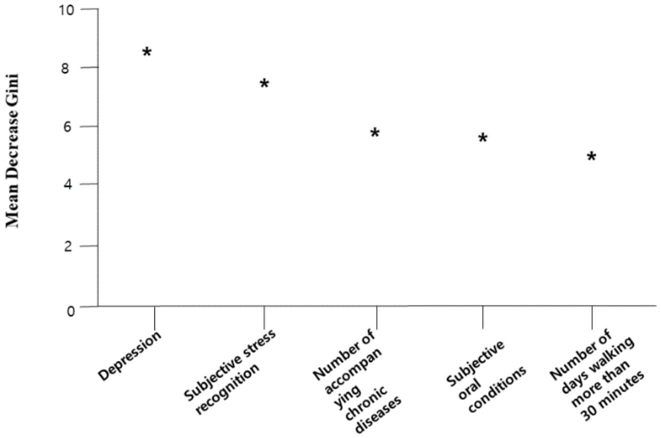


Fig. 7. The Importance of Variables in the Prediction Model for the Subjective Health of the Older-Old (Only the Top 5 Variables are Presented) Conclusions.

IV. Conclusions

This study compared the accuracy of prediction models for the subjective health of the older-old in South Korea and confirmed that the deep neural network-based prediction model had the best prediction performance among GBM, naive Bayes model, cart, deep neural network, and random forest. These results are consistent with the results of da Silva et al. [33], which predicted the health condition of the elderly and showed that the performance of deep learning was superior to that of machine learning methods such as naive Bayes, J48 Decision Tree, and SVM.

Another finding of this study is that depression, subjective stress recognition, number of accompanying chronic diseases, subjective oral conditions, and the number of days of walking more than 30 minutes were identified as the main predictors for the subjective health of the older-old. The result was different from the previous studies [34,35,36] that identified the subjective health of the elderly using regression analysis and reported that social network (e.g., the presence of a spouse), education level, job status, and household income level were the main factors for subjective health. It is believed that these results reflected the interactions of the aging effect, period effect, and cohort effect, indicating the similar socioeconomic level of the older-old in South Korea, where unhealthy elderly people have already passed away. However, Kim et al. [37] evaluated the subjective health of the elderly and reported that the higher the household income level improved the satisfaction of subjective health. Further studies are needed to identify factors associated with the subjective health of the older-old with considering the age-period-cohort effects. Furthermore, additional studies are needed to compare prediction performance such as accuracy, sensitivity, and specificity using data of various diseases in order to prove the effectiveness of deep learning in epidemiological data.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091, NRF-2019S1A5A8034211, NRF-2021S1A5A8062526).

References

[1] Korea National Statistical Office. Population projections for Korea: 2005-2050 based on the 2005 census. Korea National Statistical Office, Daejeon, 2006.

[2] M. H. Oh, Making use of older people’s human resources in an aged Korea. Health and Welfare Policy Forum, vol. 254, pp. 50-66, 2017.

[3] J. S. Kim, and Y. J. Han, The effect of household type on the medical burden of the elderly living in a local government that has entered a super-aged society. The Journal of the Korea Contents Association, vol. 17, no. 7, pp. 610-621, 2017.

[4] M. De Nardi, E. French, and J. B. Jones, Why do the elderly save? the role of medical expenses. Journal of Political Economy, vol. 118, no. 1, pp. 39-75, 2010. doi: 10.5392/JKCA.2017.17.07.610.

[5] T. Suzuki, Health status of older adults living in the community in Japan: recent changes and significance in the super-aged society. Geriatrics & Gerontology International, vol. 18, no. 5, pp. 667-677, 2018. doi: 10.1111/ggi.13266.

[6] H. S. Gweon, The effect of social participation on the life satisfaction of the elderly -focusing on the mediating effects of depression and self-reported health. Korean Journal of Human Ecology, vol. 18, no. 5, pp. 995-1008, 2009. doi: 10.5934/KJHE.2009.18.5.995.

[7] Y. Kwak, and Y. Kim, Quality of life and subjective health status according to handgrip strength in the elderly: a cross-sectional study. Aging and Mental Health, vol. 23, no. 1, pp. 107-112, 2019. doi: 10.1080/13607863.2017.1387766.

[8] S. Y. Sohn, W. Joo, W. J. Kim, S. J, Kim, Y. Youm, H. C. kim, Y. R. Park, and E. Lee, Social network types among older Korean adults: associations with subjective health. Social Science & Medicine, vol. 173, pp. 88-95. doi: 10.1016/j.socscimed.2016.11.042.

[9] M. S. Lee, and H. J. Lim, Factors related to health promoting behaviors of young-old and pld-old elderly in rural areas. Journal of Agricultural Medicine and Community Health, vol. 35, no. 4, pp. 370-382, 2010.

[10] S. H. Kim, Health literacy and functional health status in Korean older adults. Journal of Clinical Nursing, vol. 18, no. 16, pp. 2337-2343, 2009. doi: 10.5393/JAMCH.2010.35.4.370.

[11] K. S. You, H. O. Lee, J. J. Fitzpatrick, S. Kim, E. Marui, J. S. Lee, and P. Cook, Spirituality, depression, living alone, and perceived health among Korean older adults in the community. Archives of Psychiatric Nursing, vol. 23, no. 4, pp. 309-322, 2009. doi: 10.1016/j.apnu.2008.07.003.

[12] J. Y. Kim, S. G. Lee, and S. K. Lee, The relationship between health behaviors, health status, activities of daily living and health-related Quality of Life in the Elderly. Journal of the Korean Gerontological Society, vol. 30, no. 2, pp. 471-484, 2010. doi: 10.21215/kjfp.2018.8.2.220.

[13] N. C. Davis, and D. Friedrich, Knowledge of aging and life satisfaction among older adults. The International Journal of Aging and Human Development, vol. 59, no. 1, pp. 43-61, 2004. doi: 10.2190/U9WD-M79K-9HB8-G9JY.

[14] A. T. F. Beekman, D. M. W. Kriegsman, D. J. H. Deeg, and W. Van Tilburg, The association of physical health and depressive symptoms in the older population: age and sex differences. Social Psychiatry and Psychiatric Epidemiology, vol. 30, no. 1, pp. 32-38, 1995. doi: 10.1007/bf00784432.

[15] R. Choi, and B. D. Hwang, Health care utilization of age group in the elderly on the Korean health panel. The Korean Journal of Health Service Management, vol. 8, no, 3, pp. 49-61, 2014. doi: 10.12811/kshsm.2014.8.3.049.

[16] S. Suh, H. Choi, C. Lee, M. Cha, and I. Jo, Association between knowledge and attitude about aging and life satisfaction among older Koreans. Asian Nursing Research, vol. 6, no. 3, pp. 96-101, 2012. doi: 10.1016/j.anr.2012.07.002.

[17] S. Shiovitz-Ezra, and H. Litwin, Social network type and health-related behaviors: evidence from an American national survey. Social Science & Medicine, vol. 75, no. 5, pp. 901-904, 2012. doi: 10.1016/j.socscimed.2012.04.031.

[18] G. Low, A. E. Molzahn, and D. Schopflocher, Attitudes to aging mediate the relationship between older peoples' subjective health and quality of life in 20 countries. Health and Quality of Life Outcomes, vol. 11, no. 1, pp. 1-10, 2013. doi: 10.1186/1477-7525-11-146.

[19] M. Luo, D. Ding, A. Bauman, J. Negin, and P. Phongsavan, Social engagement pattern, health behaviors and subjective well-being of older adults: an international perspective using WHO-SAGE survey data. BMC Public Health, vol. 20, no. 1, p. 99, 2020. doi: 10.1186/s12889-019-7841-7.

[20] F. Desmyter, and R. De Raedt, The relationship between time perspective and subjective well-being of older adults. Psychologica Belgica, vol. 52, no. 1, pp. 19-38, 2012. doi: 10.5334/pb-52-1-19.

[21] H. Litwin, and K. J. Stoeckel, Social networks and subjective wellbeing among older Europeans: does age make a difference?. Ageing and Society, vol. 33, no. 7, pp. 1263-1281, 2013. doi: 10.1017/S0144686X12000645.

[22] H. Galenkamp, A. W. Braam, M. Huisman, and D. J. Deeg, Somatic multimorbidity and self-rated health in the older population. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, vol. 66, no. 3, pp. 380-386, 2011. doi: 10.1093/geronb/gbr032.

[23] H. Byeon, A laryngeal disorders prediction model based on cluster analysis and regression analysis. Medicine, vol. 98, no. 31, pp. e16686, 2019. doi: 10.1097/md.0000000000016686.

[24] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 36-41, 2021. doi: 10.14569/IJACSA.2021.0120105.

[25] H. Byeon, Predicting the anxiety of patients with Alzheimer's dementia using boosting algorithm and data-level approach. International Journal of Advanced Computer Science and Applications, vol. 12, no, 3, pp. 107-113, 2021. doi: 10.14569/IJACSA.2021.0120313.

[26] J. H. Friedman, Greedy function approximation: a gradient boosting machine. The Annals of Statistics, vol. 29, no. 5, 1189-1232, 2001. doi: 0.1214/aos/1013203451.

[27] M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, Illiteracy classification using K means-Naïve Bayes algorithm. JOIV: International Journal on Informatics Visualization, vol. 2, no. 3, pp. 153-158, 2018. doi: 10.30630/joiv.2.3.129.

[28] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, The CART decision tree for mining data streams. Information Sciences, vol. 266, pp. 1-15, 2014. doi: 10.1016/j.ins.2013.12.060.

[29] B. Belean, M. Streza, S. Crisan, and S. Emerich, Dorsal hand vein pattern analysis and neural networks for biometric authentication. Studies in Informatics and Control, vol. 26, no. 3, pp. 305-314, 2017.doi: 10.24846/v26i3y201706.

[30] B. Mishra, N. Kumar, and M. S. Mukhtar, Systems biology and machine learning in plant–pathogen interactions. Molecular Plant-Microbe Interactions, vol. 32, no. 1, pp. 45-55, 2019. doi: 10.1094/MPMI-08-18-0221-FI.

[31] C. Su, Z. Xu, J. Pathak, and F. Wang, Deep learning in mental health outcome research: a scoping review. Translational Psychiatry, vol. 10, no.1, pp. 1-26, 2020. doi: 10.1038/s41398-020-0780-3.

[32] H. Wang, M. Lei, Y. Chen, M. Li, and L. Zou, Intelligent identification of maceral components of coal based on image segmentation and classification. Applied Sciences, vol. 9, no. 16, pp. 3245, 2019. doi: 10.3390/app9163245.

[33] D. B. da Silva, D. Schmidt, C. A. da Costa, R. da Rosa Righi, and B. Eskofier, DeepSigns: a predictive model based on deep learning for the early detection of patient health deterioration. Expert Systems with Applications, vol. 165, p. 113905, 2021. doi: 10.1016/j.eswa.2020.113905.

[34] S. Y. Im, The association of social network and health status among the elderly in Korea. master's thesis, Public administration The Graduated School of Ewha Womans University, Seoul, 2008.

[35] K. H. Kim, and J. H. Kim, The effects of self-esteem on the relationship between the elderly depression and life satisfaction. Family and Culture, vol. 20, no. 4, pp. 95-116, 2008. doi: 10.21478/family.20.4.200812.004.

[36] T. Lampert, and J. Hoebel, Socioeconomic inequalities in health in later life. Zeitschrift fur Gerontologie und Geriatrie, vol. 52, no. Suppl 1, pp. 91-99, 2018. doi: 10.1007/s00391-018-01487-y.

[37] S. K. Kim, The socioeconomic status and the self-reported health of the aged. International Journal of Welfare for the Aged, vol. 28, 2005.

Propose Vulnerability Metrics to Measure Network Secure using Attack Graph

Zaid. J. Al-Araji[1], Sharifah Sakinah Syed Ahmad[2], Raihana Syahirah Abdullah[3]
Faculty of Information Communication Technology
Universiti Teknikal Malaysia, Melaka
Melaka, Malaysia

***Abstract*—With the increase in using computer networking, the security risk has also increased. To protect the network from attacks, attack graph has been used to analyze the vulnerabilies of the network. However, properly securing networks requires quantifying the level of security offered by these actions, as you cannot enhance what you cannot measure. Security metrics provide a qualitative and quantitative representation of a system's or network's security level. However, using existing security metrics can lead to misleading results. This work proposed three metrics, which is the Number of Vulnerabilities (NV), Mean Vulnerabilities on Path (MVoP), and the Weakest Path (WP). The experiment of this work used two networks to test the metrics. The results show the effect of these metrics on finding the weaknesses of the network that the attacker may use.**

Keywords—Attack graph; security metrics; attack path; path analysis; attack graph uses

I. Introduction

Nowadays, the use of network technology has increased [1], [2]. Nonetheless, since the network is advantageous for people to live and work in, it also carries security problems that must not be overlooked [2]. In many key computer systems and applications, security has been and will remain a major concern. A comprehensive cybersecurity attack will significantly harm the target system as well as the credibility of the businesses or organizations that use it [3]. An attacker may use such attacks to get access to private data, degrade network performance, and eventually take complete control of the targeted system. To detect or protect the network from attacks, the researchers have used many methods [4]. One of these methods in vulnerabilities analysis is an attack graph.

Attack graph has been used for the first time by Philips and Swiler [5], [6]. Since then, researchers have suggested many methods to produce an attack graph. For example, Ammann et al. (2002) proposed the generation method based on monotonicity [7], while Vaibhav Mehta et al. (2006) proposed a ranking attack graph relying on graph neural network (GNN) [8]. Furthermore, Apart from that, Yun Chen et al. (2017) proposed an attack graph generation algorithm relying on a supervised Kohonen neural network [9], while HengLi et al. (2017) introduced a searching forward complete attack graph generation algorithm depending on hypergraph partitioning [10]. Also, Bintao Yuan et al. (2020) introduced the network vulnerability assessment method depending on the graph database and elaborated its efficacy in solving state explosion and other methods [11].

An attack graph may be utilized for many reasons, with positive or negative consequences [13]. Typically, attack graphs are used by researchers to improve the network's security. One of these applications is the computation of network security metrics. Attack graphs may be employed to generate network security metrics to analyze the target network's overall security. These metrics may be utilized to assess the target network's security risk. National Institute of Standards and Technology (NIST) describes security metrics as techniques that gather, analyze, and report pertinent performance-related data to aid decision-making, maximize performance, and increase transparency [12].

There are many security metrics proposed by researchers, such as Shortest Path (SP) Metric, Mean of Paths Length (MoPL) Metric, Number of Paths (NP) Metrics, etc. Some are combined to get new metric with new features and better results, like combining NP and MoPL to get Standard Deviation of Path Lengths (SDPL) Metric proposed by [13].

In[14], the authors divided the attack graph-based security metrics into two types, which are host and network-based metrics. Host-based security measures the level of security of individual hosts in a network. The host-based is divided into two types, which are with and without probability. Meanwhile, the network-based uses the structure of a network to aggregate the network's security property. This type of metrics is classified into two categories, which is path and non-path metrics.

However, using these metrics sometimes gives misleading results, failing to sufficiently account for the number of ways an attacker violates a security policy. In this case, not only the number of the ways but the accuracy is also responsible. For example, the shortest path is not necessary to be the path used by the attacker. It also does not take into account the attack effort connected to the attack paths.

In this paper, three metrics are proposed, which are Number of Vulnerability (NV) Metric, Mean Vulnerabilities on Path (MVoP) Metric, and Weakest Path (WP) Metric to reduce the misleading of the security metrics. The NV and MVoP will view how strong the network is and indicate how much effort the attacker needs to breach network security. On the other hand, it will also measure how much effort is required by the administrator to guard the network from any attacker. Meanwhile, the WP metric views the network's weakest path, which allows the attacker to breach the network policy with minimum effort.

The remainder of the paper is laid out as follows. Section 2 presents the attack graph overview, while Section 3 gives the security metrics related work, Section 4 proposes the security metrics, Section 5 is the experiment performed and results, while Section 6 gives the conclusion.

II. Attack Graph Background

The concept of attack graph has been proposed by Philips and Swiler [15], As shown in Fig. 1. Since then, many researchers have generated attack graphs differently using different methods to improve the attack graph. It is a security model denoting the chains of vulnerabilities, where exploits in the network can be in various forms. The attack graph representation can be organized [16] as a state-oriented, exploit-oriented, or state-exploit-oriented attack graphs [17]. Attack graph generation helps merge low-level vulnerabilities to display all attack paths from source to network goals. By examining the exploited attack paths, security experts should concentrate on patches or configuration bugs that present greater risks. The probabilistic attack graph's risk assessments support such decisions even more [18].

Attack graph generation has three steps which are reachability, attack model, and core building. The attack graph reachability explores the conditions of accessibility in the network, defining whether two given devices could reach one another. The most common representation of network reachability data is a reachability matrix, in which the rows and columns represent the network's hosts. Moreover, each entry indicates the reachability condition between the hosts on the corresponding row and column, respectively [20]. Various connections between the hosts may be represented by a reachability matrix, including transport, network, physical, and application-level connections. Its spatial complexity is on the order of the square of the network's number of hosts [20].

The second phase is the attack model. Attack graph modelling deals with the modelling of attack templates, determining attack graph structure, and modeling networks. The attack template modelling comprises the representation of pre- and post-conditions for the vulnerability. It also provides a process by which information in public vulnerability and weakness databases can be extracted from these conditions for particular vulnerabilities [20].

The attack graph structure's determination involves determining which node and edge types could be contained in the attack graph. Network modelling aims to define a suitable representation of network information [6]. The third phase is the attack graph core building, which denotes the main algorithm employed to develop the attack graphs. Many paths will be pruned in this stage during creating the resulting attack graph in this process [20]. From two different viewpoints, an attack graph core building mechanism could be taken into account. One is the method of evaluating the attack paths, and the other is the method of pruning the attack paths [20].

Generating an attack graph may be utilized for various reasons comprising negative or positive impacts. According to [20], attack graph can be used in four prespectives as in Fig. 2.

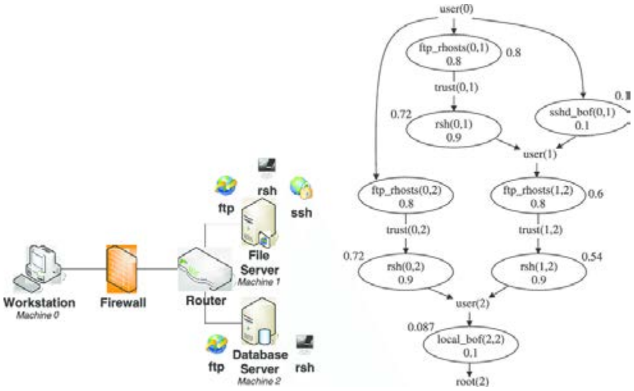


Fig. 1. An Attack Graph Sample [19].

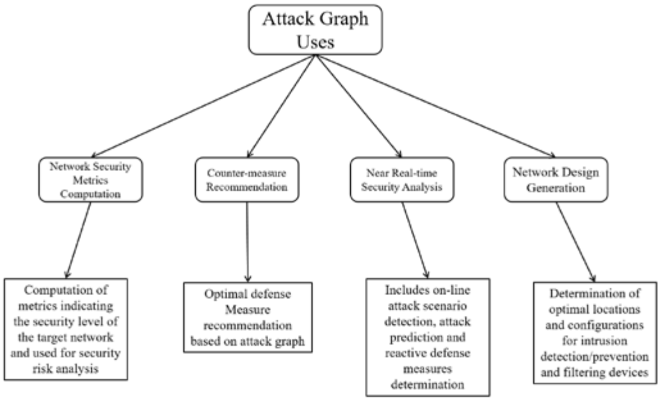

Attack Graph Uses
Network Security Metrics Computation
Counter-measure Recommendation
Near Real-time Security Analysis
Network Design Generation
Computation of metrics indicating the security level of the target network and used for security risk analysis
Optimal defense Measure recommendation based on attack graph
Includes on-line attack scenario detection, attack prediction and reactive defense measures determination
Determination of optimal locations and configurations for intrusion detection/prevention and filtering devices

Fig. 2. Attack Graph uses [20].

Attack graphs can be used for recommending near optimal security defense counter-measures. Optimal counter-measure recommendation can also find practical usage for determining proactive defense recommendations. It can also use the attack graphs generated by accounting for the goal privileges pointing to critical network resources [21].

Another use for the attack graph is network design generation can find practical use in locating the intrusion detection/prevention systems and firewalls optimally in the target network. It can also be used to determine firewall and access control rules, if necessary support for resolving conflicting rules and processing different custom rule formats is provided [17].

Attack graphs can be used for on-line security situational assessment (monitoring) and detecting ongoing attack scenarios by performing highlevel correlation and aggregation of the intrusion alerts and system logs collected throughout the target network. The detected attack scenarios can be used to perform future attack predictions and determine reactive defense measures [17].

Attack graphs also can be used to derive network security metrics used for global security assessment of the target network. These metrics can be used to perform security risk analysis for the target network. Each node (generally indicating a network state) and each edge (generally indicating a vulnerability exploit) on the attack graph can be assigned a probability of occurrence. A node can also be assigned a possible damage value, if the corresponding network state for the node indicates the compromise of some information source for a network host. From these probability and damage values, the cumulative risk values are computed for each network state on the attack graph [17]. In this paper, we will use the attack graph to derive the security metrics.

III. Security Metrics

Metrics, as defined by the NIST, are instruments that gather, analyze, and report applicable performance-related data to aid decision-making and increase performance and transparency. Comprehensive network security and CSA management necessitate the use of security metrics [12].

Security metrics have different categories. Based on Nwokedi C. Idika [22], security metric can be classified into two main classes, which are primary and secondary as in Fig. 3. The primary security metric classes are architectural-based security metrics and performance-based security metrics. The difference in the two classes stems from the type of attributes they measure. Architectural-based metrics measure internal attributes. Performance-based metrics measures external attributes. The secondary security metric classes are security metrics, complexity-based security metrics and time-based security metrics. These metrics can be applied to internal and external attributes of a network. Most of the primary class belong to secondary class as well but not all metrics belong to a primary class but not necessarily a secondary class.

Attack graph-based security metrics is a type of architectural metric [22]. It is a value produced from measuring the internal attributes of a network that affect IT security or operational security. The values are derived from generating an attack graph and subsequently deploying an analysis over the attack graph. This analysis is the measurement that produces the attack graph-based security metric [22].

According to Enoch et al. (2017) attack graph security metrics can be divided to two categories depending on the network reachability which are Host-based and Network-based as in Fig. 4 [23].

The host-level metrics are used to quantify the security level of individual hosts in a network. The host based metrics are classified to two categories which are security metrics with probability and security metrics without probability [24]. The classification had been done because sometimes it is infeasible to find a probability value for an attack, and some analysis and optimisation can be done with or without probability assignments [20].

The network-level metrics are used the structure of a network to aggregate the security property of the network. The network-based security can be classified to two categories which are security metrics path-based and security metrics non path-based. Path based metrics use the reachability information of a network to quantify the security level of the network. While in non path-based metrics, the structure and attributes of a network are not considered; instead, the security of a network is quantified regardless of the network structure [22]. Researchers had proposed many metrics. In this section, some of the previous works will be explained.

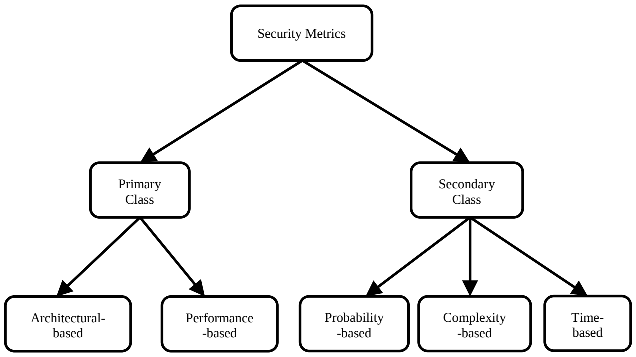


Fig. 3. Security Metrics Classification [22].

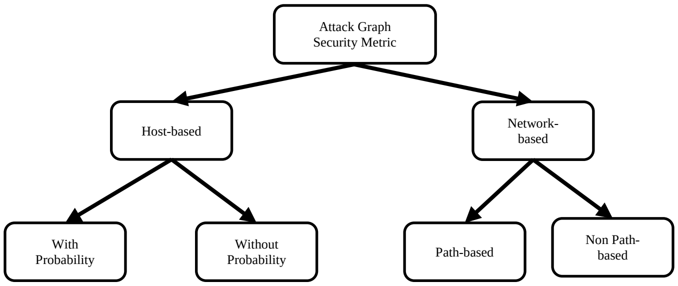


Fig. 4. Attack Graph Security Metrics Classification [23].

Phillips and Swiler (1998) proposed the Shortest Path (SP) Metric. The shortest attack path is the one that takes an attacker from his initial state to his desired goal state using the shortest distance. The length function used to calculate the distance is determined by the security engineer who conducts the attack graph analysis. Nonetheless, this metric does not signify the number of shortest paths in a network. Also, there is no guarantee that the short path is the path used by the attacker [5]. Because of the metric is depending of the length of the path, this metric is path-based metric.

Moreover, Ortalo (1999) suggested the Number of Paths metric (NP). The number of attack paths in a given attack graph is expressed by this security metric. It measures how vulnerable a network is to be attacked. A larger Number of Paths metric suggests a more exposed network. This metric, however, does not account for attack effort, implying that two networks with the same number of attack paths are considered to be of equal security [25]. This metric is path metric because it counts the network paths.

Furthermore, Idika details the Mean of Path Lengths metric (MPL), first introduced by Wei Li (2006) as the Average Path Length metric [26]. It calculates the arithmetic mean of all path lengths to reflect the typical path length. It also estimates how much effort an attacker would impose to break a network security policy. Since an attacker may not have the same perspective of known vulnerabilities as a security engineer, this metric is important. Because of this lack of experience, the attacker may pick a path that is not the shortest. Alternatively, an attacker may choose the other path because the attacker believes the security engineer is using the shortest path analysis. However, this metric cannot be applied alone because it depends on the NP metric [13]. This metric is considered as path metric because it calculate the average of the path length of the network.

These security metrics can be used to retrieve security-relevant data, but they can also produce false results. The Shortest Path and Mean of Path Lengths metrics do not account for all the possible ways an attacker would break a security policy. The attack effort related to the attack paths is not fully accounted for by the Number of Paths metric. To overcome these problems, this work proposes three metrics, which is NV, MVoP, and WP explained in the next section.

IV. Proposed Metrics

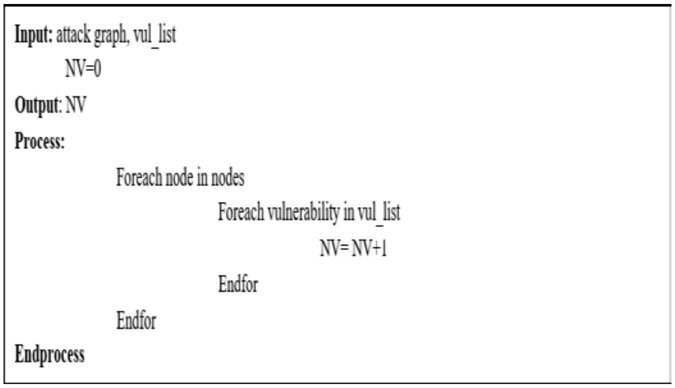
In this section, three attack graph-based security metrics will be proposed: Number of Vulnerabilities Metric (NV), Mean of Vulnerabilities on Path Metric (MVoP), and Weakest Path Metric (WP).

A. Number of Vulnerabilities (NV) Metric

The Number of Vulnerabilities (NV) Metric represents the number of weakness in each node of the network that an attacker can use to cross privilege boundaries in the network. This metric aims to understand the number of disadvantages in the network and allow the administrator to fix it and compare the security of two networks with different size and topology. The formalization of the NV metric is presented in equation 1:

$$NV = \sum V(p_1, p_2, \dots, p_n) \tag{1}$$

Here, V represents the vulnerabilities, p represents the path, and n denotes the number of nodes. Thus, the metrics will calculate the vulnerabilities for each path starting from p_1 to p_n. The pseudocode of NV metric calculation is in Fig. 5. Basically the input in the pseudocode is the attack graph to select all the nodes and the vul_list which represent the vulnerabilities list. The process is so simple is to select a node from all the nodes in the attack graph and calculate the number of vulnerabilities in that node and add them to the counter. This metric is host-based and without probability metric because this metric calculate the vulnerabilities from the host.



```
Input: attack graph, vul_list
      NV=0
Output: NV
Process:
            Foreach node in nodes
                        Foreach vulnerability in vul_list
                                    NV= NV+1
                        Endfor
            Endfor
Endprocess
```

Fig. 5. NV Metric Calculation.

B. Mean Vulnerabilities on Path (MVoP) Metrics

The Mean Vulnerabilities on Path (MVoP) metric represents the average number of the path's vulnerabilities. This metric indicates how much effort an attacker would have to put in to break a network security policy. It also provides a view for the defender to expect the attacker's move. The formalization of the NV metric is presented in equation 2:

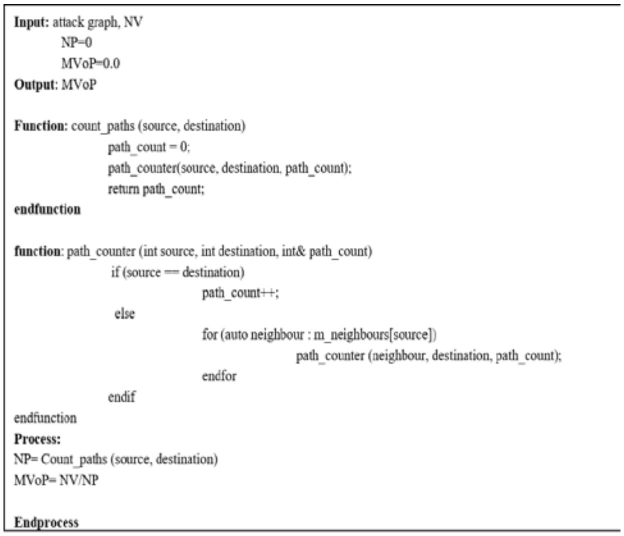
$$MVoP = \frac{NV}{NP} \tag{2}$$

where NV implies the number of vulnerabilities on the network, while NP is the number of the network paths. The formalization of NP is presented in equation 3:

$$NP = |p_1, p_2, \dots ., p_k| \tag{3}$$

where p represents the path and k represents the number of the path. Fig. 6 shows the NP and MVoP metrics calculation. To calculate MVoP metric, we need to calculate NV and NP metrics. The NV metric has been calculated above (see Section 4.1). In this section, we will calculate the NP metric.

The calculation of NP metric is depending on the edges between the nodes, basically each node has many edges with other nodes, to calculate this edges, we used edges list for each node, the we start counting the path from the source to destination using edge list. During the calculation the source change depending on the edge list until the source equal the destination which is mean it is a path. After calculating the numbe of the path, we calculate the MVoP by dividing NV on NP. This metric is considered as path-based and without probability metric because it calculate the vulnerabilies number from the host and calculate the path number from the network.



```
Input: attack graph, NV
      NP=0
      MVoP=0.0
Output: MVoP

Function: count_paths (source, destination)
            path_count = 0;
            path_counter(source, destination, path_count);
            return path_count;
endfunction

function: path_counter (int source, int destination, int& path_count)
            if (source == destination)
                        path_count++;
            else
                        for (auto neighbour : m_neighbours[source])
                                    path_counter (neighbour, destination, path_count);
                        endfor
            endif
endfunction
Process:
NP= Count_paths (source, destination)
MVoP= NV/NP

Endprocess
```

Fig. 6. MVoP Metric Calculation.

C. Weakest Path (WP) Metric

The Weakest Path (WP) Metric is similar to the shortest path metric but in another term, which is the network's strength. The path's strength does not depend on the NV only but also on the score of vulnerability itself. The vulnerability score can use the CVSS that NIST had invented. The formalization of calculating the path score using CVSS is presented in equation 4:

$$V(p) = \frac{\sum_{k=1}^{k=n} CVSS(v(a,b))}{n}, \tag{4}$$

where p represents the path, v represents the vulnerability between node a and b, k denotes the number of the nodes in the path, and n refers to the number of the nodes. Basically, the equation calculates the average score of each vulnerability in the path.

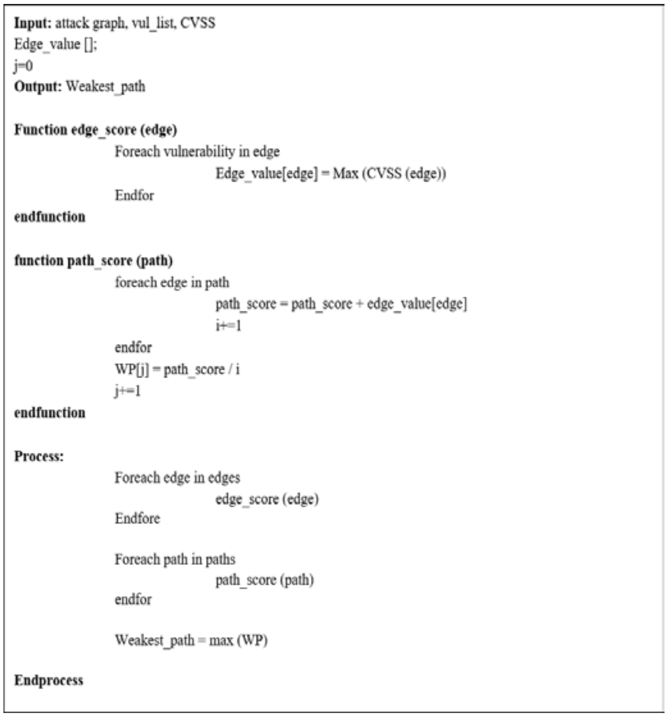
After calculating the path vulnerability score, the score results are compared between all paths in the network to represent the weakest path. The formalization of the WP metric is defined in equation 5:

$$WP = \max\big(V(p_1), V(p_2), \ldots, V(p_k)\big) \qquad (5)$$

where p represents the path, k represents the path number, and $V(p_k)$ represents the summation of the path vulnerabilities score. The calculation of the WP metric in Fig. 7.

The calculation of the weakest path is depending on the value of each edge in the path. To calculate the edge value, we need to find the vulnerabilities score of the edge, so the input will be the attack graph, vulnerabilities list and vulnerability score (CVSS). Then we calculate the edge score by finding the maximum vulnerability score in edge.

After calculating the edge value, the path score will be calculated. Basically, the value of the path will be the average number of edge value in the path. The path score will be saved to compare it with other paths score to find the weakest. The highest path score in the paths will be the weakest path in the network. This metric is considered as path-based metric because it calculate the path score.



```
Input: attack graph, vul_list, CVSS
Edge_value [];
j=0
Output: Weakest_path

Function edge_score (edge)
        Foreach vulnerability in edge
                Edge_value[edge] = Max (CVSS (edge))
        Endfor
endfunction

function path_score (path)
        foreach edge in path
                path_score = path_score + edge_value[edge]
                i+=1
        endfor
        WP[j] = path_score / i
        j+=1
endfunction

Process:
        Foreach edge in edges
                edge_score (edge)
        Endfore

        Foreach path in paths
                path_score (path)
        endfor

        Weakest_path = max (WP)

Endprocess
```

Fig. 7. WP Metric Calculation.

V. Experiment and Results

In this section, the experiment of the three metrics had been done. The experiment used two attack graphs generated with two different networks. In the following, the network's topology will be explained. Then, the results of applying the metrics will be discussed.

A. Network Design

In this experiment, we chose two networks to test the effect of our proposed metrics. The network A has two workstations, three servers, a firewall and an external attacker, as in Fig. 8. The network connectivity in this network topology depends on firewall rules given as follows:

- There is bidirectional connectivity between the webserver and other machines in the network.
- The external host is the attacker located on the internet and has access to the webserver through HTTP protocol and HTTP port.
- The two workstations and fileserver have access to each other through NFS protocol and NFS port.
- The two workstations and fileserver have access to the internet through HTTP protocol and HTTP port.
- The two workstations have access to the database server.

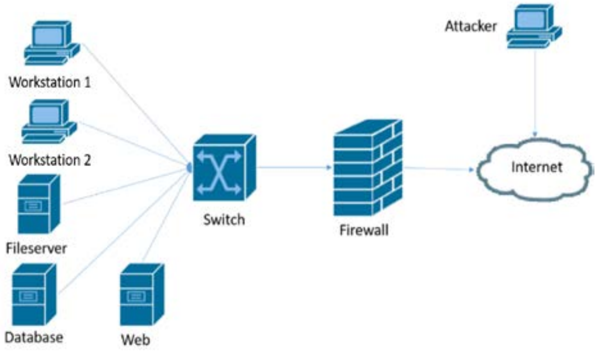


Fig. 8. Network a Topology.

Meanwhile, the network B has two workstations: a web server and database server and an attacker, as in Fig. 9. The network connectivity in this network topology depends on firewall rules listed as follows:

- There is bidirectional connectivity between the webserver and other machines in the network.
- The external host is the attacker located on the internet and has access to the webserver through HTTP protocol and HTTP port.
- Workstation 4 has access to the database server.
- All workstations have access to the internet through HTTP protocol and HTTP port.
- All workstation has a connection with each other.

The experiments evaluated the generating attack graph in the following environment. The CPU is core i5 2.0 GHz with 8 GB of RAM, the operating system is windows 10, and the coding was performed using Microsoft Visual Studio C# 2012. Also, the Common Vulnerabilities and Exposures (CVE) and National Vulnerability Database (NVD) lists were employed, which were provided by NIST to load the vulnerabilities to the attack graph.

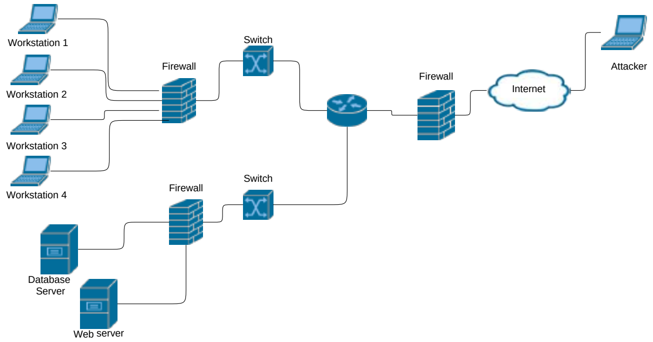


Fig. 9. Network B Topology.

B. Results

In this section, the findings of the experiment will be discussed. Before explaining the results, the paths from the started node (the attacker) to the targeted node (database server) in both networks and list of vulnerabilities must be identified.

The number of the paths in network A is four (4) paths, while the network B has 14 paths, as illustrated in Table I. The paths extracted from the networks based on the topology of the networks.

Here, A represents the word Attacker, W represents the word webserver, D represents the word database, and the numbers represent the workstations. The Attacker can reach the workstations through the webserver only. In this case the next step for the Attacker is webserver. While the webserver has bidirectional connection to the workstations, but the webserver cannot reach directly to the database server which only can be reached by both workstations in network A and workstation 4 in network B, in this case the Attacker need to conquest the workstations before can conquest the database server by exploit the vulnerabilities in the network.

A vulnerability is identified by CVE as a weakness in the computational logic (e.g., code) found in software and hardware components. Usually, the node can have more than one vulnerability because it depends on the applications that installed and the hardware is used. This section will explain some of these vulnerabilities that found in network A and network B as in Table II.

The first vulnerability that discovered is CVE-2010-0490, this weakness in Internet Explorer 6, 7 and 8 with the possibility that remote intruder can execute arbitrary code on the target machine. While CVE-2014-2510 in the outlook, this vulnerability remote authenticated users to read arbitrary files via an external entity declaration in conjunction with an entity reference. The next vulnerability is CVE-2018-15983 in adobe flash player, this vulnerability has an insecure library loading (dll hijacking) vulnerability. Successful exploitation could lead to privilege escalation. The CVE-2010-0483 vulnerability discovered in VBScript, this vulnerability might allow user-assisted remote attackers to execute arbitrary code via a long string in the fourth argument (aka helpfile argument) to the MsgBox function.

1) *NV and MVoP:*

Upon implementing the NV metric on both graphs, the results show that the network A has more vulnerabilities than the network B, as displayed in Table IV. Even the network B has more nodes than the network A, but it has fewer vulnerabilities than the network A. The reason is that the nodes can have more than one vulnerabilities at the same time. Depending in this metric network B is more secure than network A, unlike NP metric which shows than network A is more secure than network B. The reason of the different between two metrics which is the NV metric is depending on the weaknesses of the network not like NP metric which depend on the paths only.

Also, for the MVoP metric, the network A has more vulnerabilities in each path than the network B. The reason for that is the network A has few paths and more vulnerabilities than the network B, as displayed in Table III. The result of MVoP metric shows that network B is more secure than network A, unlike MPL metric which shows network A is more secure. The reason is MVoP depends on the average number of the weaknesses in the path, while MPL depends on the average length of the paths.

TABLE I. PATHS OF THE NETWORKS

No	Network A paths	Network B paths
1	A-W-1-D	A-W-4-D
2	A-W-2-D	A-W-1-4-D
3	A-w-1-2-D	A-W-2-4-D
4	A-W-2-1-D	A-W-3-4-D
5	-	A-W-1-2-4-D
6	-	A-W-1-3-4-D
7	-	A-W-2-1-4-D
8	-	A-W-2-3-4-D
9	-	A-W-3-1-4-D
10	-	A-W-3-2-4-D
11	-	A-W-1-2-3-4-D
12	-	A-W-1-3-2-4-D
13	-	A-W-2-1-3-4-D
14	-	A-W-2-3-1-4-D
15	-	A-W-3-1-2-4-D
16	-	A-W-3-2-1-4-D

TABLE II. EXAMPLE VULNERABILITIES

No	Vulnerability	CVSS Score
1	CVE-2010-0490	9.3
2	CVE-2014-2510	6.8
3	CVE-2018-15983	7.8
4	CVE-2010-0483	7.6

TABLE III. METRICS IMPLEMENTATION RESULTS

Metrics	Network A	Network B
NV	18	15
MVoP	4.5	1.07
NP	4	16
MPL	3.5	5.06

The result of NP metric and MPL metric are not accurate enough of measuring the strength of the network because they do not see the detail of the network, while NV and MVoP metrics are more accurate because of calculating the weaknesses of the network.

2) WP metric: For the third metric, WP metric, the experiment compares the path's strength between the same network paths. To calculate the path strength, the metric takes the highest score vulnerability in the link between two nodes to determine the path's score. Table IV shows the score between all links in both networks.

The results of the edges strength had been calculating depends on the highest vulnerability score in the edges. There are many vulnerabilities had been dropped in the calculation because the score of these vulnerabilities is smaller than the vulnerabilities that been calculated because we assumed that the vulnerabilities that have higher score are more danger and have a high chance to be exploit by the attacker. Based on these edges scores, the weakest path had been calculated as in Table V.

Table VI shows that path number 2 in the network A and path number 8 in the network B has the highest number of WP metric calculation, implying the weakest path in the network.

Comparing with SP metric, the results show that WP metric are more specific, accurate and effective, as in Table VI.

TABLE IV. EDGE VULNERABILITY SCORE

No	Network A links score	Link Score	Network B links score	Link score
1	A-W, W-A	9.8	A-W, W-A	9.2
2	1-2, 2-1	7.8	1-2, 2-1	5.4
3	W-1, 1-W	6.2	1-3, 3-1	8.1
4	W-2, 2-W	8.1	1-4, 4-1	7.2
5	1-D, D-1	6.8	2-3, 3-2	8.1
6	2-D, D-2	8.8	2-4, 4-2	6.4
7			3-4, 4-3	7.8
8			W-1, 1-W	7.3
9			W-2, 2-W	7.8
10			W-3, 3-W	6.1
11			W-4, 4-W	6.4
12			D-4, 4-D	8.1

TABLE V. WP METRIC SCORE FOR BOTH NETWORKS

Number	Network A path	WP metric	Network B path	WP metric
1	A-W-1-D	7.6	A-W-4-D	7.9
2	A-W-2-D	8.9	A-W-1-4-D	7.95
3	A-W-1-2-D	8.15	A-W-2-4-D	7.87
4	A-W-2-1-D	8.12	A-W-3-4-D	7.8
5	-		A-W-1-2-4-D	7.28
6	-		A-W-1-3-4-D	8.1
7	-		A-W-2-1-4-D	7.5
8	-		A-W-2-3-4-D	8.2
9	-		A-W-3-1-4-D	7.74
10	-		A-W-3-2-4-D	7.58
11	-		A-W-1-2-3-4-D	7.65
12	-		A-W-1-3-2-4-D	7.86
13	-		A-W-2-1-3-4-D	7.73
14	-		A-W-2-3-1-4-D	8.08
15	-		A-W-3-1-2-4-D	7.21
16	-		A-W-3-2-1-4-D	7.35

TABLE VI. METRICS COMPARISON

Metric	Network A	Network B
WP	- A-W-2-D	A-W-2-3-4-D
SP	- A-W-1-D - A-W-2-D	A-W-4-D

The SP metric has two paths in the network A, causing misleading results, while the WP metric has been more specific and gets one path only. In the network B, the SP and WP metrics get different paths because the SP metric counts the node number between the started and targeted node. Simultaneously, the WP is more specific and calculates the vulnerability score in each link, giving the easiest path for the attacker to reach the target.

VI. CONCLUSION

In this paper, three metrics had been proposed, which is NV, MVoP, WP metrics. These three metrics depend on vulnerabilities as a major factor to measure the security of the network. The experiment had been performed in two attack graphs generated using two different networks. The results show that the network A has more vulnerabilities, while the MVoP is higher than the network B, even though it has more nodes and paths. For the last metric, the WP metric, the network's A result shows that the shortest path is the weakest path of the network while it was not the shortest path in the network B.

VII. Future Work

Further investigation and research are still required, especially in the flowing fields:

- The work developed using the metrics and the experiments will be performed for larger graphs.
- Also, we will attempt to combine the metrics to obtain better results.

References

[1] Ramos, M. Lazar, R. Holanda Filho, and J. J. P. C. Rodrigues, "Model-based quantitative network security metrics: A survey," IEEE Commun. Surv. Tutorials, vol. 19, no. 4, pp. 2704–2734, 2017.

[2] Z. J. Al-araji, S. S. A. Syed, M. W. Al-salihi, H. A. Al-lamy, M. Ahmed, and W. Raad, "Network Traffic Classification for Attack Detection Using Big Data Tools : A Review," Intell. Interact. Comput. Lect. Notes Networks Syst. 67, pp. 355–363, 2019, doi: 10.1007/978-981-13-6031-2.

[3] A. A. Hassan, W. M. Shah, M. F. I. Othman, and H. A. H. Hassan, "Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks.," Int. J. Electr. \& Comput. Eng., vol. 10, no. 2, 2020.

[4] M. A. Mohammed et al., "A comprehensive investigation of machine learning feature extraction and classification methods for automated diagnosis of covid-19 based on x-ray images," Comput. Mater. Contin., vol. 66, no. 3, 2020.

[5] C. Phillips and L. P. Swiler, "A Graph-based System for Network-vulnerability Analysis," Proc. 1998 Work. New Secur. Paradig., pp. 71–79, 1998, doi: 10.1145/310889.310919.

[6] L. P. Swiler, C. Phillips, D. Ellis, and S. Chakerian, "Computer-attack graph generation tool," Proc. - DARPA Inf. Surviv. Conf. Expo. II, DISCEX 2001, vol. 2, pp. 307–321, 2001, doi: 10.1109/DISCEX.2001.932182.

[7] P. Ammann, D. Wijesekera, and S. Kaushik, "Scalable, graph-based network vulnerability analysis," Proc. ACM Conf. Comput. Commun. Secur., no. June, pp. 217–224, 2002, doi: 10.1145/586110.586140.

[8] V. Mehta, C. Bartzis, H. Zhu, E. Clarke, and J. Wing, "Ranking attack graphs," in International Workshop on Recent Advances in Intrusion Detection, 2006, pp. 127–144.

[9] Y. Chen, K. Lv, and C. Hu, "Optimal Attack Path Generation Based on Supervised Kohonen Neural Network," vol. 2, pp. 399–412, 2017, doi: 10.1007/978-3-319-64701-2.

[10] H. Li, Y. Wang, and Y. Cao, "Searching Forward Complete Attack Graph Generation Algorithm Based on Hypergraph Partitioning," Procedia Comput. Sci., vol. 107, no. Icict, pp. 27–38, 2017, doi: 10.1016/j.procs.2017.03.052.

[11] B. Yuan, Z. Pan, F. Shi, and Z. Li, "An Attack Path Generation Methods Based on Graph Database," in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, vol. 1, pp. 1905–1910.

[12] Y. Cheng, J. Deng, J. Li, S. A. Deloach, and A. Singhal, Metrics of Security, vol. 62. Springer International Publishing Switzerland 2014, 2014.

[13] N. Idika and B. Bhargava, "Extending attack graph-based security metrics and aggregating their application," IEEE Trans. Dependable Secur. Comput., vol. 9, no. 1, pp. 75–85, 2012, doi: 10.1109/TDSC.2010.61.

[14] M. G. and D. S. K. Simon Enoch Yusuf, Jin B. Hong, "Composite Metrics for Network Security Analysis." Journal ofSoftware Networking, pp. 137–160, 2017.

[15] C. Phillips and L. P. Swiler, "A graph-based system for network-vulnerability analysis," in Proceedings of the 1998 workshop on New security paradigms, 1998, pp. 71–79.

[16] X. Ou, "A logic-programming approach to network security analysis," Ph.D Diss., no. November, 2005.

[17] K. Kaynar, "A taxonomy for attack graph generation and usage in network security," J. Inf. Secur. Appl., vol. 29, pp. 27–56, 2016, doi: 10.1016/j.jisa.2016.02.001.

[18] M. U. Aksu, M. H. Dilek, E. İ. Tatlı, K. Bicakci, and M. Ozbayoglu, "Automated Generation Of Attack Graphs Using NVD," 24th ACM Conf. Comput. Commun. Secur., pp. 135–142, 2018, doi: 10.1145/3176258.3176339.

[19] L. Wang, T. Islam, T. Long, A. Singhal, and S. Jajodia, "An attack graph-based probabilistic security metric," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5094 LNCS, pp. 283–296, 2008, doi: 10.1007/978-3-540-70567-3_22.

[20] K. Kaynar, "A taxonomy for attack graph generation and usage in network security," J. Inf. Secur. Appl., vol. 29, pp. 27–56, 2016.

[21] K. Kaynar, "Distributed Log Analysis for Scenario-based Detection of Multi-step Attacks and Generation of Near-optimal Defense Recommendations," 2017.

[22] N. C. Idika, "Characterizing and Aggregating Attack Graph-Based Security Metrics," 2010.

[23] S. Y. Enoch, J. B. Hong, M. Ge, and D. S. Kim, "Composite metrics for network security analysis," arXiv Prepr. arXiv2007.03486, 2017.

[24] A. Roy, D. S. Kim, and K. S. Trivedi, "Scalable optimal countermeasure selection using implicit enumeration on attack countermeasure trees," in IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012), 2012, pp. 1–12.

[25] R. Ortalo, Y. Deswarte, and M. Kaâniche, "Experimenting with quantitative evaluation tools for monitoring operational security," IEEE Trans. Softw. Eng., vol. 25, no. 5, pp. 633–650, 1999, doi: 10.1109/32.815323.

[26] W. Li and R. B. Vaughn, "Cluster security research involving the modeling of network exploitations using exploitation graphs," Sixth IEEE Int. Symp. Clust. Comput. Grid Work. 2006. CCGRID 06, no. July, 2006, doi: 10.1109/ccgrid.2006.1630921.

Creativity Training Model for Game Design

Raudyah Md Tap[1], Nor Azan Mat Zin[2], Hafiz Mohd Sarim[3], Norizan Mat Diah[4]
Faculty of Information Science and Technology, The National University of Malaysia, 43600 Bangi, Malaysia[1, 2, 3]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia[4]

***Abstract*—The popularity of digital games is increasing with a global market value of RM197.6 billion. However, the game produced locally still has no impact. One reason is that there is no emphasis on the game design process in the game development education program. Games designed have a problem in terms of creativity, and there is still no specific method of training creative thinking. This study aims to identify and validate game design's creativity components and develop a Creativity Training Model for Game Design (LK2RBPD Model) verified through the Game Design Document Tool (GDD Tool) prototype. This research has four main phases: the requirements planning, design, development, implementation, and testing phases. In the requirements analysis phase, the component of LK2RBPD Model was identified. The LK2RBPD Model contains elements from industry practices of game designing, creative and innovative thinking skills, creativity dimensions, Sternberg Creativity, and Cultural Activity theories. The GDD Tool prototype implementing the model was developed and tested. The LK2RBPD Model was evaluated using questionnaire survey, SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis, and verification of ideas in the GDD Tool prototype. Evaluation using a five-point Likert scale shows that GDD Tool prototype is effective in implementing 19 components. Expert verification on the results of game design ideas and creativity building using Cohen Kappa calculations is 0.94, indicating an excellent agreement. The results show that the LK2RBPD Model can be effectively used to train creativity in game design. This research's contributions are LK2RBPD Model, creative game design ideation process guideline, and GDD Tool prototype design.**

***Keywords*—Creativity training; game design; creative ideas; creative thinking**

I. Introduction

Technological change is very significant, especially in information and communication technology (ICT), which has led to rapid development in various fields and aspects of life [1]. Changes in ICT have been particularly noticeable when the explosion of electronic and digital devices has dominated today’s technology market. Digital games have also become a large industry, generating billion dollars worldwide. Despite the expansion, this industry is facing a number of problems [2]. The popularity of digital games is increasing, with a global market value of RM197.6 billion. However, the game produced locally still has no considerable impact. One reason is that there is no emphasis on the game design process in the game development education program.

Many digital games on various platforms have been created, either for entertainment or for other purposes. Various studies have shown the potential and effectiveness of computer games for education, they are found to be more efficient, beneficial, and of quality [3] [4] [5]. Regardless of the purpose of the computer games, the fundamental emphasis should be on the game design and the design process's creative aspect.

The game design process plays a crucial role in producing a good and creative game. According to [6], the process of game design is divided into two main phases that need to be emphasized, namely ideation and conceptualization. Ideation refers to the process of generating creative and innovative game design ideas, while conceptualization is to outline the concept in the form of game prototypes that involve programming, architecture, sound, and interface. According to [6], creating an interactive and effective computer game is a huge issue and challenge because various assumptions in the creating of a computer game can have a significant impact on users. This is further reinforced by [7] [8] in their studies, that three important elements are emphasized in the construction of computer games, namely, a creative resource involving graphics and sound, files involving creative graphic rules, and a game engine involving the control of every graphic and sound. According to [9], the game design process needs to have certain creativity based on the practice of human game designers and an advanced analogy with creativity in science, acceptable to computational achievements in the form of discovery systems. In addition, [10] [11] [12] also agreed on the need to generate game ideas that require creative strategies in game designing. According to a recent study by [13], an effort to create or adopt tools that increases team communication is necessary to maintain designers and programmers’ working in a strong collaborative atmosphere to produce creative game designs. Thus, it can be concluded that the computer games design needs to be emphasized in the game development education.

Creativity is defined as a process that occurs to a human being and a quality that is natural and learned. According to [14] [15], creativity is the ability of those who excel in creative endeavors, reflecting the idea of the human imagination as an intuitive notion of skill. The author in [5] considered creation as the process of finding sufficient facts, problems, ideas, and solutions, with cognition, imagination, and judgment, i.e. creative problem-solving. The author in [14] deemed creativity as the creation for particular needs or purposes by incorporating elements connected to form new relationships. Issues in a game product involve issues of creative thinking in game design and development. Creativity and skills are among the most important aspects for a digital game developer and designer. Eighty percent of games are expected to fail in meeting their business objectives due to poor designs and lack of creativity [16]. Creativity is an essential constituent in the game design process, which requires creative ideas, storytelling, character design, interface design, environmental design, and animation [17]. Also, game design demands a continuous process which incorporates many important aspects, including discipline and creativity [18] [19].

Game designers still face difficulty determining players' needs; therefore, the game design process desperately needs a model to help solve innovation and creativity problems [20] [18] [21]. This indicates that the gaming industry can use a model that will help improve the game's quality. According to [22], a software product's design requires a good planning process and a systematic methodological guide to software development. According to [23], innovative and creative processes are among the essential elements of the game business model. Furthermore, the most effective activities toward achieving business success are innovative and creative processes [24].

Meanwhile, the game design process requires more quality resources for the gaming industry's creative development. According to the Official Portal of the Multimedia Development Corporation (MDEC 2019), despite sufficient resources to meet demands, existing talent sources indicate serious problems with quality and skills. Thus, a creative thinking learning model is needed to address the talent gap inherent in the country's creative industry and show that everyone needs to practice creative skills.

II. Research Methodology

This study uses a mixed-method design approach that combines qualitative and quantitative study design to collect systematic data [25] [26]. Data collection includes course document analysis, interviews, questionnaire surveys, SWOT analysis, and analysis of game design documents. This research used Rapid Application Development (RAD) methodology, which includes requirements planning, design, development, implementation, and testing phases. The data collected were from the activities of each phase.

A. *Requirements Planning Phase*

The requirements planning phase involved the collection of the required information at the beginning of the study. In this phase, issues and problems together with elements of the model were identified. The phase activities include unstructured interviews for game design practices from six game designers and three game design lecturers, analysis of game development course documents and program structure, students' final project evaluation reports, students' observations on the game design process, and the literature review. Fig. 1 shows a summary of the activities performed in this phase.

B. *Design Phase*

In the design phase, the elements identified are grouped and mapped into different components of the LK2RBPD Model. The four phases of design activities are (i) excluding/removing elements not suitable for use in creative skills training for game design, (ii) grouping of similar elements by categorizing them into four components categories of designer, knowledge, skills, and technology support, which are essential to the model based on the four dimensions of creativity (person, environment, process, and product) [27][28], (iii) mapping of components by determining their relationship and visualizing via diagrams, based on the Input, Process, and Output model [29][30]; (iv) Expert verification to validate the components in the model using the Inter Rater Reliability (IRR) technique by four experts from the gaming industry and academia [31][32]. Fig. 2 summaries the activities performed in the design phase.

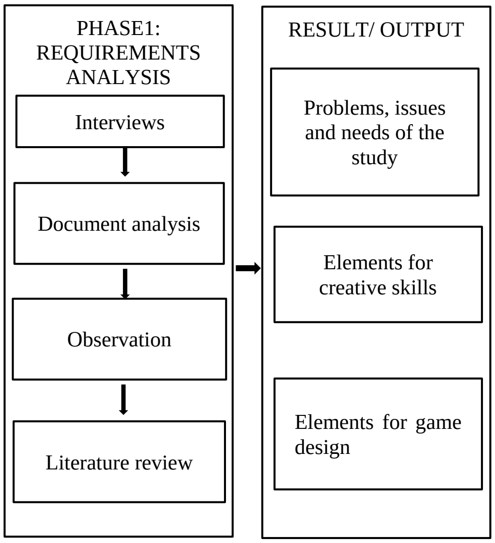

PHASE1: REQUIREMENTS ANALYSIS
Interviews
Document analysis
Observation
Literature review
RESULT/ OUTPUT
Problems, issues and needs of the study
Elements for creative skills
Elements for game design

Fig. 1. Requirement Planning Phase.

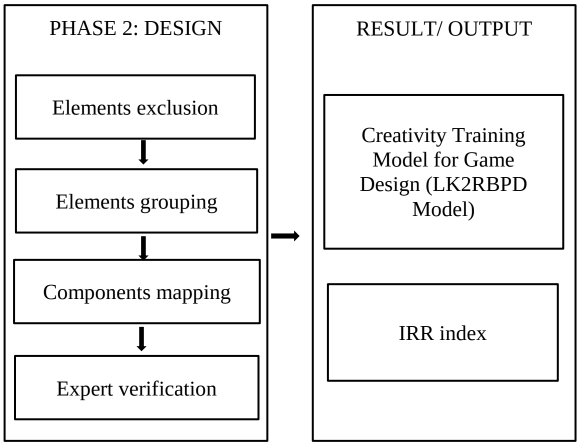


Fig. 2. Design Phase.

C. *Development Phase*

The development phase involves five activities. The first activity is prototype development with implementation of the LK2RBPD Model - the GDD Tool. Second activity is verification of the LK2RBPD Model component in the GDD Tool prototype, conducted by five (5) experts (Numbers (N) = 5: Male (M) = 1; Female (F) = 4) comprising of two game industry experts and three academic specialists / lecturers of game design and development. Semi-structured interview method using a checklist and a low fidelity prototype were used. The third activity is the prototype test, conducted using Thinking Aloud Testing [33] with five students majoring in game design. In the test's procedure, testers would be logged in as system users, and they would follow the instructions given to test the GDD Tool's features: choosing roles, generating ideas, brainstorming, and downloading game design documents. Additionally, respondents would need to provide feedback on the low prototype design of the GDD Tool in the checklist provided. The fourth activity is the questionnaire (instrument) development for content validation on the implementation of the LK2RBPD Model component. A 5-point Likert's Scale (1 = strongly disagree, 2 = strongly disagree, 3 = neutral, 4 = agree, 5 = strongly agree) questionnaire consisted of 83 questions has three main sections: part A (student information), part B (question items according to the LK2RBPD Model component), and part C (SWOT Analysis). Five experts (N = 5: M= 1; F = 4) from various fields of expertise validated the questionnaire. Finally, a pilot study was conducted with 15 students for reliability testing of the instrument. Fig. 3 summarizes the activities in the development phase.

D. Implementation and Testing Phase

In this phase, the GDD Tool prototype was evaluated through 45 students-perception questionnaire survey. Descriptive mean score used three-level indicators: 0.00 - 1.67 = low, 1.68 - 3.34 = medium, and 3.35 - 5.00 = high [34] [35]. Finally, creativity and innovation was evaluated based on observations on ideation process and the game design documents (GDD) produced through GDD Tool. GDD were assessed by five experts, who are game designers. The evaluation was carried out in two sessions. The first session was conducted without using the GDD Tool, in the form of a regular discussion. The researcher provided the students with instructions on the production of game design documents and gave them some ideas on game design. The second session required the GDD Tool, so the testing procedure was initiated with a demonstration that provided the respondents with instructions on how to use the GDD Tool. Fig. 4 shows a summary of the activities in this phase.

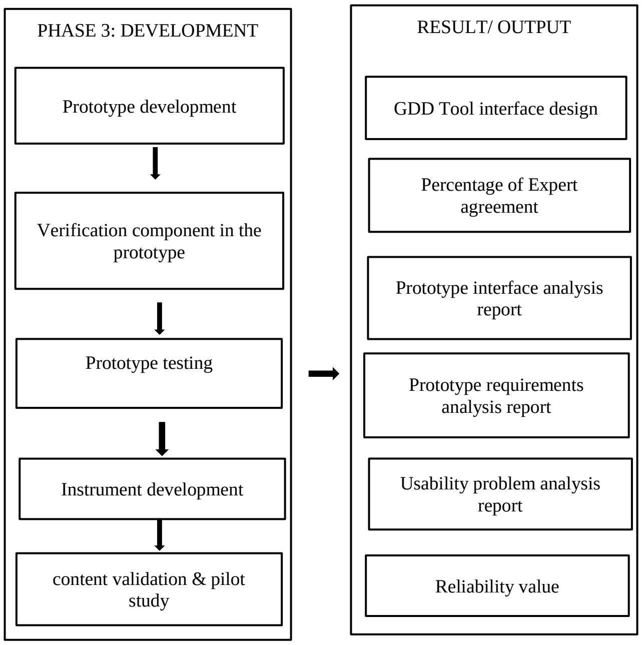

PHASE 3: DEVELOPMENT
Prototype development
Verification component in the prototype
Prototype testing
Instrument development
content validation & pilot study
RESULT/ OUTPUT
GDD Tool interface design
Percentage of Expert agreement
Prototype interface analysis report
Prototype requirements analysis report
Usability problem analysis report
Reliability value

Fig. 3. Development Phase.

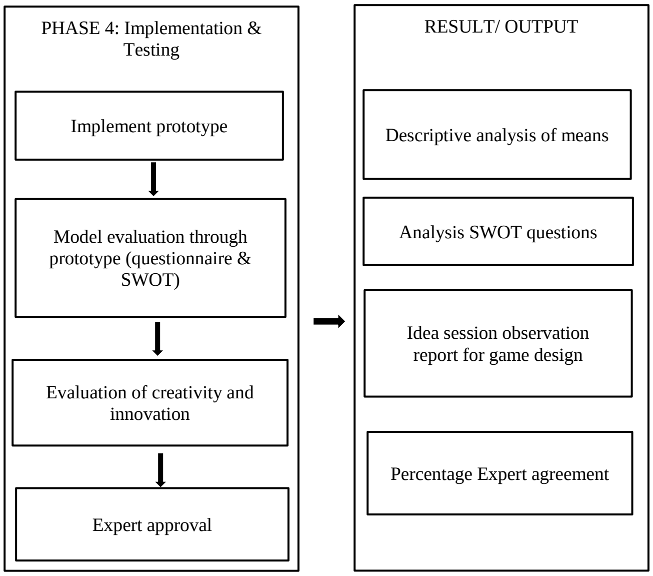

PHASE 4: Implementation & Testing
Implement prototype
Model evaluation through prototype (questionnaire & SWOT)
Evaluation of creativity and innovation
Expert approval
RESULT/ OUTPUT
Descriptive analysis of means
Analysis SWOT questions
Idea session observation report for game design
Percentage Expert agreement

Fig. 4. Implementation and Testing Phase.

III. Results and Discussion

Results identified from requirements analysis were 19 elements essential for training creative skills in game design, consisting of industry practices, creative and innovative thinking skills, creativity dimensions, Sternberg's theory of creativity, and cultural activity theory. These elements were grouped into four components: (1) the designer component consists of elements (intelligence, thinking style, motivation and personality) while (2) the knowledge component consists of the elements (experience, game genre, environment and storyline) next (3) the skills component consists of the elements (linking, synthesizing, imitation, game analysis, generating ideas, inventing and play centric) and last (4) the technology support component that consists of elements (goals, rules, community, and distribution of tasks), forming the LK2RBPD Model.

The IRR approval rate for expert validation of the model is 94 percent, an acceptable level of trust agreement. IRR values from 75 percent to 90 percent indicate acceptable levels of agreement when using a percentage of consent [31] [32] [35]. Fig. 5 shows the LK2RBPD Model with 19 components (intelligence, thinking style, motivation, personality, experience, game genre, environment, storyline, correlate, synthesis, imitation, game analysis, generate idea, invent, play centric, goals, rules, community and task distribution. Model construction and validation has been discussed in detail in [36].

The model evaluation is done via the GDD Tool low fidelity prototype. Results of the experts' agreement on the implementation of the LK2RBPD Model components in the GDD Tool's low-fidelity prototype is good (Cohen Kappa coefficient = 0.84) [37]. However, only 17 components of the model were identified and agreed upon by all the experts (Table I). Table II shows the experts' feedback on GDD Tool prototype interface. Experts' feedback using the provided checklist was then used to develop the final prototype, based on the system requirements (Table III).

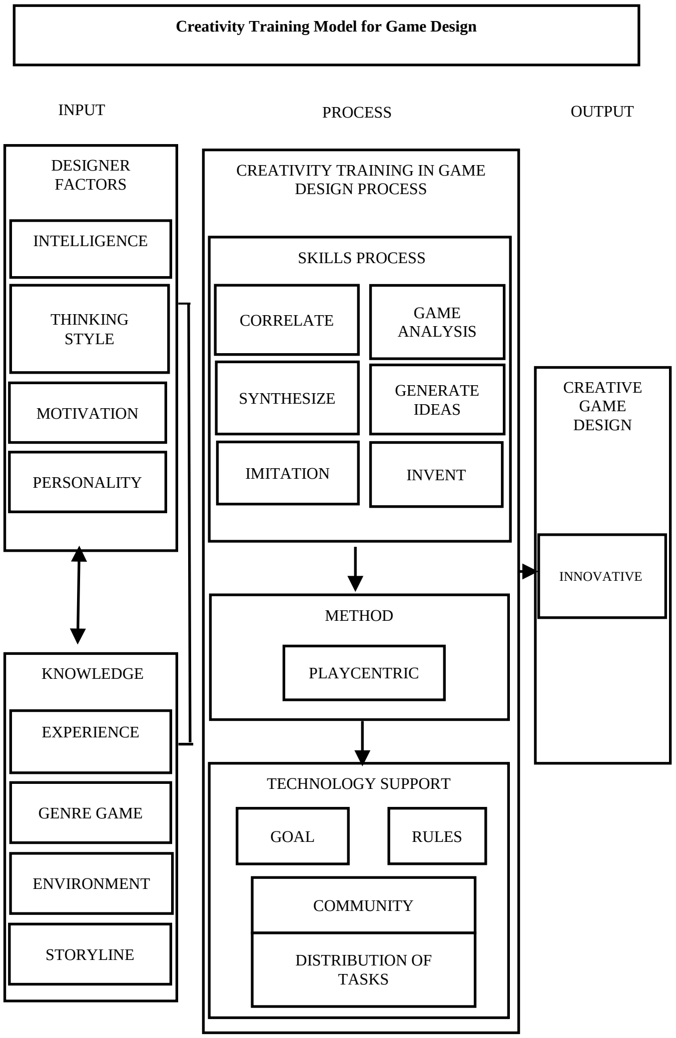


Fig. 5. LK2RBPD Model.

TABLE I. EXPERTS' AGREEMENT ON THE IMPLEMENTATION THE LK2RBPD MODEL COMPONENTS

No	Components	Expert1	Expert 2	Expert 3	Expert 4	Expert 5
1.	Intelligence	/	/	/	/	/
2.	Thinking style	/	/	/	/	/
3.	Motivation	X	X	X	X	X
4.	Personality	/	/	/	/	/
5.	Experience	/	/	/	/	/
6.	Game Genre	/	/	/	/	/
7.	Environment	X	X	X	X	X
8.	Storyline	/	/	/	/	/
9.	Correlate	/	/	/	/	/
10	Synthesise	/	/	/	/	/
11.	Imitation	/	X	/	/	X
12	Game analysis	/	/	/	/	/
13.	Generate ideas	/	/	/	/	/
14.	Invent	/	/	/	/	/
15.	Play centric	/	/	/	/	/
16.	Goal	/	/	/	/	/
17.	Rules	/	/	/	/	/
18.	Community	/	/	/	/	/
19.	Distribution of tasks	/	/	/	/	X

TABLE II. EXPERT FEEDBACK ON GDD TOOL PROTOTYPE INTERFACE

No	Elements of Prototype interface	Suggestion for improvement
1.	Colour usage	• Change the background colour to make it a more vibrant, engaging and effective design. • Ensure appropriate and consistent use of colours and combinations on each interface. • Consistent use of colour for each menu button on the interface.
2.	Text/Fonts style and size	• Change text style and size for communicating purposes, communicating ideas and facts. • Sensitive font: for computer display. • Sans serif font types are more commonly used because it's more accurate for inside use of computer resolution. • The choice of style and text size should be more appropriate to explain how the generated application works, to guide the user to browse the application, and to send information to the designed application. • Consistent use of text size for page title, description of words in words or paragraphs and instructions for operating the application. • Use of the title of each interface in accordance with the functions and processes that will be used by the GDD Tool user. • Consistent use of English.
3.	Menus	• Add menu buttons to the Comment Page and Best Ideas interface for the purpose of providing the client with a flow of games in the diagram (for better understanding). • Add a critic menu button to the Client's Idea interface. • Add an Edit button to the Game Design Document interface. • Use a consistent and appropriate menu button design to ensure that users are aware of the existence of buttons and functions. • The menu button position is placed in the centre of the interface. • The position of the client name is arranged in a more organised manner on the Client Name interface.
4.	Sequence of Action	• The Rules interface is placed on the first interface of the GDD Tool application so that users can better understand the rules when using the GDD Tool.

TABLE III. GDD TOOL PROTOTYPE REQUIREMENTS

No	Prototype Requirements	Suggestion and improvement
1.	Input	• The instructions should be clear for each interface that requires text input from the user. • Suggest options for emoticons or Likes on the Comment Page interface. • The Best Ideas interface reward/likes function to enhance user motivation.
2.	Process	• Functional enhancements to the Role Select interface according to the MDA Framework game design framework which is Mechanics designer, Dynamics designer and Aesthetics designer. • Enhanced chat between clients.
3.	Output	• Need to display instructions or information about the GDD Tool on the first interface that is represented by the Help/About/Rules button. • Replace the rule sentence with the illustration/flow chart/video/animation in the Rules interface. • Need statement or instructions for user to download document in doc/pdf format.
4.	Data source	• Must have a stable domain for easy access to developed applications.
5.	Control	• Functions added to the Game Design Document interface - only hosts are given the functionality to edit, update, and submit. • Extension of time given to client on Comment Page interface according to eureka concept.

System requirements analysis is part of the initial study to identify system-specific problems and requirements, including input requirements, process requirements, output requirements, data handling, and system control resources. System-specific requirements are specifications of what the system will do when implemented [38]. Table III shows an analysis of the GDD Tool prototype requirements.

Usability Testing allows users to try using a real prototype or an application of a particular task. The goal is to find out how well the designed application or service can be used, so that any arising problem can be highlighted or discovered during testing [33]. For usability testing (pilot test) of the GDD tool prototype, five tasks were assigned to 5 users, and 8 problems were identified. The list of usability issues is shown in Table IV.

The agreement on the prototype usability problems, the Cohen Kappa coefficient is 0.54, which indicates a medium-scale level usability problem for the GDD Tool. This value implied that although improvements to the usability problems or issues should still be made, the GDD Tool prototype can be used by the user without major complications. Many of the problems encountered in the test were related to the respondents' misunderstanding the purpose of the GDD Tool application or their inability to identify the next steps in the application process, similar to the problems found by [39] [40] [41] in studies done with more experienced users. Additionally, the reliability value of the Cronbach's alpha was 0.84, indicating that the questionnaire items had reliability [37], so can be used for final evaluation.

TABLE IV. USABILITY PROBLEMS OF GDD TOOL PROTOTYPE

No	Tasks (T)	Code Problen (P)	Problems	User (U)
1.	Task 1: User needs to join the idea brainstorming session (T1)	P1	Users do not know how to start the application.	U1, U2, U3, U4, U5
2.	Task 2: User needs to select a role in the GDD Tool. (T2)	P2 P3	Users are not clear about their role in the selection process. Host is not clear about the role of the GDD Tool.	U1, U2, U4 U5
3.	Task 3: Users need to generate ideas in game design (T3)	P4	Users require more time beyond the allocated 15 minutes to generate ideas.	U1, U2, U3, U4, U5
4.	Task 4: Users need to brainstorm and choose creative game design ideas (T4)	P5 P6 P7 P8	Users do not understand the instructions provided. User shows or expresses impatience by clicking on the objects that respond slowly. Users show a tendency to take actions randomly (intentionally or otherwise). Button disappears when clicked.	U1, U2, U4 U1 U1, U2 U4, U5
5.	Task 5: Users need to create and download Game Design Document GDDD (T5)		No issue	U1, U2, U3, U4, U5

The final results from questionnaire survey show that all the components of the LK2RBPD Model have high ratings. Overall, the analysis showed that 19 components of the LK2RBPD Model have high mean reading scores of 4.11 to 4.30, as shown in Table V.

Results of SWOT analysis of the GDD Tool prototype is shown in Table VI. Theme analysis was used to sort out answers based on themes of strengths, weaknesses, opportunities, and threats.

The results of the SWOT analysis show that the prototype has strengths and opportunities. However, some threats and weaknesses in the areas of technology need to be improved, such as internet accessibility, poor information sharing, misuse of application features, and competition from social applications.

Finally, the result of creativity and innovation evaluation based on the ideation session process is presented in Table VII while experts' evaluation of the game design documents using the GDD Tool prototype is in Table VIII.

TABLE V. MIN SCORES AND STANDARD DEVIATIONS FROM SURVEY QUESTIONNAIRE IMPLEMENTATION MODEL LK2RBPD

No	Construct	Min Scores	Standard Deviations
1	Intelligence	4.21	0.63
2	Thinking style	4.30	0.59
3	Motivation	4.20	0.57
4	Personality	4.16	0.71
5	Experience	4.11	0.67
6	Game Genre	4.20	0.64
7	Environment	4.28	0.58
8	Storyline	4.25	0.63
9	Correlate	4.28	0.59
10	Synthesise	4.30	0.59
11	Imitation	4.23	0.67
12	Game analysis	4.33	0.58
13	Generate ideas	4.26	0.62
14	Invent	4.21	0.59
15	Play-centric	4.24	0.55
16	Goal	4.23	0.61
17	Rules	4.23	0.57
18	Community	4.23	0.65
19	Distribution of tasks	4.32	0.54

TABLE VI. SWOT ANALYSIS

Team	Answer
Strengths	"Can play more than once anywhere" “Able to come up with ideas and create stories quickly” “It really helps to come up with ideas” “Can improve a lot of ideas” “Lots of creative ideas” “Being able to talk online with a group of friends to build the best game design” “GDD tool helps me to know more about the process of creating game designs, game design documents” “Being able to share ideas and stimulate creative ideas” “GDD Tool can improve creative thinking skills” "I can think a lot" “Being able to connect with friends and talk about game design ideas” “Each user can play their own role” “Get lots of creative ideas” “Developing ideas for the better
Weaknesses	“The interface can be improved to make it more interesting” “Needs strong internet access” "Host role may expand" “No rules on the voting site need complete instructions” “Add community members for more ideas” “No images led to the development of game ideas”
Opportunities	"It's accessible everywhere but it's still a one-on-one session for developing creative ideas in game design" "Be able to train yourself to think creatively" “Can be used many times” “Can easily create GDD” “Can add other functions like audio and other” “Allows sharing of ideas anywhere” "Can be used by everyone" "Students get ideas for game design" “Final semester students will gain more and more knowledge in game design when using the GDD Tool “ “Can expand the use of the GDD Tool to other students / institutions” “Can introduce this GDD Tool to other colleges that have gaming programs” "Can be a training tool for train creative thinking" "Allows many people to use it well"
Threats	"Too easy to disconnect" "Requires a strong internet connection" “Theft of ideas can happen”

TABLE VII. IDEATION SESSION REPORT

Discussions without using the GDD Tool:	Discussions using the GDD Tool:
• There is no fixed time on when to start the discussion.	•The discussion session is conducted at a time that is set and controlled by a Host.
• Discussions take long time and waste of time without any brainstorming activity.	• The discussion is clearly set out in accordance with the goals and rules set in the GDD Tool.
• Does not achieve the actual goal of the discussion.	• Community members have the time and platform to generate ideas.
• All members of the community want to talk at the same time.	• Ideas are generated based on the keywords provided in the GDD Tool.
• There is no platform for channelling ideas.	• Ideas can be created creatively.
• Ideas are written on paper and difficult to keep.	• Each member’s ideas can be shared with other community.
• The discussion took a long time because there was no proper monitoring and distribution of tasks among the members	• Using the GDD Tool can train creative thinking for game ideas, while game designers can develop their imagination and introduce new game ideas that can be produced.
• Difficult to generate ideas. • The situation cannot be controlled and it is difficult to obtain conclusion on the game design.	• Generate formal ideas by developing mechanics, dynamics and aesthetics elements in game design.
• Ideas can be produced, but not made in a complete game design document, instead just drawings and notes that cannot be referenced.	• Ideas are improved as they are generated based on game objectives, user targets, game levels, and game names.
• Each member’s ideas cannot be shared with other community partners.	• A complete and creative game design document can be created as soon as the GDD Tool usage session ends.
	• Game design documents can use as a reference for the game development phase.

A total of nine (9) game design documents were produced from the use of the GDD Tool session by 45 respondent focus groups of students majoring in game design and development. GDDs analysis produced a total of 72 data (ideas). The experts' agreement on the creative game design ideas is 97 per cent. For comparison, the level of expert agreement is calculated using Cohen Kappa formula, K: 0.94 (94%), which signifies excellent agreement [42]. The evaluation results confirm that the LK2RBPD Model could be used effectively to develop and train creativity in game design. Table VIII and Table IX show the experts agreement on game design ideas derived from the final test of the GDD Tool prototype.

TABLE VIII. EXPERTS' APPROVAL

Game Design Document	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**(%)**
Creative Game Design Ideas	72	72	72	72	72	100
Agreement	67	70	72	72	69	70
Percent Agreement	93	97	100	100	95	97

TABLE IX. CALCULATION OF EXPERTS' APPROVAL USING THE COHEN KAPPA FORMULA

<table>
<tr><th>Step</th><th colspan="6">Calculation</th></tr>
<tr><td>Step 1</td><td colspan="6">Formula:
$K=\frac{fo-fc}{N-fc}$
fo- Agreement
fc- 50 Expectation percent agreement
N – Number of game design ideas</td></tr>
<tr><td rowspan="5">Step 2</td><td colspan="6">Calculation</td></tr>
<tr><td>Expert 1</td><td>Expert 2</td><td>Expert 3</td><td>Expert 4</td><td>Expert 5</td><td>Expert Agreement</td></tr>
<tr><td>fo = 67
fc = 36
N = 72</td><td>fo = 70
fc = 36
N = 72</td><td>fo = 72
fc = 36
N = 72</td><td>fo = 72
fc = 36
N = 72</td><td>fo = 69
fc = 36
N = 72</td><td rowspan="3">$K = \frac{0.86 + 0.94 + 1.0 + 1.0 + 0.91}{5}$
$K = \frac{4.71}{5}$
<b>K = 0.94</b></td></tr>
<tr><td>$K=\frac{fo-fc}{N-fc}$</td><td>$K=\frac{fo-fc}{N-fc}$</td><td>$K=\frac{fo-fc}{N-fc}$</td><td>$K=\frac{fo-fc}{N-fc}$</td><td>$K=\frac{fo-fc}{N-fc}$</td></tr>
<tr><td>$K=\frac{67-36}{72-36}$
$K = \frac{31}{36}$
K = 0.86</td><td>$K=\frac{70-36}{72-36}$
$K = \frac{34}{36}$
K = 0.94</td><td>$K=\frac{72-36}{72-36}$
$K = \frac{36}{36}$
K = 1.0</td><td>$K=\frac{72-36}{72-36}$
$K = \frac{36}{36}$
K = 1.0</td><td>$K=\frac{69-36}{72-36}$
$K = \frac{33}{36}$
K = 0.91</td></tr>
</table>

IV. CONCLUSION AND IMPLICATIONS

This study successfully identified 19 components of the LK2RBPD Model and implemented them in the GDD Tool prototype. Evaluation results of the LK2RBPD Model components in the GDD Tool prototype, and the assessment of creativity and innovation in the game design confirmed that the LK2RBPD Model could be used for training the creative skills of game design to produce game. The results of this study successfully identified 19 components of the LK2RBPD Model and implemented them in the GDD Tool prototype. Evaluation result of the LK2RBPD Model components in the GDD Tool prototype, and the assessment of creativity and innovation in the game design confirmed that the LK2RBPD Model could be used for training the creative skills of game design to produce game design documents. Evaluation result of the LK2RBPD Model components through the use of the GDD Tool prototype indicated a rating higher than 4.0, which means that the GDD Tool prototype can be used to train creative skills in game design. As for the results of experts' agreement on creativity and innovation in game design, creative game design documents showed an excellent agreement of 97 per cent or K: 0.94. This shows that creative skills can be trained by involving the important and necessary components of the LK2RBPD model in producing a creative game design. Apart from that, technological aids/tools also help train creative skills in game design. The results of this study have identified three (3) guidelines that need to be in place to generate a creative idea and can be used as a reference for game designers in the process of generating creative game design ideas. This study found that creative thinking can be trained on each individual in particular, to produce a creative game design document by following these three guidelines: (i) ideas are generated through a selection process of some of the best ideas, ii) the ideas produced need to be targeted such as the target user, game name, game objectives and game level and iii) idea generation needs to be more formal that is according to the rules of form, characteristics and certain ways that are acceptable and considered appropriate according to the rules or steps necessary in a game design. The five experts in the field of game design agreed that creative skills training is effective in developing creativity and innovation in the ideation phase of game design. Individuals could be trained successfully in creative thinking, as well as to develop creative ideas by implementing the 19 components of the LK2RBPD Model. The results demonstrate that the objectives of the study have been achieved and the LK2RBPD Model can help solve some problems in this research area. However, further research can be carried out with the use of elements involving creativity and innovation as well as other appropriate training and learning theories according the needs to train creative skills in game design. In addition, the improvement of GDD tools in various platforms with the use of more interactive elements such as chat rooms, image uploads, videos, animations and software functions according to the latest trends that can be used as a communication platform to game designers in producing a creative game design.

ACKNOWLEDGMENT

We acknowledge the assistance from Metro Polytechnic Kuala Lumpur and the Polytechnic and Community College Department, University Kebangsaan Malaysia (UKM) for supporting the research under grant GUP-2019-059 and HLP Federal Training Specialist, Ministry of Higher Education Malaysia for funding the study.

REFERENCES

[1] Chernyakov, M., & Chernyakova, M. Technological risks of the digital economy. Корпоративные финансы, 12(4), 2018.

[2] Sharif, M., Zafar, A., & Muhammad, U. Design patterns and general video game level generation. International Journal of Advanced Computer Science and Applications, 8(9), 393-398, 2017.

[3] De Freitas, S. Are games effective learning tools? A review of educational games. Journal of Educational Technology & Society, 21(2), 74-84, 2018.

[4] Vlachopoulos, D., & Makri, A. The effect of games and simulations on higher education: a systematic literature review. International Journal of Educational Technology in Higher Education, 14(1), 22, 2017.

[5] Barr, M. Student attitudes to games-based skills development: Learning from video games in higher education. Computers in Human Behavior, 80, 283-294, 2018.

[6] Athavale, S., & Mohan, A. Understanding Game Ideation Through The Lens Of Creativity Model. In DS 89: Proceedings of The Fifth International Conference on Design Creativity (ICDC 2018), University of Bath, Bath, UK (pp. 176-182), 2018.

[7] Kasurinen, J., Palacin-Silva, M. & Vanhala, E. What Concerns Game Developers? A Study on Game Development Processes, Sustainability and Metrics. International Workshop on Emerging Trends in Software Metrics, WETSoM (May): 15–21, 2017.

[8] Byun, J., & Loh, C. S. Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. Computers in Human Behavior, 46, 129-138, 2015.

[9] Ritchie, G. The evaluation of creative systems. In Computational Creativity (pp. 159-194). Springer, Cham, 2019.

[10] Giannakos, M. N., & Jaccheri, L. From players to makers: An empirical examination of factors that affect creative game development. International Journal of Child-Computer Interaction, 18, 27-36, 2018.

[11] Guzdial, M., Liao, N., & Riedl, M. Co-creative level design via machine learning. arXiv preprint arXiv:1809.09420, 2018.

[12] Machado¹, T. L. D. A., Ramalho, G. L., Alves, C. F., Garcia, V. C., Araujo¹, L. F., Lemos¹, V., ... & do Recife, S. A. Game development guidelines: Practices to avoid conflicts between software and design. Software Engineer, 3, 8, 2010.

[13] Henriksen, D., Mishra, P., & Fisser, P. Infusing creativity and technology in 21st century education: A systemic view for change. Educational Technology & Society, 19(3), 27-37, 2016.

[14] Chee, C. M., & Wong, D. H. T. Affluent gaming experience could fail gamification in education: a review. IETE Technical Review, 34(6), 593-597, 2017.

[15] Hendrik, B., Ali, N. M., & Nayan, N. M. Robotic Technology for Figural Creativity Enhancement: Case Study on Elementary School. International Journal Of Advanced Computer Science And Applications, 11(1), 2020.

[16] Hook, K. Designing with the body: somaesthetic interaction design. MIT Press, 2018.

[17] Friesike, S., Flath, C. M., Wirth, M., & Thiesse, F. Creativity and productivity in product design for additive manufacturing: Mechanisms and platform outcomes of remixing. Journal of Operations Management, 65(8), 735-752, 2019.

[18] Kanode, C. M. & Haddad, H. M. Software engineering challenges in game development. ITNG 2009 - 6th International Conference on Information Technology: New Generations 260–265, 2019.

[19] O'Hagan, A. O., Coleman, G., & O'Connor, R. V. Software development processes for games: A systematic literature review. In European Conference on Software Process Improvement (pp. 182-193). Springer, Berlin, Heidelberg, 2014.

[20] Refai, J. J., Bateman, S., & Fleming, M. W. External Assistance Techniques that Target Core Game Tasks for Balancing Game Difficulty. Front. Comput. Sci. 2: 17. 2020.

[21] Politowski, C., Fontoura, L., Petrillo, F., & Guéhéneuc, Y. G. Are the old days gone? A survey on actual software engineering processes in video game industry. In Proceedings of the 5th International Workshop on Games and Software Engineering (pp. 22-28), 2016.

[22] Kasurinen, J., Palacin-Silva, M., & Vanhala, E. What concerns game developers? a study on game development processes, sustainability and metrics. In 2017 IEEE/ACM 8th Workshop on Emerging Trends in Software Metrics (WETSoM) (pp. 15-21). IEEE, 2017.

[23] Harel, R., Schwartz, D., & Kaufmann, D. The relationship between innovation promotion processes and small business success: the role of managers' dominance. Review of Managerial Science, 1-24, 2020.

[24] Rutberg, S., & Bouikidis, C. D. Focusing on the fundamentals: A simplistic differentiation between qualitative and quantitative research. Nephrology Nursing Journal, 45(2), 209-213, 2018.

[25] Yusoff, S. R. M., & Zin, N. A. M. Design and evaluation of collaborative learning management system (clms) framework for teaching technical subject. In International Conference on Web-Based Learning (pp. 79-89). Springer, Berlin, Heidelberg, 2012.

[26] Garces, S., Pocinho, M., de Jesus, S. N., & Viseu, J. The impact of the creative environment on the creative person, process, and product. Avaliação Psicologica, 15(2), 169-176, 2016.

[27] Widana, I. W., Sumandya, I. W., Sukendra, I. K., & Sudiarsa, I. W. Analysis of conceptual understanding, digital literacy, motivation, divergent of thinking, and creativity on the teachers skills in preparing hots-based assessments. Journal of Advance Research in Dynamical & Control Systems, 12(8), 459-466, 2020.

[28] Subiyakto, A. A., Ahlan, A. R., Putra, S. J., & Kartiwi, M. Validation of information system project success model: a focus group study. SAGE Open, 5(2), 2158244015581650, 2015.

[29] Mansikka, H., Harris, D., & Virtanen, K. An input–process–output model of pilot core competencies. Aviation Psychology and Applied Human Factors, 2017.

[30] McDonald, N., Schoenebeck, S., & Forte, A. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-23, 2019.

[31] Belur, J., Tompson, L., Thornton, A., & Simon, M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociological methods & research, 0049124118799372, 2018.

[32] Alhadreti, O., & Mayhew, P. Rethinking thinking aloud: A comparison of three think-aloud protocols. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12), 2018.

[33] Hollon, S. D., & Kendall, P. C. Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. Cognitive therapy and research, 4(4), 383-395, 1980.

[34] Ruiz, F. J., Suárez-Falcón, J. C., & Riaño-Hernández, D. Validity evidence of the Spanish version of the automatic thoughts question nnaire–8 in Colombia. The Spanish Journal of Psychology, 20, 2017.

[35] Wilhelm, A. G., Rouse, A. G., & Jones, F. Exploring differences in measurement and reporting of classroom observation inter-rater reliability. Practical Assessment, Research, and Evaluation, 23(1), 4, 2018.

[36] Raudyah Md Tap,Nor Azan Mat Zin and Hafiz Mohd Sarim,"Creative Game Design Training Requirements," International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 1, pp. 64-71, 2021.

[37] Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In Proceedings of the 31st European Conference on Cognitive Ergonomics (pp. 207-214), 2019.

[38] Podvezko, V. Determining the level of agreement of expert estimates. International Journal of Management and Decision Making, 8(5-6), 586-600, 2007.

[39] Stachtiari, E., Mavridou, A., Katsaros, P., Bliudze, S., & Sifakis, J. Early validation of system requirements and design through correctness-by-construction. Journal of Systems and Software, 145, 52-78, 2018.

[40] Morey, S. A., Barg-Walkow, L. H., & Rogers, W. A. Managing heart failure on the Go: Usability issues with mHealth apps for older adults. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 61, No. 1, pp. 1-5). Sage CA: Los Angeles, CA: SAGE Publications, 2017.

[41] Srikesavan, C., Williamson, E., Cranston, T., Hunter, J., Adams, J., & Lamb, S. E. An online hand exercise intervention for adults with rheumatoid arthritis (mySARAH): design, development, and usability testing. Journal of medical Internet research, 20(6), 2018.

[42] Zec, S., Soriani, N., Comoretto, R., & Baldi, I. Suppl-1, M5: high agreement and high prevalence: the paradox of Cohen's Kappa. The open nursing journal, 11, 211, 2017.

Spiritual User Experience (iSUX) for Older Adult Users using Mobile Application

Nahdatul Akma Ahmad[1*]
Department of Computer Science
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus, Malaysia

Zirawani Baharum[2*]
Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur
Bandar Seri Alam, Johor
Malaysia

Azaliza Zainal[3]
Department of Computing
Faculty of Communication, Visual Art and Computing
Universiti Selangor
Bestari Jaya, Selangor, Malaysia

Fariza Hanis Abdul Razak[4], Wan Adilah Wan Adnan[5]
Department of Information Technology
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam Campus, Selangor, Malaysia

Abstract*—The increasing number of aging populations worldwide versus vast developments in mobile technology creates questions on how older adults adapt and apply mobile technology in their daily life. This research focused on spiritual user experience for older adult users because older adults are claimed to be more spiritually inclined as they aged. Despite high profile calls for research in the area of spirituality, the research pertaining spirituality in HCI is still in infancy state. Recent literatures discover most studies focus on design for spiritual user experience and evaluation of spiritual application for adult users, but fundamental of spirituality and its elements from the view of user experience is limited. Therefore, this study employs qualitative method approach within an interpretive paradigm to propose a model for Spiritual User Experience from the perspective of Islamic older adult users. The Geneva Emotional Musical Scale (GEMS) was adopted as a theoretical lens in order to gain deeper insights on the spirituality elements. A single case study was conducted with the total of 11 participants to research on the spirituality user experience elements among older adults. The triangulation of qualitative data collection through 3E diary, interviews and observations was conducted. All data were analyzed verbatimly using thematic analysis. Six themes emerged from the analysis which are effectiveness, efficiency, learnability, satisfaction, sublimity and vitality. These themes are further categorized into 10 attributes; effectiveness (accessibility features), efficiency (simplicity and portability), learnability, satisfaction (attractiveness and reliability), sublimity (transcendence and peacefulness) and vitality (energy and joyful activation). These are embedded into a model known as Spiritual User Experience (iSUX) which are evaluated by the Islamic religious experts, user experience expert and older adult's representatives. This model could be a reference for spiritual model development apps among developers and provide understanding for researchers in the HCI domain. In conclusion, the Spiritual User Experience (iSUX) is hope to increase the understanding of spirituality from the domain of user experience.

Keywords*—*Techno-spiritual; user experience; human computer interaction; Geneva emotional musical scale; 3e diary; older people

I. Introduction

Spirituality is the most important part of life for every people. Spirituality is the feeling of connectedness to the higher power or consciousness and it also develop individual beliefs, practices and rituals [1]. Nowadays, people start to use technology in many ways to support spirituality. There are a lot of technological applications including the ICT tools such as email, SMS, computer software, websites and mobile applications for supporting spirituality practices. The use of technology in spirituality is also high, for instance, [2] reported that there are almost 6,000 iPhone and iPad applications exist in the market to assist human spirituality from various perspectives. Additionally, [3] highlighted that there is an increasing number of Islamic websites created daily on the internet to spread teachings and services related to spirituality. Apparently, people are searching for spiritual technologies to support for their spiritual experiences with the help from the technology.

Spiritual experience is a type of user experience. Spiritual experience is a rich body of work that addresses user experiences related to the feeling of transcendent, connectedness to God and consciousness toward higher power [4]. The activity of using the spiritual technology can evoke various kind of affect and emotions. These emotions are evoked by the interaction while engaging with spiritual technology. Individuals will respond emotionally to these activities because of their personal spirituality values. For example; a user feels happy when reciting Quran using mobile Quran application; or a user feel satisfied with achievement of dzikir since personal dzikir informatics can keep track of their dzikir counter. Since then, HCI has expanded in scope to move beyond usability and focusing on this subjective spiritual experience. There is, for example, a little work being done in HCI on affect and emotion for spirituality [2]. Most of other HCI studies focusing on the design element that can trigger the spirituality feelings and emotions but did not mention specifically the spirituality elements from the perspectives of

*Corresponding Author

user experience. Moreover, the statistics reveals that researchers are more interested in focusing on the adults as a target user whereas the study on older people study is way behind [5]. Apparently, research for older people in spirituality domain is important since a theory proves that people will become more spiritualized as they age [6]. Therefore, this research aims to answer the following research questions:

1) What are the spirituality elements from the context of user experience?

2) How can the identified spirituality elements be used as a guideline in spiritual user experience application development?

II. Research Background

Numerous models and frameworks have been found in the literature to explain about UX. Apparently, UX has been used widely across many research fields such as business, social science, medical, design and also information technology [7],[8]. In consequence, it shows that UX is an important element especially when it comes to product, system or services delivery towards users. In spite of various UX model there is one comprehensive UX model proposed by [9] which specifies three UX core components and the relationship among it known as Component Model of User Experience (CUE-Model). In 2008, the model has been justified by [10] and the enhancement of the model lead towards the new model. The model addresses instrumental and non-instrumental quality as well as emotions responses by user towards technology. The instrumental quality in the model refers to usability which is the effectiveness of a system [9] and non-instrumental quality involves the hedonic [11] elements such as visual aesthetics and haptic quality [9].

Music plays a great role in human everyday lives. Obviously by listening to music could evoke human positive and negative emotions such as happy, sad, amaze, motivated and energetic. Many research has been done to explore on the links between emotive states evoked by music. A study by [12] proposes the Geneva Emotional Music Scale or GEMS which is widely used. The GEMS scale taken from [12] can be categorized into three major emotion components which are the vitality, unease and sublimity. The first component which is vitality is a positive emotions and includes emotions such as power and joyful activation. The second component is unease that is negative emotions which include tension and sadness elements. And the last emotion component is sublimity which is a positive component and includes elements such as wonder, transcendence, tenderness, nostalgia and peacefulness.

III. Methods

A. The Participants

Kompleks Warga Emas Seksyen 24 (Golden Citizen Complex) located in Shah Alam, Selangor, Malaysia is being selected as a case study of this research. This complex is a center which gathered older adult members from within Shah Alam city to promote beneficial programs or community activities in addition to reduce gap between young and old generations. This complex was chosen as the research area because Shah Alam is the state capital of Selangor in which Selangor has the highest population density among all states in Malaysia [13]. Muslim older adults were chosen in this research since Malay is the highest race composition in Malaysia and majority of Malay is Muslims. Total of 11 Muslim older adult participants were selected based on purposive sampling from among the participants who volunteered to participate in the Learning Quran Bahasa Melayu mobile application workshop using Tablet PC conducted by researchers.

B. Model Development Process

This research comprises of five processes as depicted in Fig. 1 which are research planning, pilot study, UX evaluation, UX model development and expert evaluation. Detail explanation for each processes were discussed further in the following section.

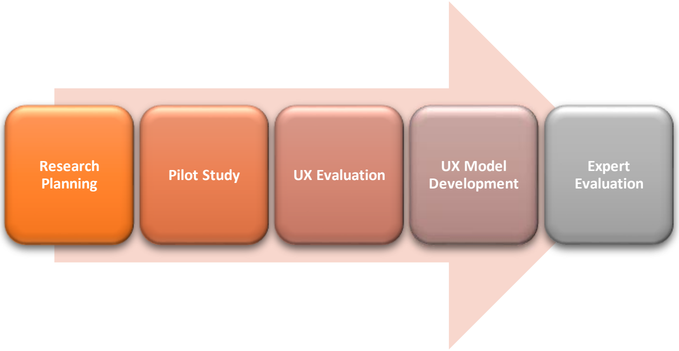


Fig. 1. Research Design.

1) Research planning: Research planning involves steps such as initiation of the research and also development of the research plan. It is important in this study to first understand the concept of spirituality in the context of user experience. Systematic literature review involves selected journals and proceedings from top listed HCI indexed papers listed by google scholar metric raking is used. Review from previous prominent scholars provides guidance to obtain definition and identify the spirituality elements.

2) Pilot study: This pilot study aims to identify data collection procedure and to test diary and interview instruments to be used for the older adults.

Three types of diary design as shown in Fig. 2 were tested with older adult participants which are daily dairy, 3E diary and structured diary. Two older adults' participants which is a male, 62 years old and a female, 60 years old were selected using snowball sampling to be interviewed and involved in a diary study for seven consecutive days. The selected case study is within Selangor state. From the pilot study reveals that 3E diary design is the most suitable diary tool to explore spirituality experience compared to the other two diaries since 3E diary allows older adults to draw facial expressions to express emotional status furthermore encourage older adults to explain it in their writing. For instance, in spirituality context, the older adults may draw crying face to express the feeling of sadness and repentance when the older adult users feel the connection with God. Further explanation on pilot study is as discussed in previous paper [14]. The selected tool will next be used in the next process which is UX evaluation.

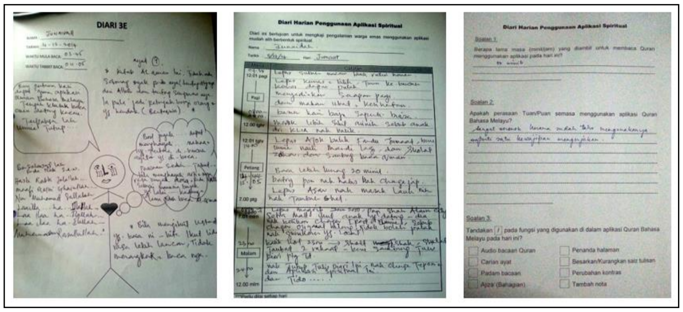

DIARI 3E
Diari Harian Penggunaan Aplikasi Spiritual
Diari Harian Penggunaan Aplikasi Spiritual

Fig. 2. Three Types of Diary Design Tested with Participants.

3) UX evaluation: This research implemented single case study research strategy to obtain spiritual user experience data from the older adult users. Fig. 3 shows the procedures involved in user experience evaluation. The following brief explanation is for the procedures implemented in user experience evaluation with older adult users.

- Set appointment and asking for participant's approval. Relationship between researchers and complex manager is good since researchers have joined several community activities with older adults in the past. Researchers went to the complex to meet the management committee two weeks before the workshop day to book for workshop room and to set the workshop date. To create rapport, researchers also participate in the series of Al-Quran recitation classes at the complex.

- Workshop I & II. During the workshops, ll participants were seated according to their preferences. Program started with speech from the Program Director explaining about the objectives of the workshop. Next, a montage and tutorial video on how to use Tablet PC and Quran Bahasa Melayu application were presented to all audiences. Older adult participants were then divided into three small groups by researchers and the division of group is made according to their gender either male group or female group. Each group consists of either 2 or 3 persons. Firstly, the facilitator demonstrates the method to use the Tablet PC to participant 1. Next, participant 1 need to teach the other team members alternately in a small teaching team. Teaching process is finish when every member in the group understands all features and functions in the Quran Bahasa Melayu application. Teaching team procedures is as shown in Fig. 4. The following Table I and Table II presented the demographic profile of participants in workshop I and workshop II.

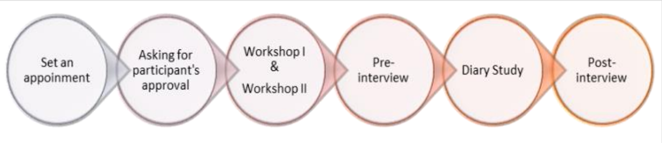


Fig. 3. User Experience Evaluation Procedures.

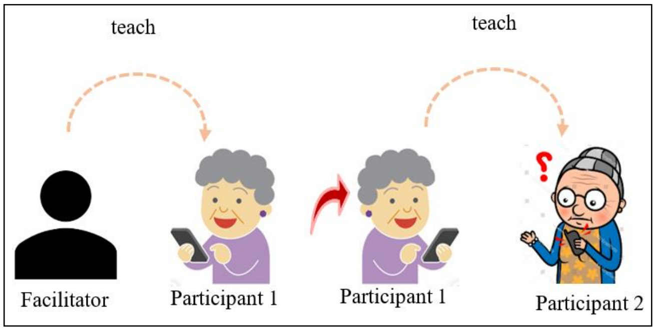


Fig. 4. User Experience Evaluation Procedures

TABLE I. LIST OF PARTICIPANTS FOR WORKSHOP I

Participant ID	Age	Gender	Employment Background
P1	68	Male	Retiree
P2	61	Female	Retiree
P3	68	Female	House wife
P4	71	Female	Retiree
P5	74	Male	Retiree
P6	61	Male	Working (part time)

TABLE II. LIST OF PARTICIPANTS FOR WORKSHOP II

Participant ID	Age	Gender	Employment Background
P7	65	Male	Working
P8	68	Female	Retiree
P9	61	Female	Retiree
P10	63	Male	Retiree
P11	67	Female	House wife

User experience evaluation workshop has been executed twice. The reason for conducting the workshop twice is due to the lack of participant's number in workshop I. Since the workshop was held on weekend, there were fewer participants present because they had to attend the wedding feast, visit children's house and taking care of daughter in confinement period. Moreover, the workshop is tentatively done in small scale number of participants (around 10 to 15 participants only), so that participants can learn in a conducive pace. The second series of workshops were also conducted using the same procedure.

- Pre-interview. After completing the teaching session, five participants were randomly selected to be interviewed. Interview sessions were conducted in the workshops one by one by three facilitators. Each interview session lasted about 30 to 40 minutes for each participant and the conversation was audio and video recorded for reporting purposes. The aim of this pre-interview is to obtain participant's background information, participant experience in using mobile phones and also their experiences in using any spiritual mobile applications. This information is essential, especially to know the level of exposure towards spiritual mobile applications usage among older adults in their daily lives.

- Diary Study. The diary study was conducted based on the method solicited from [15] which previously studied about older adults and hospitalization experience. After pre-interview session has completed, all participants from the workshop session was asked for their willingness to participate in the diary study. Each participant was asked in advanced for their willingness to participate in the study to prevent them from being burdened and this is to ensure that the collected data is reliable. Each of the older adult participants was provided with Tablet PC set with charger, a set of 3E diary, a pen, a user manual and a paper bag. The Tablet PC was equipped with Quran Bahasa Melayu application installed in it, meanwhile, the diary set consist of seven pieces of printed 3E diary template. User manual contains the guidance to the participants on how to start the Tablet PC until how to use Quran Bahasa Melayu application. The paper bag is also provided to keep all the probe kits and bring home for a week. Each participant's was asked to report their spiritual feelings in 3E dairy after each application's usage to capture their emotions. To keep encourage and remind the participant about writing the diary, researchers will contact the participants through phone call and short message service (SMS) at about 2 to 3 times a week as suggested by [15], [16]. The diary was collected back by researchers after a week through collecting them at participant' house or in complex. Each of the diary was interpreted by researchers before the post-interview session.

- Post-interview. Once the interpretation has completed in approximately two weeks to a month, a post-interview question was set up with each of the participant's. Post-interview objectives are to clarify the data gathered from the 3E diary and to explore participant's experience after using the application. This method is also done to avoid misunderstandings of the interpreted data by researchers. Each interview lasted in about 1 to 2 hours where the interview was done at participant's house or in complex. Further explanation on the UX evaluation tools used in this research is as discussed in previous paper [18].

4) UX model development. Once the data from pre-interview, diary, post-interview and observation were collected from the participants, thematic analysis is used to analyze all the data's. Seven steps adopted from [17] were used to analyze all data. Data were thoroughly processed step by step starting from familiarizing with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes and lastly producing the report. Triangulation which involves multiple sources of data is used to ensure the credibility of output.

Expert evaluation session with three different roles of informants which are the Islamic religious experts, the UX expert and the older adult users were implemented. All three Islamic religious experts have at least five years of experience in the field and has minimum Master's level of Islamic religious education. For UX expert, a senior researcher who currently engaging project pertaining user experience in the field of elderly and cultural was selected. Meanwhile, all three older adult users who have been recruited have at least five years' experience of using spiritual mobile applications. All experts were either emailed or face to face interviewed to review each of the themes and categories in the iSUX Model. Each of the expert panels were provided a list of operational definition for each elements in the iSUX model to understand further on the model. Panels then interviewed to give their opinion on each of the elements. Finally, all opinions and comments from expert panel were evaluated thematically and the iSUX model where then amended.

IV. Results and Discussion

A. Participants Feedbacks

All 11 participant's responses through pre-interview, diary study, post-interview and observation were analyzed using thematic analysis. The results were grouped into themes and categories (sub-theme) as following:

1) Effectiveness. Effectiveness theme consist of one category which is simplicity.

- Simplicity: is being defined as the quality or condition of system, product or services of being easy to use or understand. The simplicity element in HCI is widely known especially in evaluation and cannot be disregarded especially when designing with older adult users [7],[8]. Quran application was reported by participants to have simplicity element in terms of its function, presentation, content and appearance. Using Quran mobile application perceives as difficult at first, however, the perception was changed after usage.

2) Efficiency. Efficiency theme consists of two categories which are accessibility features and portability.

- Accessibility features: it provides evidence for reporting that one of the Quran mobile application's strength is on the application's function. Majority of older adults' participants reported in interview and 3E diary on the advantage of using Quran application such as the recitation audio. The audio recitation function allows elderly users to keep listening to Quran recitation despite doing other tasks such as doing house chores and while in vehicle.

- Portability: Portability can be defined as the ability of the application to be easily moved from one environment to another environment. The result shows that Quran spiritual mobile application is easy to be used and carry everywhere makes it easier to recite Quran even in car or airplane as well as in other places such as in hospital.

3) Learnability. Learnability could be defined as of how competent older adults' user may use Quran mobile application without too much effort. Result shows that some elderly users rely on Quran mobile application to learn Quran especially on the correct way to pronounce each of the verses. Furthermore, Quran recitation audio is very much helpful for Quran novice in learning Quran recitation.

4) Satisfaction. Satisfaction theme consists of two categories which are reliability and attractiveness.

- Reliability: The interviews and 3E diary proves that elderly users are aware and very cautious about Quran mobile application. Among aspect emphasized by participants including the accuracy of Quran application content especially on the spelling and punctuation. It shows that the participants are very cautious upon choosing the correct version of Quran applications from application store.

- Attractiveness: Cognitive and vision impairment caused elderly to hardly focusing on devices for long duration. Cognitive ability to focus in longer duration decreased whereas vision ability like eyes sensitivity becomes less. All these changes should be taken into account while delivering an application for older adult users. For instance, in the interview session, an elderly participant did address on the importance of using colourful text for the Quran application to avoid from feeling boredom.

5) Sublimity. Sublimity theme consists of two categories which are transcendence and peacefulness.

- Transcendence: Majority of participants addressed on the subjective experience related to Islamic spirituality that meet the criteria of transcendence. Codes which are identified as the feeling of transcendence are centered on a sense of divine. Transcendence in this study relates to the context of the relationship with God or transcendent being, individual transcendence as well as relationship with others including the nature. Participants shared the inner feeling of connection to God while engaging Quran Bahasa Melayu application by addressing the servitude and remorse feelings, being guided, feel fear of God through diary and interview.

- Peacefulness: The definition of peacefulness as according to [12] is when there happen to be a separation of threats either in physical or psychology with the real world. The emotional states of calmness and feeling relaxed were reported frequently in respond to engage Quran mobile application among elderly participants. Reciting Quran using Quran application offers positive emotional effect as reported by participants in the diary where it is able to soothe their mind while releasing stress.

6) Vitality. Vitality theme consists of two categories which are energy and joyful activation.

- Energy: Reciting Quran application is found to be uplifting while boosting participants' energy. Specifically, participants spoke about their eagerness to use Tablet PC, which eventually motivated participants to recite Quran often using mobile application. In fact, one participant indicated in the diary that with reciting Quran using application is capable to eliminate the unwilling and laziness feeling to recite Quran.

- Joyful Activation: Several participants who shared similar positive experiences noted that the Tablet PC encourages their curiosity to recite Quran up to long duration. In addition, one participant stressed that it is very happy to recite using application for it has many good features for instance, the tajweed marker and the translation which resulting in a better understanding of the Quran meaning.

The following Table III shows several older adult users feedbacks from the collected data. All users' feedbacks were analyzed thoroughly to produce meaning.

TABLE III. PARTICIPANT'S FEEDBACKS

Participant's ID	**Feedback**	**Category**	**Theme**
P7	It is a good and **simple application**. Easy to use and like to use it.	Simplicity	Effectiveness
P11	I never used a tablet and IT. At first I thought it is very hard, but **it is OK**.		
P2	It is fun to use the tablet, we just listen to it when we switched it ON. It will play and we just listen to it… **if we switched it ON we can do other work**.	Accessibility features	Efficiency
P4	After *Isyak* prayer, ON the tablet and listen to *surah Al-Maidah*. I often **use a tablet while I am in a vehicle**, while appreciating its translations.		
P4	For myself, I like to **use the table when I am outside of the house** - for example **in a car** while travelling. Also while relaxing. Can be use **while cooking** also.	Portability	
P9	With tablet, if we do not know, we just **turn ON the sound and follow**; with the book, right or wrong, nobody can correct it. Those who do not know how to read can also learn	Learnability	
P11	Yes, sometimes I think that the **Quran need to be checked**. For example, when it has been printed, there might be a missed mark.	Reliability	Satisfaction
P11	The **letters have colours**…because we are old and it catches our interest. If black and white, its look normal...older people can be that long.	Attractiveness	
P2	If I switched it ON, it feels sad because we are listening to people reading, feel enjoyable… feel "high". Feels like we are interacting with Allah. **Like we are talking to Allah. The message is absorbed into the soul.**	Transcendence	Sublimity
P5	A deep feeling. More focused, more concentrate. By using the tablet, it **feels more close to Allah**… **Feel the fear to Allah, horrified**.		
P4	After listening to the Quran – **the heart and soul feel calm.** No negative thoughts. Only good thinking.	Peacefulness	

P6	**Does not get bored** using the tablet even though already read many verses.	Energy	Vitality
P2	**Enjoy reading** until I realize it has been quite some time and feel tired.	Joyful Activation	
P9	**Fun to use**, just switch it ON.		

Findings from the research presents ten spiritual user experience categories identified for older adults as shown in Fig. 5.

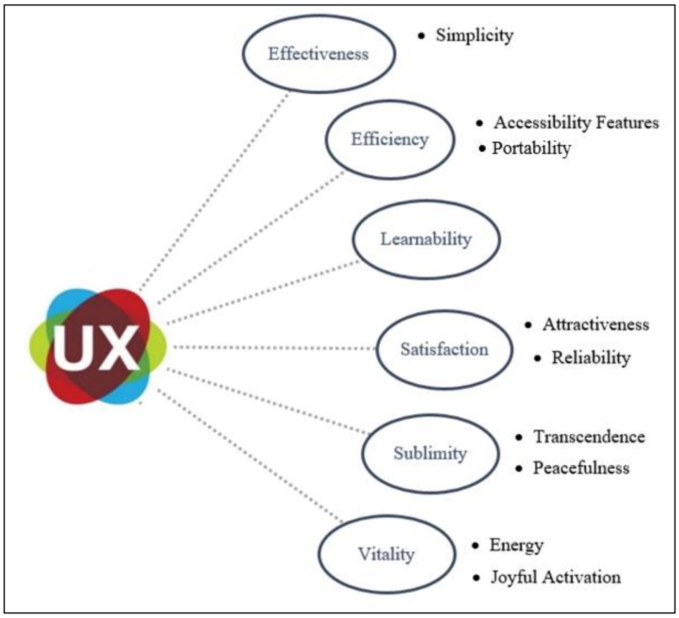


Fig. 5. List of Identified Categories and Themes.

The lists of the categories are simplicity, portability, accessibility features, learnability, reliability, attractiveness, transcendence, peacefulness, energy and joyful activation. These ten categories were grouped into six spiritual user experiences themes for older adults namely; efficiency, effectiveness, learnability, satisfaction, sublimity and vitality.

B. Spiritual User Experience (iSUX) Model

iSUX Model can be categorized into three central user experience (UX) components as highlighted by [9] which are the instrumental qualities, non-instrumental qualities and the emotional reactions. The first component, which is the instrumental qualities focused on user experience towards the ease of use and functionality of the application. This component is related to the usability aspects of using spiritual applications. The second component, which is non-instrumental qualities, focus on the look and feel of the application in which this component is related to the subjective user experience of the applications. Non-instrumental qualities concerns are on the affect and emotional aspects, where it talks around the evocation of emotions while using the spiritual applications. Finally, the third component, which is the emotional user reactions is actually the emotional responses from users while using the spiritual applications. Based on the research findings, it can be concluded that spirituality elements exist in both instrumental qualities and non-instrumental qualities components as depicted in Fig. 5. The older adults use spiritual mobile application with the purpose to fulfill their spirituality needs. From the non-instrumental qualities aspect, the activity of using the spiritual mobile applications observably can evoke the spiritual emotions. One can experience joy when the application is unexpectedly easy to use or when the older adult's successfully operating the application. Older adults will respond emotionally especially when they have personal spirituality values such as crying when they listen to the verses translation pertaining the day of judgment, heaven and hell. While from the instrumental qualities aspect, it shows that the accessibility features of spiritual application helps the older adults a lot in using the technology. The portability of spiritual mobile application makes older adults easy to recite Quran everywhere even at the public places or while in the vehicles.

The perception of both qualities influences the third component of UX which is the emotional user reactions. For example, the audio recitation and text resizing function in spiritual mobile application may affect perceived effectiveness and lead to motivation to use the application in future. In addition, a similarity of mobile Quran design with the original Al-Quran may impact on transcendental and cause emotional responses such as fear to God, calm and sense of conviction. In conclusion, the ability of designing and developing spirituality applications that could evokes positive experiences can be increased by formulating spiritual value profiles in the early stage. Therefore, the iSUX Model is proposed to discover spiritual elements to help pave the way for a study of spirituality in the domain of user experience in future.

V. Conclusion

This research is hoped to provide significance understanding to the body of knowledge by amplifying the relatively limited research studies done on spiritual mobile applications towards the researchers in the field of HCI. In this research, the model for spiritual user experience was presented. This research identifies key elements of spirituality for user experience with based on the data gathered from the older adults. The model namely iSUX was then developed based on the data findings supported with expert evaluation as well as theoretical views from the field of musical and theological. Even though spirituality elements fall under both qualities in iSUX model, however, the non-instrumental quality is more prominent where it discusses more about the derivation of emotions to the Creator. Emotional responses can result through physical outwardities such as crying, happy, calm and scared. Further investigation on subjective spiritual user experience is very interesting to explore more. This iSUX model is useful to practitioners and researchers as it provides spiritual user experience elements that may be considered in developing spirituality application for older adult users. Designing the needs of older adult users is perceived designing for all.

Indeed, this research was conducted in an emerging and important domain, within the area that is still in infant state and need attention from numerous HCI scholars to embark research on spirituality. In conclusion, it is hoped that this research could provide significance contribution to the body of knowledge and can act as a sound basis for further research in spiritual user experience. Finally, it is also researcher's hope that this study could inspire new ideas that in turn create a spark of more research in spirituality domain in future.

ACKNOWLEDGMENT

This work was supported by Universiti Teknologi MARA (UiTM), and also funded by the Research and Innovation, Universiti Kuala Lumpur. Special thanks to all older adult participants from Kompleks Warga Emas, Seksyen 24, Shah Alam for their willingness and cooperation in this study.

REFERENCES

[1] M. Ahmad and S. Khan, "A Model of Spirituality for Ageing Muslims," J. Relig. Health, no. March 2015, p. 14, 2015.

[2] E. Buie and M. Blythe, "Spirituality: there's an app for that! (but not a lot of research)," CHI'13 Ext. Abstr. Hum. Factors …, pp. 2315–2324, 2013.

[3] M. Aliyu, M. Mahmud, A. O. Md Tap, and R. Mohammad Nassr, "A Preliminary Investigation of Islamic Websites' Design Features that Influence Use: A Proposed Model," Electron. J. Inf. Syst. Dev. Ctries., vol. 58, no. 5, pp. 1–21, 2013.

[4] E. Buie, "Transcendhance: A Game to Facilitate Techno-Spiritual Design," Proc. CHI 2016, pp. 1367–1374, 2016.

[5] N. A. Ahmad, "Islamic Spiritual User Experience (iSUX): A Case Study of Muslim Older Adults using Al-Quran Mobile Application," Universiti Teknologi MARA, 2018.

[6] N. Tohit, C. J. Browning, and H. Radermacher, "'We want a peaceful life here and hereafter': healthy ageing perspectives of older Malays in Malaysia," Ageing Soc., vol. 32, no. 03, pp. 405–424, May 2011.

[7] O. V. Bitkina, H. K. Kim, and J. Park, "Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges," Int. J. Ind. Ergon., vol. 76, pp. 1–11, 2020.

[8] R. Pinto, "Study of User Experience Design of Digital Financial Services," 2020.

[9] M. Thuring and S. Mahlke, "Usability, aesthetics and emotions in human-technology interaction," Int. J. Psychol., vol. 42, no. 4, pp. 253–264, 2007.

[10] S. Mahlke, "Visual aesthetics and the user experience," Study Vis. Aesthet. Human-Computer Interact., no. 2000, 2008.

[11] M. Hassenzahl and N. Tractinsky, "User experience - a research agenda," Behav. Inf. Technol., vol. 25, no. 2, pp. 91–97, 2006.

[12] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement.," Emotion, vol. 8, no. 4, pp. 494–521, 2008.

[13] Department of Statistics, "Anggaran Penduduk Semasa, Malaysia, 2018-2019," 2019.

[14] N. A. Ahmad, A. Zainal, F. Hanis, A. Razak, W. Adilah, and W. Adnan, "A Pilot Study of Using Diaries Method for Collecting Spiritual Experiences Data Among Older Adults," ARPN J. Eng. Appl. Sci., vol. 10, no. 23, pp. 17690–17697, 2015.

[15] C. S. Jacelon and K. Imperio, "Participant diaries as a source of data in research with older adults.," Qual. Health Res., vol. 15, no. 7, pp. 991–7, Sep. 2005.

[16] D. Zimmerman and D. Wieder, "The Diary: Diary-Interview Method," Urban Life, vol. 5, no. 4, pp. 479–498, 1977.

[17] V. Braun and V. Clarke, "Using thematic analysis in psychology Using thematic analysis in psychology," Qual. Res. Psychol., no. October 2012, pp. 37–41, 2008.

[18] N. A. Ahmad, A. Zainal, F. H. Abdul Razak, W. A. Wan Adnan, and S. Osman, "User Experience Evaluation of Mobile Spiritual Applications for Older People: An Interview and Observation Study," J. Theor. Appl. Inf. Technol., vol. 72, no. 1, pp. 76–85, 2015.

Reversible Data Hiding using Block-wise Histogram Shifting and Run-length Encoding

Kandala Sree Rama Murthy[1], V. M. Manikandan[2]
Department of Computer Science and Engineering
SRM University-AP, Andhra Pradesh, India

***Abstract*—Histogram shifting-based Reversible Data Hiding (RDH) is a well-explored information security domain for secure message transmission. In this paper, we propose a novel RDH scheme that considers the block-wise histograms of the image. Most of the existing histogram shifting techniques will have additional overhead information to recover the overflow and/or the underflow pixels. In the new scheme, the meta-data that is required for a block is embedded within the same block in such a way that the receiver can perform image recovery and data extraction. As per the proposed data hiding process, all the blocks need not be used for data hiding, so we have used marker information to distinguish between the blocks which are used to hide data and the blocks which are not used for data hiding. Since the marker information needs to be embedded within the image, we have compressed the marker information using run-length encoding. The run-length encoded sequence is represented by an Elias gamma encoding procedure. The compression on the marker information ensures a better Embedding Rate (ER) for the proposed scheme. The proposed RDH scheme will be useful for secure message transmission also where we are also concerned about the restoration of the cover image. The proposed scheme's experimental analysis is conducted on the USC-SIPI image dataset maintained by the University of Southern California, and the results show that the proposed scheme performs better than the existing schemes.**

Keywords—Histogram shifting; run-length encoding; secure message transmission; overflow; Elias gamma

I. Introduction

Data hiding is a well-explored information security area in which a hidden message can be safely transmitted by embedding it in a digital media mask [1, 2]. Digital images are often used as a cover medium for data hiding purposes. The conventional data hiding scheme makes permanent changes to the cover image pixel values during data hiding process and at the receiver side, only the hidden message will be extracted and not concerned about the cover image. The image obtained after data hiding is termed as stego image. The sender embeds the secret message into the cover image and communicates the stego image and a key for extraction to the sender. If the receiver can extract the hidden message and recover the original image completely then the scheme is called Reversible Data Hiding (RDH) [3]. For the past two decades, RDH schemes are widely studied and a number of algorithms are proposed in this domain. The overview of a RDH scheme is graphically shown in Fig. 1.

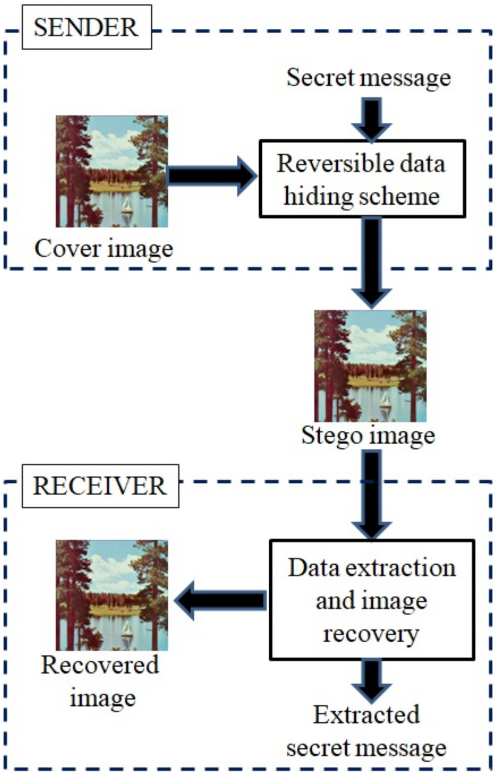


Fig. 1. Overview of RDH.

II. Literature Review

Most of the existing RDH schemes are centered around the idea of compressing a bit plane of image to create extra space for the secret message [3-5]. Difference expansion techniques [6-8] and histogram shifting based approach is discussed in [9-13]. There are a number of algorithms that use entirely different approaches for data hiding are also available in the literature [14-16].

In medical image transmission, patient records can be securely sent along with medical images using RDH methods which enable us to recover both the image and message without any loss of data. The reversible watermarking schemes are widely used for authenticating medical images in which a watermark will be embedded in the medical image instead of a secret message. The cloud service providers commonly use RDH schemes to hide some additional metadata on the images uploaded by the clients.

In this research, we explored a histogram shifting based RDH scheme and introduced a block-wise histogram shifting based RDH scheme with a better embedding rate by compressing the marker information using run-length encoding. The marker information is additional information that helps to distinguish the blocks which are used for data hiding and the blocks which are not used for data hiding.

For a better understanding of the proposed scheme, the basic histogram shifting based RDH scheme is detailed in Section III. The proposed scheme is detailed in Section IV. The experimental study and result analysis is detailed in Section V. A few other alternatives that we tried along with the proposed scheme are briefed in Section VI. The conclusion and a few insights to future work are given in Section VII.

III. Preliminaries

In this section, we briefly discuss the basic histogram shifting based RDH scheme [9] since we extended the same in this manuscript. An overview of the block-wise histogram shifting based RDH scheme introduced in [13] is also discussed in this section for better understanding of the proposed scheme.

```
Algorithm I: Basic histogram shifting based RDH algorithm

Input   :  Cover image $I$ and the secret message $D$
Output  :  Stego image $S$
Step 1  :  Find the histogram $H$ of the cover image I
Step 2  :  Identify the peak pixel value $P$ from the histogram $H$
Step 3  :  Increment all the pixels which are greater than $P$ by one to get the new image $S$. No need to do any changes on the pixels having intensity value P or less than $P$.
Step 4  :  Read the pixels in $S$ in a predefined order (either row-wise linear order or in a pseudo-random order based on a data hiding key) and if we find a pixel at location ($X$, $Y$) with the pixel value $P$ in which we can hide one bit $Q$ from the secret message $D$. Add the $Q$ value with the pixel at location ($X$, $Y$) to perform data hiding. Note that the new pixel value will be ($P$+1) if we hide the secret bit 1 and the pixel will be unchanged if we embed secret bit 1. Do this process to embed all the bits from secret message to get the final stego image $S$.
Step 5  :  Return the stego image $S$
```

The data extraction and image recovery process in histogram shifting based RDH process is given in Algorithm II.

The major challenges with the basic histogram shifting based RDH scheme are listed below:

- Embedding Rate (ER) (the number of bits that can be embedded per pixel) is very low. For more information please refer to section IV. a.
- Overflow is quite common when the original cover image contains some pixel value 255.

```
Algorithm II: Data extraction and image recovery process

Input   :  Stego image $S$, the pixel value $P$ used for data hiding
Output  :  Recovered image $I$ and the extracted message $D$
Step 1  :  Read one pixel $K$ from $S$ in a predefined order. Do the following to extract the hidden message $D$ and recover the original image I:
           Case 1: $K<P$
                   Just keep $K$ as it is in the recovered image $I$.
           Case 2: $K==P$
                   Extract a bit 0 and append it with the secret message $D$ and no need to do any change for the pixel value keep as it is in the recovered image $I$.
           Case 3: $K==P+1$
                   Extract a bit 1 and append it with the secret message $D$ and to get the recovered pixel value, decrement the pixel value by one and keep it in recovered image $I$.
           Case 4:$K>P+1$
                   Decrement the pixel value by one and keep it in $I$ to recover the original pixel value and do not extract any secret message bit.
Step 2  :  Return the recovered image $I$ and extracted secret message $D$
```

A scheme is introduced in [17] to handle the overflow in the histogram shifting based on RDH. A block-wise histogram shifting based scheme is introduced by us [13] in which the overflow handling technique is used. In addition to that, block-wise histogram shifting based approach is used to improve the ER. The main idea is that instead of generating the histogram of the whole image, we divided the image into fixed size blocks and applied the histogram shifting algorithm on each block. The embedding rate in the histogram shifting algorithm is proportional to the peak value of a pixel, and by dividing the image into blocks, the overall embedding rate is increased significantly without affecting the visual quality of the stego image. We also eliminated the need of sending a key separately to the receiver by embedding the key, which is the peak pixel value, within the first a few pixels of each block using the least significant bit substitution.

A major hurdle in histogram shifting is the problem of overflow bits. When we are shifting the pixels by one position, the end pixel value can't be shifted ahead. The solution we provided is to make a marker list of all the end pixel values before and after shifting and store this information within each block by embedding it along with the message. On the retrieval of the message, this marker information can be utilized to distinguish the overflow bits and deal with them separately. To store this marker information within a block along with a portion of the secret message, each block's peak value must be greater than the number of overflow bits plus the binary size of the key. This condition disqualifies a few blocks to be used for data embedding. The marker information is embedded in the first block of the image and thus the entire first block is omitted in the rest of the algorithm.

The scheme discussed in [13] divides the original image into blocks of size *B*×*B* and we need to find the histogram of each block. If the peak in the histogram of the selected block is capable to provide a sufficient embedding rate to hide the peak pixel value, overflow information, and at least one secret message bit then it will be treated as a block suited for data hiding. All the blocks will be verified at the beginning itself to generate marker information in which 0 corresponded to a block indicates that the block is not usable for data hiding and 1 indicates that the block can be used for data hiding.

Let us assume that the original image is of the size *R*×*C* pixels then the original image will have *N* non-overlapping blocks of size *B*×*B* where *N* is defined as follows:

$$N = \left\lfloor \frac{R}{B} \right\rfloor \times \left\lfloor \frac{C}{B} \right\rfloor \tag{1}$$

All remaining (N-1) blocks have been used for data hiding in the current scheme except for the first block, and the size of marker information M will be N-1. The marker bits are embedded in the first block of the cover image by replacing the N-1 least significant bits (LSB) from the first N-1 pixels in the first block of the cover image. It may be noted that at the receiver side, we must recover the original image as it is. For this purpose, the LSBs of the first *N*-1 pixels are embedded in the image itself while hiding the secret message. If there is a possibility to reduce the number of bits in the marker information then it will help to improve the actual embedding rate. It should be noted that the actual embedding rate is defined based on the number of secret message bits that we can embed in the image without considering all the other overhead such as embedding marker information, the embedding of overflow information, etc.

The proposed scheme explored the possibility of reducing the number of bits required for marker information. During the existing scheme's experimental study, it is observed that in most of the images most of the blocks are used for data hiding. In such cases, the marker information will be a sequence of bits in which very rarely a bit 0 will be present. Note that if the J^{th} bit is 0 in the marker information, it indicates that the J^{th} block is not used to hide data.

In this manuscript, we propose an RDH scheme that compresses the marker information using run-length encoding. The run-length encoded sequence is encoded using a variable length encoding scheme called Elias gamma. The run-length encoding sequence is popularly used to compress the data when it contains redundant information. We observed that the marker information is always having a high amount of redundancy (1's are coming consequently as most of the blocks are used for data hiding) which motivated us to use run-length encoding to compress the marker information. The run-length sequence should be converted into a sequence of bits and we have used Elias gamma encoding process in the proposed scheme. The Elias gamma is a well-known variable length encoding scheme. For better understanding, the run-length encoding process is briefly described here with an example. Let us assume *C* is a bit sequence as follows:

C: 1000000000000000111111100000000000000000

The corresponding run-length sequence *L* will be

L: 1, 1, 15, 7, 17

In run-length sequence we will keep the starting bit of the *C* as it is. In the above example the first bit is 1, so we kept it as it is in the run-length sequence. The second value 1 indicates that 1 is repeating 1 time in *C*. The third value 15 indicates that 0 is repeating 15 times in *C*. The next value 7 indicates that 1 repeats 7 times in *C* and so on.

The run-length encoded sequence *L* can be converted into sequence of bits *E* by using Elias gamma encoding scheme. The Elias gamma encoding scheme is given Algorithm III. The input will be run-length sequence and the output will be *E*.

```
Algorithm III: Elias gamma encoding on run-length sequence
Input    :  Run-length sequence $L$ which consists of $N$ values
Output   :  Elias gamma encoded sequence $E$
Step 1   :  Take the starting bit value (first value from $L$) and
            keep it in $E$
Step 2   :  $X$=1
Step 3   :  While $X<N$
Step 4   :  $D=L_X$ //$X^{th}$ value from $L$
Step 5   :  Convert $D$ into binary, say $B$
Step 6      Find the number of bits $K$ in $B$
Step 7   :  Append ($K$-1) 0's to E then append $B$ to $E$.
Step 8   :  $X=X+1$
Step 9   :  EndWhile
Step 10  :  Return the Elias gamma encoded sequence $E$
```

If we apply the run-length encoding procedure on the run-length sequence that we have obtained in the previous step, the corresponding binary sequence *E* will be the following:

E: 11000111100111000010001

The total number of compressed data bits is 23 and we achieved a compression ratio of 30/23=1.30.

We have used run-length encoding with Elias gamma to compress the marker information. Run-length encoding is a lossless compression scheme that will help us to decompress the information without any loss.

In the next section, the proposed algorithms are given where we have used all the basic concepts we have discussed here.

IV. Proposed Scheme

This segment addresses the proposed scheme. The sequence of operations using the run-length encoding scheme in the proposed block-wise histogram-shifting based RDH is defined:

Algorithm IV: Proposed histogram shifting based RDH using run-length encoding

```
Input   : The cover image $I$ and the secret message $D$
Output  : The stego image $S$
Step 1  : Divide the cover image into blocks (non-overlapping) of size $B×B$ pixels.
Step 2  : Access one block at a time and check whether or not the block is suitable for hiding data. To check this for the $K^{th}$ block, compute the histogram of the block and find the peak intensity $P_K$ value and its corresponding count $C_K$.
          If $C_K>F+8$
                $M_K=1$ //$K^{th}$ marker bit
          Else
                $M_K=0$
          Where $K=1, 2,…N-1$ and $F$ is the number of 255's or 254's in the $K^{th}$ block which indicates the overflow information required. 8 bits are needed to keep the LSB's of the first 8 pixels which will used to store the peak intensity value.
Step 3  : Apply the run-length encoding process on the marker information $M$ to get the run-length run-length sequence $L$
Step 4  : Apply Elias gamma encoding procedure on $L$ to get compressed binary sequence $E$
Step 5  : Find the length of $E$, say $Z$
Step 6  : Access the first block and extract Z LSB's from the first Z pixels, say T is the list of Z LSB's.
Step 7  : Replace the LSB's of first $Z$ pixels by the bits in $E$ and keep the modified block as the first block of the stego image $S$.
Step 8  : The secret message $D$ should be appended at the end of $T$
Step 9  : Access the $K^{th}$ block $G_K$ for data hiding purpose in the predefined order. Find the peak pixel value from the block $G_K$ by considering the histogram of the image block and store the 8-bit peak pixel binary value at the LSB position of the first 8 pixels of the block, if and only if the block is suitable for hiding the data. The LSBs of the first 8 pixels W, along with the overflow information, should be embedded in the same block itself. The overflow handling bits $O$ are computed by traversing the whole block one by one and the 0 bit will be used to mark the pixels originally with 254 and 1's will be used to mark the pixels with overflow (originally 255).
          The same process that is described in Algorithm I is used for data hiding but the only difference is that we will be applying the same thing on a block (except the first 8 pixels) and the bits that are going to hide consists of LSB information of the first 8 pixels, the overflow information and some bits from actual secret message. The altered GK blocks will be positioned at the corresponding position in I.
Step 10 : Step 9 should be repeated for all the blocks in the image $I$ to get final stego image $S$
Step 11 : Return the stego image $S$.
```

The sender just needs to send the stego image S to the receiver for data retrieval and image recovery during the data hiding. The marker information about the order followed during data hiding is embedded in stego image. In our experimental study, we have followed row-wise linear order during data hiding phase. The image recovery and data extraction is possible from the stego image S. The process is described in Algorithm V.

Algorithm V: Image recovery and data extraction from the proposed scheme

```
Input   : The stego image $S$
Output  : The extracted message $D$ and the recovered image $I$.
Step 1  : Divide the stego image $S$ into blocks (non-overlapping) of size $B×B$ pixels.
Step 2  : Find the number of non-overlapping blocks $N$ in the image $S$.
Step 3  : Keep on extracting the LSB's of the pixels from the first block and apply the Elias gamma decoding and run-length decoding procedure on the fly. Continue this process until we are getting a bit sequence $M$ having size ($N$-1). This will be the actual marker information which is used identify the blocks used for data hiding. Let us assume that we have accessed $V$ number of pixels to get a bit sequence $M$ of $N$-1 bits.
Step 4  : Access the $K^{th}$ block $G_K$ from stego image S and process it if $M_K$ is 1. Otherwise just ignore it.
Step 5  : If the $G_K$ is used during data hiding then extract the first 8 bits from the first 8 pixels and convert into corresponding binary $P_K$. Apply data extraction and image recovery process (almost same as Algorithm II) from the block $G_K$ (excluding first 8 pixels). The first 8 bits from the extracted message should be used to recover the first 8 pixels by replacing the LSBs. Next $W$ number of bits $O$ will be used to retrieve the pixels having value 255. The size of the $W$ will be nothing but number of 255's in the current block $G_K$. The bit 0 in $O$ says that the 255's should be decremented by one and the bit 1 in $O$ says that the 255 should keep as it in the recovered image. The remaining bits extracted from this will be keep on appending with a bit sequence $D$'. Repeat this process for all the blocks in the image and store the values in $I$.
Step 6  : Extract the first $V$ number of pixels from $D$' and replace it in the LSB's of first $V$ pixels to recover the first block and keep the block as the first block of $I$.
Step 7  : Remove the $V$ number of bits from $D$' to get $D$ which will be the actual extracted message.
Step 8  : Return extracted message $D$ and recovered image $I$.
```

V. Experimental Study and Result Analysis

The experimental research and analysis of outcomes is performed on every image from the USC-SIPI image dataset

[18] managed by the Southern California University. Since the dataset consists of both color images and grayscale images of various sizes, to attain uniformity we have converted all images into 8-bit grayscale images of size 512×512 pixels. Our algorithm can be implemented on color images as well without any preprocessing. In all the images we have embedded the maximum possible number of bits as secret message. A pseudo-random bit sequence is generated as the secret message. The USC-SIPI image dataset has four different categories of images: aerials, textures, sequences and miscellaneous (misc.). The optimum block size is empirically decided as 16×16 pixels and all the experimental study of the proposed scheme is performed by taking this block size. We have conducted an experimental with various block sizes such as 128×128, 64x64, 32x32, 16x16 and 8x8 and picked the block size which provides the maximum ER. We have analyzed the following efficiency parameters during the study:

1) Embedding rate
2) Peak signal to noise ratio (PSNR)
3) Structural similarity index (SSIM)
4) Natural image quality evaluator (NIQE)
5) Blind image spatial quality evaluator (BRIQUE)

A. Analysis of Embedding Rate

Embedding rate (*ER)* is defined as follows:

$$ER = \frac{K}{R \times C} \tag{2}$$

where K is the maximum number of bits in the message that can be embedded in a R×C size image. The embedding rate is usually denoted by bits per pixel (bpp). The embedding rate of the scheme proposed is purely dependent on the distribution of pixels in the image. In the proposed scheme we are utilizing some of the actual embedding possibilities to hide the pixel recovery information. So we basically considered the effective embedding rate by deducting all the overhead information. Table I provides the average embedding rate obtained from all four types of images from the proposed scheme.

If the image consists of too much smooth regions with less possibility of overflow then it will be capable to provide a high embedding rate. From Table I, it can be observed that the average ER from all four categories of images is better than the existing schemes. Some additional overhead bits are reduced in the proposed scheme through the use of run-length encoding process. The embedding rate from the four well-known images such as airplane, boat, baboon and peppers is given in Table II. It also shows the embedding rate from the existing schemes [9], [13].Table II shows that for all the four well-known images, the embedding rate of the proposed scheme is better than the embedding rate of the existing schemes mentioned in [9] and [13].

B. Analysis of PSNR and SSIM

Two image quality measurement tests are the PSNR and SSIM values, which help to analyze the image's quality deterioration based on a reference image. The low quality images maybe an indication of some hidden images and the attackers may try to do steganalysis on such images to extract the hidden messages. So the researchers are very much concerned about the stego image quality while proposing new RDH scheme. If the original image is exactly same as the stego image, then PSNR will be ∞ and SSIM will be 1. But this will not happen since we are supposed to do some alteration in the pixels to hide the secret message and such changes will lead to reduction in PSNR and SSIM measure. If the PSNR and SSIM measure are high it indicates that there is low quality degradation on the stego image due to data hiding process. Table III presents a comparison between the PSNR for well-known images from the existing schemes and the proposed scheme.

The average SSIM value for all four categories of images from the existing schemes and the proposed scheme is given in Table IV.

Table V contains the SSIM values from the existing schemes and the proposed scheme for well-known images.

TABLE I. COMPARISON OF AVERAGE ER FROM ALL CATEGORY OF IMAGES

Image category	**Scheme in [9]**	**Scheme in [13]**	**Proposed scheme**
Misc.	0.0388	0.1081	0.1089
Textures	0.0230	0.0521	0.0529
Sequences	0.0266	0.0765	0.0774
Aerials	0.0252	0.0588	0.0597

TABLE II. COMPARISON OF ER (BPP) FOR WELL-KNOWN IMAGES

Image category	**Scheme in [9]**	**Scheme in [13]**	**Proposed scheme**
Baboon	0.0105	0.0201	0.0210
Peppers	0.0105	0.0438	0.0447
Boat	0.0221	0.0442	0.0451
Airplane	0.0317	0.0903	0.0912

TABLE III. COMPARISON OF PSNR (IN DB)

Image name	**Scheme in [9]**	**Scheme in [13]**	**Proposed scheme**
Baboon	50.4139	50.9370	50.9405
Peppers	50.0807	51.0778	51.0815
Boat	51.9712	50.9220	50.9259
Airplane	53.9372	52.2065	52.2121

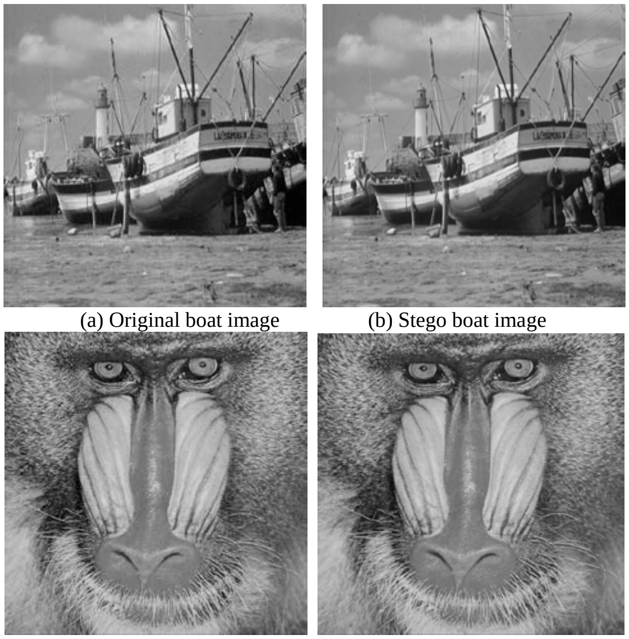
TABLE IV. COMPARISON OF AVERAGE SSIM FROM PROPOSED SCHEME AND EXISTING SCHEMES

Image category	**Scheme in [9]**	**Scheme in [13]**	**Proposed scheme**
Misc.	0.9993	0.9987	0.9987
Textures	0.9996	0.9994	0.9994
Sequences	0.9997	0.9989	0.9989
Aerials	0.9996	0.999	0.999

TABLE V. COMPARISON OF SSIM FROM PROPOSED SCHEME AND EXISTING SCHEMES FOR WELL-KNOWN IMAGES

Image category	Scheme in [9]	Scheme in [13]	Proposed scheme
Baboon	0.9998	0.9996	0.9996
Peppers	0.9998	0.9989	0.9989
Boat	0.9996	0.9991	0.9991
Airplane	0.9996	0.9987	0.9987

The results shown in Table III, Table IV and Table V indicate that from all the well-known images we are getting a very high PSNR and SSIM measure. In general, the data hiding schemes that can generate stego image with a PSNR of 50 dB or more is treated as a good scheme. The SSIM values are very close to 1 also. Two sample cover image and the corresponding stego image obtained after data hiding process is given in Fig. 2.



(a) Original boat image (b) Stego boat image

(c) Original baboon image (d) Stego baboon image

Fig. 2. Sample Results.

C. Analysis of NIQE and BRISQUE

The NIQE and BRISQUE are two well-known no-reference image quality measures [19, 20]. In addition to the reference based image quality assessment techniques such as PSNR and SSIM, we have analyzed the quality of the stego images using no-reference image quality assessment approaches such as NIQE and BRISQUE. .In Fig. 3 and Fig. 4, the NIQE and BRISQUE measures of well-known images obtained from the proposed scheme and the existing schemes are given.

In Fig. 3, NIQE-1 indicates the NIQE measure from the original image and NIQE-2 is the NIQE measure from the stego image. A high NIQE measure means that the image quality is good.

The data shown in Fig. 3 and Fig. 4 indicates that the stego image quality is not much deviated from the quality of the original image.

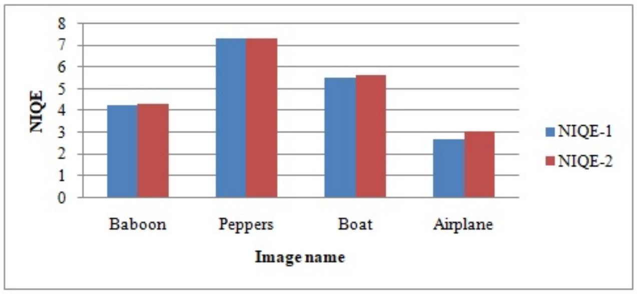


Fig. 3. NIQE Measure from Original Image and Stego Image.

The BRISQUE image quality measures of original image and stego image are given in Fig. 4. It may be noted that the low BRISQUE measure indicates high image quality. In Fig. 4, BRISQUE-1 is the BRISQUE measure from the original image and BRISQUE-2 is the stego image's BRISQUE measure.

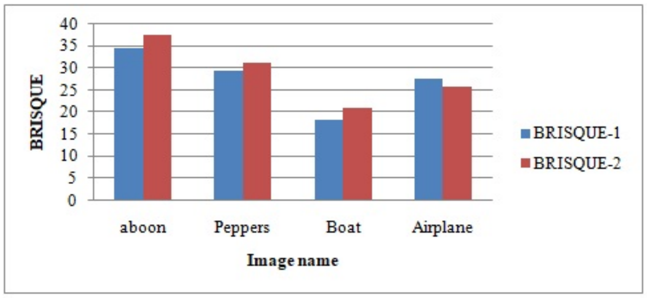


Fig. 4. BRISQUE Measure from Original Image and Stego Image.

VI. ALTERNATE APPROACHES FOR IMPROVEMENT

In addition to the Algorithms discussed in Section II, we have attempted a few other alternatives to improve the proposed scheme. The histogram of each block can also be left-shifted instead of the default right shift operation based on the following considerations:

1) The number of pixel values less than the peak pixel value is less than the number of pixels greater than the peak value. This demands us to change fewer pixel values and thus the stego image quality will be improved.

2) The number of underflow marker bits (0's and 1's) is less than the number of overflow marker bits (255's and 254's). This reduces the marker bits that need to be embedded along with the message and thus the embedding rate will increase.

In all the methods we have proposed so far, we have omitted the entire first block of the image for storing marker information. Since we have compressed the number of marker information that needs to be stored in the first block, we no more need nearly more than half of the first block for storing the marker bits. Hence we can also utilize the first block for data hiding which will marginally improve the embedding rate. We have implemented the above-mentioned techniques also but didn't incorporate them with the proposed algorithm as there were only marginal and mixed improvements in results.

VII. Conclusion

A block-wise histogram shifting based RDH scheme with a high embedding rate is proposed in this manuscript. In the proposed scheme, marker information is generated by checking the suitability of a block for data hiding and this information is compressed by using run-length encoding with Elias gamma encoding procedure. In order to facilitate the data extraction and image recovery process, the Elias gamma encoded marker information is embedded in the first block of the cover image. A better embedding rate without losing the visual quality of the stego image is attained in the experimental analysis of the proposed scheme. The visual quality of the stego images is measured using a metric of reference image quality, such as the peak signal to noise ratio and the structural similarity index. Most of the stego images gave a PSNR value greater than 50 dB and SSIM values close to 1. In addition, the visual accuracy of the stego image is measured using non-reference image quality tests such as natural image quality evaluator and blind image spatial quality evaluator, and the results show that the quality of the stego image does not differ significantly from the quality of the original image. The future works can be focused to apply histogram shifting on the histogram of regions generated through segmentation rather than on the non-overlapping blocks.

Acknowledgment

We kindly acknowledge the endless support received from the parent institute for successfully completing the project.

References

[1] Bender, Walter, et al. "Techniques for data hiding." IBM systems journal 35.3.4 (1996): 313-336.

[2] Swanson, Mitchell D., Bin Zhu, and Ahmed H. Tewfik. "Robust data hiding for images." 1996 IEEE Digital Signal Processing Workshop Proceedings. IEEE, 1996.

[3] Celik, Mehmet Utku, et al. "RDH." Proceedings. International Conference on Image Processing. Vol. 2. IEEE, 2002.

[4] Abbasi, Rashid, et al. "Efficient lossless compression based RDH using multilayered n-bit localization." Security and Communication Networks 2019 (2019).

[5] V. M. Manikandan, V. Masilamani, "A Novel Bit-plane Compression based RDH Scheme with Arnold Transform", International Journal of Engineering and Advanced Technology, vol. 9, issue 5, pp. 417-423, 2020.

[6] Tian, Jun. "Reversible data embedding using a difference expansion." IEEE transactions on circuits and systems for video technology 13.8 (2003): 890-896.

[7] Hu, Yongjian, et al. "Difference expansion based RDH using two embedding directions." IEEE Transactions on Multimedia 10.8 (2008): 1500-1512.

[8] El-sayed, Hala S., S. F. El-Zoghdy, and Osama S. Faragallah. "Adaptive difference expansion-based RDH scheme for digital images." Arabian Journal for Science and Engineering 41.3 (2016): 1091-1107.

[9] Ni, Zhicheng, et al. "RDH." IEEE Transactions on circuits and systems for video technology 16.3 (2006): 354-362.

[10] Jia, Yujie, et al. "RDH based on reducing invalid shifting of pixels in histogram shifting." Signal Processing 163 (2019): 238-246.

[11] Li, Yuanzhang, et al. "A high-imperceptibility and histogram-shifting data hiding scheme for JPEG images." IEEE Access 7 (2019): 73573-73582.

[12] Manikandan, V. M., and V. Masilamani. "Histogram shifting-based blind watermarking scheme for copyright protection in 5G." Computers & Electrical Engineering 72 (2018): 614-630.

[13] Murthy, Kandala Sree Rama, and V. M. Manikandan. "A Block-wise Histogram Shifting based RDH Scheme with Overflow Handling." 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2020.

[14] Nguyen, Thai-Son, and Chin-Chen Chang. "A RDH scheme based on the Sudoku technique." Displays 39 (2015): 109-116.

[15] Parah, Shabir A., et al. "Hiding clinical information in medical images: a new high capacity and RDH technique." Journal of biomedical informatics 66 (2017): 214-230.

[16] Weng, Shaowei, et al. "Dynamic improved pixel value ordering RDH." Information Sciences 489 (2019): 136-154.

[17] Manikandan, V. M., and P. Renjith. "An Efficient Overflow Handling Technique for Histogram Shifting based RDH." 2020 International Conference on Innovative Trends in Information Technology (ICITIIT). IEEE, 2020.

[18] USC-SIPI image dataset: http://sipi.usc.edu/database/, accessed on 05-11-2020

[19] Mittal, A., R. Soundararajan, and A. C. Bovik. "Making a Completely Blind Image Quality Analyzer." IEEE Signal Processing Letters. Vol. 22, Number 3, March 2013, pp. 209–212.

[20] Mittal, Anish, Anush K. Moorthy, and Alan C. Bovik. "Blind/referenceless image spatial quality evaluator." 2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR). IEEE, 2011.

A Bird's Eye View of Natural Language Processing and Requirements Engineering

Assad Alzayed[1], Ahmed Al-Hunaiyyan[2]
Computer Science and Information Systems Department, College of Business Studies
Public Authority for Applied Education and Training, (PAAET), Kuwait

***Abstract*—Natural Language Processing (NLP) has demonstrated effectiveness in many application domains. NLP can assist software engineering by automating various activities. This paper examines the interaction between software requirements engineering (RE) and NLP. We reviewed the current literature to evaluate how NLP supports RE and to examine research developments. This literature review indicates that NLP is being employed in all the phases of the RE domain. This paper focuses on the phases of elicitation and the analysis of requirements. RE communication issues are primarily associated with the elicitation and analysis phases of the requirements. These issues include ambiguity, inconsistency, and incompleteness. Many of these problems stem from a lack of participation by the stakeholders in both phases. Thus, we address the application of NLP during the process of requirements elicitation and analysis. We discuss the limitations of NLP in these two phases. Potential future directions for the domain are examined. This paper asserts that human involvement with knowledge about the domain and the specific project is still needed in the RE process despite progress in the development of NLP systems.**

Keywords—Automated text understanding; natural language processing; requirements engineering; requirements elicitation

I. Introduction

Natural Language Processing (NLP) has a significant functional value in many application fields. NLP is especially useful in the requirements engineering (RE) domain. RE is a vital part of software engineering and considered as the first phase in the software development life cycle. It consists of several activities, including elicitation, analysis, documentation, validation, and management of requirements [1]. RE is a complicated process that is both time-consuming and error-prone, especially for large projects [2]. The RE process defines all of the requirements that a new system needs to complete to be successful. In addition, RE process collects the necessary and appropriate domain knowledge that comprises the requirements of the stakeholders (customers, end-users, domain experts). The requirements elicitation and requirements analysis tasks are performed incrementally and iteratively to accomplish this goal. These tasks use both informal natural language (NL) and formal modeling language. Using NL helps communicate with the stakeholders. It is the universal format language understood by end-users and stakeholders from all involved disciplines [3]. However, NL can be ambiguous and result in misunderstandings concerning the definition of requirements.

NLP can improve the communication between requirement engineers and domain experts (i.e., the users) by creating suitable RE specification documents [3]. NLP can also improve computer understanding of natural language text written by humans [4].

This paper presents a survey of how NLP supports current RE approaches. Specifically, the following research questions are addressed:

Q1: What is the current state of the practice for elicitation and analysis phases in RE using NLP as the platform?

Q2: What are the activities for the requirements elicitation and analysis in RE using NLP as the platform?

Q3: Are NLP systems improving the requirements elicitation and analysis for RE?

Q4: What are the current limitations of using NLP in requirements elicitation and analysis?

A literature review was conducted to address these research questions. This literature review summarized the research reporting use of NLP to support RE activities. The literature review was divided into data preparation, data collection, and data analysis stages. First, the literature search criteria were developed based on the research questions. Second, literature searches were conducted over a predefined collection of databases, including Springer Link, Scopus, IEEEXplore, Google Scholar, Science Direct, and ACM Digital Library. The literature search results were evaluated by title and abstract for all four study questions. This literature review was not an exhaustive review. Instead, this literature review provides a snapshot of the state-of-the-art practices based on the Kitchenham guidelines [5] for conducting systematic literature reviews.

This paper provides an overview of the current state of practices and challenges associated with the elicitation and analysis phases of RE employing NLP techniques. This study focused on three important aspects:

- Providing an overview of the challenges facing the requirements elicitation and the requirements analysis (i.e., ambiguity, inconsistency, incompleteness, and requirements classification).

- Providing an overview of the state-of-the-art approaches to NLP support of current RE operations. Precisely, this study focused on how the available NLP tools and techniques support requirements elicitation and requirements analysis.
- Providing an overview of the limitations of NLP use in requirements elicitation and requirements analysis.

This paper is organized as follows. Sections II and III provide background discussion about NLP and RE. Section IV provides an overview of RE activities where NLP was used to support the requirements elicitation and analysis processes. Section V discusses current requirements elicitation and analysis practices in NLP. Section VI discusses the current limitations of using NLP in requirements elicitation and analysis. Section VII provides general discussions. Finally, Section VIII concludes the paper.

II. Requirements Engineering

The elicitation, analysis, and management of requirements on the basis of semantics during the system development process are difficult due to a large number of requirements for large system engineering projects. Experts normally face numerous constraints during the processes of elicitation and analysis. These constraints include time restraints, insufficient human cognitive capacity to understand the full scope of the processes, and the volume of data to be processed [6]. These considerations make it difficult to manage and maintain the quality of the software requirements specification (SRS) document.

Requirements elicitation involves understanding the objectives and motivation for proposed system software. This phase usually begins with fundamentally informal knowledge and involves unfamiliar people with the processes for developing a software system. Thus, data interpretation is influenced by misunderstandings between the analysts and the consumers [7]. In RE, it is essential to have an excellent semantic understanding of the situation before beginning the process of requirements elicitation [8]. Traditionally, requirements were communicated to the elicitation team using NL text to prevent misunderstandings related to variations in terminology. However, ambiguity is a problem related to NL [9]. The elicitation team may misinterpret or misunderstand the stated requirements specified using NL due to the method used to communicate these requirements. Additionally, problems can occur that are associated with the elicitation of functional requirements (FRs) and the numerous sub-categories of the non-functional requirements (NFRs) [10]. These problems stem from differences in computer jargon and terminology to describe the requirements between the stakeholders and the requirements engineers [11]. The lack of consistency in the requirements documentation process makes the requirements classification process difficult and prone to error [12]. Cordes and Carver (1989) [13] stated that requirements are not created by a single human individual but are the result of common needs from multiple communities. These requirements can introduce uncertainty and inconsistencies. They further state that different participants in the requirements elicitation process have different interpretations of the meaning of the requirements. The participants have different opinions about the design of the new system based on their interpretation of the requirements. The resulting requirements will be ambiguous, contradictory, and incomplete if the participants in an elicitation process do not have a common semantic understanding of the requirements.

Requirements analysis involves understanding and assessing the documented requirements. This phase is concerned with checking the set of elicited requirements for qualifying conflict, omission, duplication, ambiguity, and inconsistency criteria. Common requirements analysis practices involve using checklists for analysis, prioritization, and sorting of requirements, using interaction matrixes to define differences and overlaps, and developing a risk evaluation of the requirements [14].

Requirements elicitation and requirements analysis are usually interlinked processes. Requirements are identified during elicitation, and then analyses are performed. If issues are identified, they are addressed and solved using the source of the requirements [14]. Once a requirement has been elicited, modeled, and analyzed, the SRS document should be clearly and unambiguously documented [15]. The SRS is part of a contract and must simply, accurately, and unambiguously define the requirements of the user and the system. An SRS that is inconsistent, unmanaged, vague, incorrect, or unclear inevitably leads to cost overruns and missed deadlines [16], [17] and [18]. A noteworthy research issue in RE is ambiguity which is described as “a statement having more than one meaning.” Ambiguities may be lexical, syntactic, semantic, pragmatic, vague, generic, or linguistic [19].

The use of NLP in RE is critical because requirement specifications are developed in collaboration between the software analysts and the end users. End users, consumers, and customers will not sign a contract if the requirements are written in the formal language. [20].

Several projects have demonstrated that the RE process can be automated or semi-automated by using NLP [21], [22], and [23]. Furthermore, NLP can support requirements elicitation and requirements analysis by automatically eliminating the ambiguity barrier.

III. Natural Language Processing

NLP is technically one of the sub-fields of artificial intelligence. NLP implements computational and linguistics techniques to assist computer understanding. Additionally, NLP can create human language in the form of texts and speech/voice [24]. The processing of NL is difficult and involves different techniques from those used in artificial languages [25]. NLP approaches are usually based on machine learning (ML). For NLP, the ML process is composed of two tasks: natural language understanding (which is the task of understanding the text) and natural language generation (NLG) (which is the task of generating text with a syntax that is widely used by humans) [26] and [27]. Another study [28] identified three types of NLP technologies: NLP techniques, NLP tools, and NLP resources. An NLP technique is a functional method, approach, process, or procedure for conducting a specific NLP task. NLP techniques include part-of-speech (PoS) tagging, parsing, and tokenization. An NLP tool is a software system or software library that continues to support one or more NLP techniques. Examples of NLP tools include Stanford CoreNLP7, NLTK8, and OpenNLP9. An NLP resource is a linguistic data resource that assists NLP techniques or tools. An NLP resource can be a language lexicon (i.e., dictionary) or a corpus (i.e., a collection of texts).

A lexical analysis can be included in a requirements document to specify pre-built dictionaries, databases, and rules. The goal of lexical analysis is to analyze the meaning of specific words. Five key techniques can be used in a lexical analysis: sentence splitting, PoS tagging, tokenization, morphological analysis, and parsing. Sentence splitting involves separating the text into different sentences. During this process, the NL text is evaluated to determine sentence boundaries.

Tokenization involves dividing a sentence into meaningful components, called tokens. Depending on the form of the text, which is partially determined by the sentence splitting, the tokens are assigned to a category, including punctuation marks, numbers, symbols, and words. The PoS tagging method involves tagging each token with its grammatical group, depending on its meaning and context. Each token is designated with a tag, including noun, verb, adjective, or determinant [29].

The morphological analysis is the initial stage in syntactic analysis. The goal of syntactic analysis is to define the origin of a compound word. Compound words are quickly stemmed and lemmatized. Stemming is the process that reduces a compound word into its ground form or origin. Lemmatization is the process that searches for the ground form of a word [29].

Parsing is a process that analyzes a sentence by taking each word and determining its structure based on its constituent sections. Two components are required to parse a piece of text: a parser and a grammar. Ambiguous sentences may require several different types of analyses in the grammar of NLs [30]. There are two main parsing approaches: dependency parsing and phrase structure parsing. Dependency parsing focuses on the connections between the words in a sentence. Phrase structure parsing involves construction of a parse tree using probabilistic context-free grammar.

The output of a lexical analysis serves as the input for a syntactic analysis. This method performs a sentence analysis of the words to determine the grammatical structure of the sentence. It requires both grammar and a parser. This level of processing results in a representation of the sentence that shows the structural relationship of the dependence between words [4].

Semantic processing defines the potential meaning of a sentence by focusing on associations between word-level meanings in a sentence [4]. Semantic processing builds a description of the objects and actions identified in a sentence and include the details given by the adjectives, adverbs, and prepositions [31].

The goal of categorization is to automatically assign new documents to categories that are already defined [32]. In RE, the NLP method is used to collect requirements from a text; analyze the consistency, linkages, similarities, and ambiguities in the text; and automatically group the text. It also classifies requirements for specific purposes that may be useful during software development. Work associated with the classified requirements may be split between different software development teams, with each team assigned a different class of requirements [33].

IV. NLP for Requirements Elicitation and Analysis

Traditionally, requirements elicitation and requirements analysis are manually processed, expensive, time-consuming, and resource-demanding [17]. Researchers have applied NLP techniques and tools to support a range of linguistic analysis tasks performed at various phases to produce complete requirements documents without ambiguity and inconsistency [34]. The requirements can be illustrated for the stakeholders in a semi-automated or automated way [35], [36], and [37]. Requirements may appear in different forms, including lists of single words, phrases, paragraphs, short texts, and special characters.

Generally, requirement engineering problems are primarily caused by heavy dependence on humans use of NL [38]. NL is syntactically ambiguous and semantically inconsistent. A systematic analysis of literature from 1995 to 2016 indicates that assembly of ambiguous requirements remains one of the most critical problems in software engineering [39]. In response, researchers have attempted to use NLP systems to solve the ambiguity challenges of NL. NLP systems have also been used to support the communication process between system users and stakeholders during the development stages of a system [15]. Communication techniques may focus on pre-selected tools (e.g., Stanford Parser), preferred methods (e.g., rule-based and ontological-based), or degree of automation. The work of [40] provides a detailed discussion about the current approaches to ambiguity in the field of requirements. This paper evaluates empirical work on NLP tools and techniques for dealing with different types of requirement ambiguity [40]. These studies indicate that a significant number of current software implementations solely rely on ambiguity recognition. However, compensation is the responsibility of the stakeholders.

An interesting research area is using NLP for eliciting and analyzing domain requirements based on developed domain ontologies. Ontologies provide a standardized means of organizing information among stakeholders in RE. Thus, ontologies may significantly enhance the quality of the elicited requirements [41]. For example, [42] used a domain ontology and meta-model requirements to generate and elicit requirements. Similarly, [43] describes three core features of domain ontologies ideal for elicitation requirements: explicit relational expression, competent relationship recognition, and explicit temporal and spatial expressions. For the creation of certain domain ontologies, A rule-based approach is recommended for the creation of certain domain ontologies from NL technical documents [44]. The research [45] used NLP to derive formal representations of the requirements based on object-oriented designs using intermediate models.

NLP is recognized as general assistance in analyzing requirements for ambiguity defection [46]. NLP techniques were used to retrieve information and synthesize models. For example, [47] produced unified modeling language (UML) templates (e.g., use-case, analysis class, collaboration, and design class diagrams) from the requirements of natural language using a collection of syntactic reconstruction rules. In addition, [48] proposed a tool-supported approach to promote the process of requirements analysis and the retrieval of class diagrams from textual requirements that support NLP and domain ontology techniques.

Emerging software paradigms, including social networks, mobile computing, and cloud computing, has expressed a growing interest in using NLP techniques. Additionally, NLP techniques are being explored for extensive data analysis to enable data-driven RE [49] and crowd-based RE [50]. Requirements articulated in user stories have been presented as an interesting application of NLP to support agile methodology [51].

V. Current State of Practice of NLP in RE

Applying NLP to RE is an area of research and development that implements NLP tools, techniques, and resources to a range of requirements documentation to facilitate various linguistic analysis activities performed at different RE phases. These tasks include detecting language problems, defining core domain terms, and creating traceability links between requirements [28].

Currently, most NLP tools are used for solving problems in the elicitation phase. NLP tools are also used to extract NL text by the process model, based on parsers and tagging [52]. For example, [53] used 2PoS tagging for preprocessing during the development of conceptual models. This paper proposed an automated solution called Visual Narrator based on NLP. Visual Narrator derives a conceptual model based on user story requirements. In this process, the PoS tagging is used to define the linguistic pattern of the sentence. If the PoS tagging is determined to be a requirement, it is collected from the text corpus and gathered for the next steps in the methodology. The automated approach enables identifying dependencies, redundancies, and inconsistencies between requirements based on a comprehensive and understandable view created from long textual requirements.

The latest trend in requirements elicitation uses NLP is to mine accessible databases (e.g., social media, requirement documents, or Apple Store feedback). The mining process is carried out with the help of ML techniques, NLP, and text mining [54]. Recently, a growing body of research has assessed the use of NLP techniques to extract requirements based on different types of user feedback for the requirements elicitation process [55]. For example, [56] created a tool for detecting ambiguous words in translated SRS.

Currently, an approach has been used to automate requirements elicitation and classification criteria which utilizes an intelligent conversational chatbot. For example, [57] used ML and artificial intelligence to develop a chatbot that interacts with stakeholders using NL and creates formal system requirements based on conversation. This chatbot then classifies the elicited requirements into functional and non-functional system requirements. Additionally, chatbots are widely used in web applications to provide help or information requested by users. For example, CORDULA is a framework that uses chatbot technology to establish contact with end-users for requirements elicitation and understand users' needs. CORDULA guides the users to their desired outcome with minimal effort required by the end-user [58].

Domain ontology has been widely used to improve the elicitation and analysis of functional and NFRs. For example, [59] used NLP to extract NFRs for natural language documents. Furthermore, [60] used an ontology-based approach to support the collection of knowledge to identify possible solutions for eliciting NFRs. Additionally, NLP has been used in similarity analyses to identify functional and NFRs from user app reviews [61] and [62]. The study of [63] proposed an ontology-based approach to support software requirements traceability, which makes it possible for a development team to effectively manage the evolution of the requirements for a software product.

New requirements analysis tools based on NLP are emerging. These tools should significantly reduce the cost of fixing requirement errors by faster identification, thus freeing domain experts from tedious, time-consuming tasks. For example, QuARS (Quality Analyzer for Requirements Specifications) is a tool that analyzes NL requirements in a comprehensive and automated manner using NLP techniques. QuARS emphasizes detecting potential linguistic weaknesses (i.e., ambiguity) that can create issues with interpretation at the next stage of software development [64]. This tool partially assists with the analysis of accuracy and completeness by grouping requirements based on specific concerns. However, user interaction is recognized as a crucial factor adversely affecting the performance and approval of the entire processing. The study of [65] proposed an NLP technique that uses a classification method to automatically handle redundancy and inconsistency problems in a requirement document.

An annual workshop called the NLP4RE (Natural Language Processing for Requirements Engineering) was established to explore interests in NLP applications related to RE issues [66]. The goal of NLP4RE is to help requirements analysts perform multiple linguistic analysis activities for RE phases. This workshop has produced numerous publications and gained broad interest from diverse cultures. About 42.7% of NLP4RE studies focused on the analysis phase. These analysis phase studies used detection as the core linguistic analysis activity and requirements specifications as the processed document type [28].

In the industrial sector, many companies have begun developing NLP tools for RE. For example, Qualicen developed Requirements Scout, a tool that analyses requirement specifications, detects requirement ambiguity, and requirement “smells.” The ThingsThinking system offers several tools under the brand name, Semantha®. This system includes a tool that classifies requirements and identifies the associated risks. It also has a tool that performs document comparison on a semantic level, which can be used for analysis of requirements documents created by multiple stakeholders. QRA Corp developed QVscribe, a tool that checks the quality and consistency of requirements documents. OSSENO Software developed ReqSuite, which is a tool to support the writing and review of specifications. IBM recently developed the IBM Engineering Requirements Quality Assistant, which is an application that leverages the advanced NLP capabilities of IBM Watson for automated requirements analysis and management.

VI. Current Limitations of NLP in Requirements Elicitation and Requirements Analysis

An automated means of enabling software engineers and project managers to develop and refine their NL requirements is needed in RE. NLP can reduce the human effort in making NL requirements clear, consistent, unambiguous, and easy to understand by all stakeholders before moving into modeling and design phases. There are still numerous limitations on the capabilities and rationales for using NLP techniques, despite research developments on NLP for RE. Various challenges for NLP use within requirements elicitation and analysis still exist, including the followings:

- Coreference resolution: Coreference resolution is the task of extracting several expressions in a sentence or text that refer to the same entity/actor in a requirements document. It is especially employed at the semantic/pragmatic level when two nouns are treated the same [67]. Coreference resolution is a key challenge of NLP, not only in English but also for all other languages [68].

- Emotion Detection (ED): While the use of NLP system in requirements elicitation and feedback techniques are well defined, there are no current state-of-the-art techniques that combine emotionally driven features and the capture of user feedback on these features [69]. Emotion recognition may also be used to evaluate social media data or to spot fake news [70]. A major challenge in ED is that the cultural affiliations of an individual may significantly impact their expressed feelings in a situation [69]. However, progress is being made as several methods have been developed to solve this problem. These methods include the use of a knowledge-enriched transformer [71], focusing on latent representation [72], and building new datasets that focus on emotions [73] and [74].

- Unimodal LNP: Current NLP systems are primarily unimodal. Thus, they are limited to process and analysis of linguistic inputs [75]. However, humans are multimodal. They use diverse combinations of visual, auditory, tactile, and other inputs. Humans do not handle each sensory model in isolation but rather simultaneously. This process incorporates each sensory model to enhance the quality of awareness and understanding [76]. Therefore, from a computational point of view, NLP needs to have these same abilities to achieve human-level ground and understanding in a variety of AI tasks. NLP must be assisted by multimodal control interfaces, identification and understanding of human behavior, and collaborative decision-making between the system and individuals or groups to understand the requirements of the customer and other stakeholders [77]. Visual question answering is a method that addresses the challenging unimodal aspect of NLP systems [78]. Many other methods are used to integrate multimodality into NLP structures, including declarative learning-based programming [79], multimodal datasets [80], procedural reasoning networks [81], and unified attention networks [82].

- Ability to recognize requirement sentences that contain contextual information rather than merely describing the process steps [28]. The inherent ambiguity of NL can lead to differing interpretations of the same sentence [83].

- Domain ontologies approach: [43] found that requirement analysts were more likely to misidentify concepts and relationships when using a domain ontologies approach. Thus, domain ontologies need to be investigated to develop a deeper understanding of the requirements and their respective relationships [84].

- NLP accuracy in extracting the correct requirements must be improved. NLP must be enhanced by other methods (e.g., ML) to eliminate errors. Accuracy must be substantially improved if NLP to be seriously considered for use with RE [85].

- The algorithms for detecting ambiguity need improvement and fine-tuning while simultaneously avoiding over-fitting. These improvements are needed to evaluate whether the use of domain ontologies can lead to a deeper understanding of the requirements and their relationships [86].

- PoS detection is generally considered a challenge that has been resolved. However, there are still issues with incorrect POS tags [87].

Attempting to resolve all ambiguities in a requirements specification is a time-consuming process that cannot be fully automated. Human interaction is needed to overcome dynamic ambiguities that are dependent on domain knowledge. Controlled language is helpful in identifying or avoiding ambiguities in SRS. However, the input must be written in the constraint language, and lexical and syntactic ambiguities must be addressed. Furthermore, methods that use knowledge-based, ML, and ontology techniques may produce precise outcomes by detecting semantic ambiguities in the requirements specifications [88], [89], [90].

VII. Discussion

Our research focused on a specific field and evaluated a range of trends explained and summarized in this section. The objective of this research was to provide a state-of-the-art summary of NLP performed in various RE activities. This research is intended to be an overview for domain experts and serve as an entry point for researchers in this area. We present results based on the conducted literature review. As previously discussed, the literature review was not entirely systematic.

Thus, our findings may be revised and/or expanded by future studies within this domain.

Table I and Section IV provide answers to RQ1 (“What is the current state of the practice for elicitation and analysis phases in RE using NLP as the platform?”) and RQ2 (“What are the activities for the requirements elicitation and analysis in RE using NLP as the platform?”). Twenty-five articles (i.e., the “Contributions” column in Table I) were closely analyzed to assess the state-of-the-art of NLP the use NLP in RE activities. We assume that most studies are preliminary proposals because there are more academic research papers than actual software projects for industrial applications. As listed in Table I, the “NLP Tasks” and “RE Support.

TABLE I. CONTRIBUTIONS AND RE ACTIVITIES SUPPORTED BY NLP

N	Contribution	Publication Year	Objective/Purpose	NLP Tasks	RE Activities Support	Tools +Techniques + Resources	User interaction
1	Shah (+)	2015	Compare NLP approaches to resolve ambiguity	Detection	Elicitation + Analysis	-	-
2	Omoronyia et al., (-)	2010	Reducing the effort of building a domain ontology for requirements elicitation by analyzing NL text from technical standard	Extracting semantic graphs	Elicitation	POS tags Sentence parser	Medium
3	Ibrahim et al., (*)	2010	Using NLP and domain ontology to extract UML class diagram from informal NLP by using the RACE tool to produce class diagram from requirements	Extracting	Analysis	PoS tags, Open NLP parser	Medium
4	Maalej et al., (*)	2015	The authors discussed how software development organizations can employ user feedback to identify, prioritize, and manage requirements	Classification	Elicitation + Prioritization	-	-
5	Bano (+)	2015	The findings of a mapping analysis are discussed in a collection of empirical work from the last two decades that discuss the concept of complexity in RE using NLP methods and techniques	Ambiguity Detection	-	-	-
6	Robeer et al., (+)	2016	Extracting conceptual modules from user stories requirements using NLP	Modeling	Analysis	spaCy tagger	Low
7	Bordignon et al. (-)	2018	Using NLP PoS tagging and Parsing in the first three BPM phases to extract ambiguity	Classification	Elicitation + Analysis	PoS Tagging	Low
8	Lucassen et al., (-)	2017	Proposing an automated approach called Visual Narrator that is based on extraction of conceptual models from user story requirements	Classification + Clustering	Elicitation +Analysis	PoS tagging in order to further optimize the results	Low
9	Groen et al., (*)	2015	Proposes crowd-based RE that integrates elicitation and analysis techniques using a crowd sourcing concept where individual tasks are aggregated to provide a final list of identified requirements	Clustering -	Elicitation		
10	Cruz et al. (-)	2017	Proposes two approaches to identify vague words and phrases in the requirements document in a multilingual language	Classification	Elicitation	-	Low
11	Friesen et al. (*)	2018	Developed a chatpot to help end-users revise current extraction and classification requirements	Detection	Elicitation	POS Tagging	High
12	Hollis et, al., (*)	2017	Proposes a method to capture verbal discussion and translate it into a text transcript, a real time audio-to-text conversion software		Elicitation and Specification	Text mining	
13	Murtazina and Avdeenko (*)	2019	Proposes a method to convert knowledge domain into ontology language	Classification	Traceability	-	Medium
14	Zhong, et, al., (+)	2019	Proposes a Knowledge Transformer (KET) to detect emotions in a conversation	Detection + Classification	Sentiment analysis	-	-
15	Ernst and J. Mylopoulos (-)	2010	Eliciting FR and NFR requirements	Classification	Elicitation	-	-

16	Hu, et, al., (*)	2016	Investigates the lack of consistency in RE and the feasibility of using Human Error Taxonomy to support the SRS inspection process	Detection + Classification	Analysis	-	Low
17	Gnesi et, al., (*)	2019	Using NLP techniques with an emphasis on detection of potential linguistic weaknesses (ambiguity)	Classification Clustering	Analysis	Syntax Parser	High
18	Acheampong, at, al., (+)	2020	A survey of state of the art NLP techniques that combines both emotionally-driven features and the capture of user feedback on these features	Sentiment	Analysis	-	-
19	Abdul-Mageed, et, al., (-)	2017	Using Twitter information to create a large dataset of fine-grained emotions for deep learning purposes	Classification	Analysis	-	-
20	Kordjamshidi, et, al., (+)	2017	Integrated multimodality into NLP structures, including declarative learning-based programming	Classification	Analysis	-	Low
21	Van der Aa et, al., (-)	2018	Describes an application of NLP in the BPM context and illustrates how NLP has the ability to maximize the advantages of BPM activities at multiple stages	-	-	-	-
22	Dalpiaz, et, al., (-)	2018	Proposes a visual model to detect ambiguity in requirements that maybe triggered by the use of different words to refer to the same concept	Classification	Analysis	Semantic Similarity Algorithms	Low
23	Femmer, et, al., (-)	2018	NLP toll to detect quality findings and checks on SRS smells in the requirements	Classification	Analysis	is Word- and Sentence Splitting, Morphologic analysis, Lemmatization (and sometimes stemming), PoS tagging, and syntactic parsing	Low
24	Zhao, et, al., (-)	2020	Presents a comprehensive overview of the applications of NLP in RE	-	Elicitation Analysis	-	-
25	Mezghani, et, al., (-)	2018	Proposes an approach using NLP to detect technical business terms associated with the requirements documents	Classification + Clustering	Analysis	PoS tagging + Noun chunking	

Legends: (+) improves state of the art; (*) no information related to state of the art; (-) comparable with state of the art

Activities” columns address RQ2 and describe the various RE tasks that can be assisted by NLP techniques. These RE tasks include traceability, ambiguity detection, and requirements classification. The available techniques and tools developed to support each RE task are presented (e.g., PoS, tagging, and tokenization). The “Contributions” column in Table I also includes a partial response to RQ 3 (“Are NLP systems improving the requirements elicitation and analysis for RE?”). The answer to RQ3 appears to be preliminary; as indicated by the lack of comparison with state of the art in Table I. Table I also include information related to the quantity the NLP data and user expectations, emotions, and experiences. This rich data set may be used by software developers to assess better their product users' needs, experiences, and sentiments. Mining NLP, especially user opinions, can yield valuable information for product upgrades by software development organizations. However, it is often difficult to extract user requirements from massive amounts of data. Opinions are often shared without regard to grammar or style. This issue has recently caused problems with corpus processing. As a result, we assume that the collected data are unstructured. Software developers focus on user feedback for requirements elicitation and analysis. However, the trustworthiness of comments, tweets, or feedback remains a major problem for the software development community. In section VI, we discussed a variety of limitations in this domain. Other challenges or limitations that the RE community faces by using NLP as the source for eliciting user requirements include user privacy and personalization [55]. These findings address RQ4 (“What are the current limitations of using NLP in requirements elicitation and analysis?”).

Table I lists the NLP methods used in the articles reviewed in this paper. Most researchers use NLP techniques for the identification of ambiguity in RE. Furthermore, due to its wide role in RE, the analysis process was the phase with the most attention in the research. We observed that NLP was primarily used in the preprocessing phase to convert data into a format that was consumable by all stakeholders. Most of the papers in our survey claim that the vast amount of imprecise data generated by NL users may provide tremendous benefits to software development organizations if processed with an NLP system. Most of the articles also indicate that NLP use with RE is still in its early stage; however, this research topic is rapidly expanding. Although NLP can compensate for many of the requirements’ ambiguity, inconsistency, and incompleteness, there are still circumstances where interaction with end-users is required for clarification [58]. This finding is supported by entries in the “User Interaction” column in Table I. Most of the papers about the use of NLP in requirements elicitation and analysis indicate that parsing requirement texts and classifying the information stored in them are difficult for humans. Thus, these activities should be automated as much as possible.

Furthermore, we discovered that most of the analyzed studies obtained their requirement datasets from external sources (e.g., Twitter, or an app store) rather than from existing documents for requirement elicitation. We found that the current state of the art in this area indicates that the first two phases prioritize researchers and practitioners. Table I shows most of the NLP techniques employed parsers and taggers to explain the tools used in text processing. This finding is supported by the results from [53]. In that study, NLP techniques are applied to sentence segmentation, tokenization, PoS tagging, shallow parsing, dependency parsing, word stemming, lemmatization, and role labeling.

Within the RE field, NLP is primarily used to analyze requirements and schedule them for further processing. This is primarily focused on developing models from elicited requirements and improving the consistency of the SRS. Both SRS and NLP share three common tasks: PoS tagging, rule-based analysis, and syntactic parsing. Two areas that lag behind the other RE (sub-)phases in the scope are requirements documentation (e.g., drafting of the SRS) and requirement prioritization. We assume that the writing requirements are supported by the NLP, including implementation of a specific template, spell checking, and explicitly resolving any possible ambiguities. These writing requirements may be beneficial to requirements engineers. However, since these tools necessitate live contact with requirement engineers, this research is not regularly performed than less advanced tools (e.g., solely for ambiguity checking).

Classification, model extraction, and detection seem to have more advancement than other areas of research, based on the number of publications and reported NLP tasks. This assumption is also valid for ambiguity detection.

VIII. Conclusion

This paper reviewed the current status of using NLP and its limitations in requirements elicitation and analysis processes. With the need for faster speed, lower cost, and higher quality in software engineering, there is an increasing need for automated support for all processing requirements of elicitation and analysis in the RE artifacts. While the pressure from industrial customers is obvious, extensive work is still needed to create automated NLP-based processes in RE. NLP tools and techniques have been proposed to automatically or semi-automatically detect syntactic, semantic, and pragmatic ambiguities in the requirements. Many solutions have been proposed from academia and industry to evolve the use of NLP in RE. Despite progress in NLP, there are still limitations to the NLP system. There is extreme pressure from the industry to improve the accuracy of NLP used to extract system requirements. Therefore, NLP must be enhanced for the elimination of residual errors. The accuracy must be substantially increased if it is to be seriously considered for use in RE. Based on the results from this paper’s review and due to its limitations, NLP systems cannot be considered as a solution that can fix all RE issues. Nevertheless, NLP can be used to assist RE analysts. Findings from this study indicate that NLP can be used in real-world applications. Future research may produce more specialized NLP tools that can help consolidate the domain model and serve as translators between different RE documents and structured models.

These findings provide an understanding of the state of the art in this field of study and are useful for developing an analytical framework for complete systematic literature reviews. From the standpoint of RE, it may be important to investigate how NLP task combinations can be streamlined to fully perform additional tasks. The same is true for NLP in requirements management where requirements are managed within a software system, from elicitation to implementation to reuse. The NLP tasks described in this paper is not a comprehensive list. A possible extension of this research may examine NLP activities that have not yet been used in the field and how they may be used in the future applications.

Acknowledgment

The authors wish to thank the Public Authority for Applied Education and Training for providing the resources needed to conduct this research.

References

[1] T. Ambreen, N. M. Ikram, M. Usman, and M. Niazi, “Empirical research in requirements engineering: trends and opportunities,” Requirements Engineering, vol. 23, pp. 63–95, 2018.

[2] R. Vlas, and W. N. Robinson, “A rule-based natural language technique for requirements discovery and classification in open-source software development projects” 44th Hawaii Intl Conf. on Syst. Sci., vol. 2011, pp. 1–10, 2011.

[3] C. Huyck, and F. Abbas, “Natural language processing and requirements engineering: a linguistics perspective” in Proc 1st Asia Pac. Conf. on Software Quality, 2000.

[4] E. D.Liddy, Natural Language Processing. Encyclopedia of Library and Information Science, 2001.

[5] B. Kitchenham, Procedures for Performing Systematic Reviews. Keele, UK: Keele University, 2004, pp.1-26.

[6] Ö Albayrak, and J. C. Carver, “Investigation of individual factors impacting the effectiveness of requirements inspections: a replicated experiment,” Emp Softw Eng, vol. 19, pp. 241-266, 2014.

[7] E. Dubois, J. Hagelstein, and A. Rifaut, Formal Requirements Engineering. “with ERAE,” Philips J Res, vol. 43, pp. 393-414, 1988.

[8] D. Zowghi, and C. Coulin, “Requirements elicitation: A survey of techniques, approaches, and tools” in Eng Manag Soft Requirements. Springer, pp. 19-46, 2005.

[9] F. D. Milsom, “Using CORE with SDL to specify requirements” in Proc. Seventh Intl. Conf. on Software Engineering for Telecommunication Switching Systems: SETSS 89 (Conf. Publ. No. 306), vols. 137-141. London: Institution of Electrical Engineers (IEE), 1989.

[10] N. A. Ernst, and J. Mylopoulos, “On the perception of software quality requirements during the project lifecycle,” Lecture Notes in Computer Science. Berlin Heidelberg: Springer, pp. 143–157, 2010.

[11] Z. S. H. Abad, A. Shymka, S. Pant, A. Currie, and G. Ruhe, “What are practitioners asking about requirements engineering? an exploratory analysis of social q a sites” in 24th IEEE Int. Requirements Engineering Conf. Workshops (REW), pp. 334–343, 2016.

[12] W. Hu, J. C. Carver, V. K. Anu, G. S. Walia, and G. Bradshaw, “Detection of requirement errors and faults via a human error taxonomy: a feasibility study” in Proc. 10th ACM/IEEE Intl. Symp. on Empirical Software Engineering and Measurement, pp. 1-10, 2016, Sept.

[13] D. W. Cordes, and D. L. Carver, “Evaluation method for user requirements documents,” Information and Software Technology, vol. 31, pp. 181-188, 1989.

[14] I. Sommerville, and P. Sawyer, Requirements Engineering: A Good Practice Guide, Wiley, 1997.

[15] U. S. Shah, and D. C. Jinwala, “Resolving ambiguities in natural language software requirements: A comprehensive survey. ACM SIGSOFT Software Engineering Notes,” vol. 40, pp. 1–7, 2015.

[16] G. C. Belev, “Guidelines for specification development” in Proceedings., Annual Reliability and Maintainability Symposium (pp. 15-21). IEEE, 1989, Jan.

[17] M. G. Christel, and K. C. Kang, 1992, Issues in Requirements Elicitation (No. CMU/SEI-92-TR-12). Pitts Burgh, PA: Software Engineering Institute, Camegie Mellon University.

[18] D. Firesmith, “Common requirements problems, their negative consequences, and the industry best practices to help solve them,” J Object Technol, vol. 6, pp. 17-33, 2007.

[19] D. M. Berry, E. Kamsties, and M. M. Krieger, 2003, From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity. Univ. of Waterloo, Tech. Rep.

[20] A. Lash, K. Murray, and G. Mocko, “"Natural language processing applications in requirements engineering." ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.” American Society of Mechanical Engineers, 2012.

[21] P. Rayson, R. Garside, and P. Sawyer, Language Engineering for the Recovery of Requirements from Legacy Documents, REVERE project report. Lancaster University, May 1999, 1999.

[22] S. Delisle, K. Barker, and I. Biskri, 1999, Object-Oriented Analysis: Getting Help from Robust Computational Linguistic Tools. na.

[23] P. Sawyer, P. Rayson, and R. Garside, “REVERE: support for requirements synthesis from documents,” Information Systems Frontiers, vol. 4, pp. 343-353, 2002.

[24] G. G. Chowdhury, “Natural language processing,” Annu Rev Inf Sci Technol, vol. 37, pp. 51-89, 2003.

[25] Tj. Eggebraaten, “Rj Stevens, and Ew will. Natural language processing (’NLPoverview’)”. US Patent 8,639,495, pp. 1–19, 2014.

[26] B. Langner, S. Vogel, and A. Black, Evaluating a Dialog Language Generation System: Comparing the MOUNTAIN System to Other NLG Approaches, 2010, pp. 1109–1112.

[27] R. Perera, and P. Nand, “Recent advances in natural language generation: A survey and classification of the empirical literature,” Comput Inform, vol. 36, pp. 1–32, 2017.

[28] L. Zhao, W. Alhoshan, A. Ferrari, K. J. Letsholo, M. A. Ajagbe, E. Chioasca, and R. T. Batista-Navarro, Natural Language Processing (NLP) for Requirements Engineering: A Systematic Mapping Study. Available at: arXiv:2004.01099, 2020.

[29] B. Arendse, A thorough comparison of NLP tools for requirements quality improvement [Master’s thesis], 2016.

[30] Euiwhan Kim Gi-Name Wang (eds.) Bhabani Shankar Prasad Mishra, Satchidananda Dehuri. Techniques and Environments for Big Data Analysis: Parallel, Cloud, and Grid Computing. Studies in Big Data 17. Springer International Publishing, 1st edition, 2016.

[31] A. Reshamwala, D. Mishra, and P. Pawar, “Review on natural language processing,” IRACST Engineering Science and Technology: an International Journal (ESTIJ), vol. 3, pp. 113–116, 2013.

[32] D. Yuret, and F. Türe, “Handbook of natural language processing,”, Proc. Main Conf. on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics -, pp. 328–334, 2006.

[33] F. Mokammel, E. Coatanéa, J. Coatanéa, V. Nenchev, E. Blanco, and M. Pietola, “Automatic requirements extraction, analysis, and graph representation using an approach derived from computational linguistics,” Sys Eng, vol. 21, pp. 555-575, 2018.

[34] K. Ryan, “The role of natural language in requirements engineering”. IEEE Intl Symp. on Requirements Eng., pp. 240–242, 1993.

[35] F. Fabbrini, M. Fusani, V. Gervasi, S. Gnesi, and S. Ruggieri, “Achieving quality in natural language requirements,”, Proc. Int. SW Quality Week, pp. 4–5, 1998.

[36] O. Laitenberger, C. Atkinson, M. Schlich, and K. El Emam, “An experimental comparison of reading techniques for defect detection in UML design documents,” J Sys Soft, vol. 53, pp. 183–204, 2000.

[37] L. Mich, and R. Garigliano, “Ambiguity measures in requirements engineering,”, Proc. ICS 16th IFIP WCC, pp. 39–48, 2000.

[38] U. Ahmed, “A review of knowledge management in requirements engineering”. Intl Conf. on Eng. and Emerg. Technol. (ICEET), pp. 1–5, 2018.

[39] S. Besrour, L. B. A. Rahim, and P. D. D. Dominic, “A quantitative study to identify critical requirement engineering challenges in the context of small and medium software enterprise” in 3rd Intl Conf. on Comput. and Inf. Sci. (ICCOINS), vol. 2016. IEEE, pp. 606-610, 2016, Aug.

[40] M. Bano, "Addressing the challenges of requirements ambiguity: a review of empirical literature" in 5th Intl Workshop on Empirical Requirements Eng. (EmpiRE), vol. 2015. IEEE, pp. 21-24, 2015.

[41] M. S. Murtazina, and T. V. Avdeenko, “An ontology-based approach to support for requirements traceability in agile development,” Procedia Computer Science, vol. 150, pp. 628-635, 2019.

[42] Y. Lee, and W. Zhao, “An ontology-based approach for domain requirements elicitation and analysis,”, Proc. First International Multi-Symposiums on Computer and Computational Sciences, pp. 364–371, 2006.

[43] G. I. Omoronyia, T. G. Sindre, S. T. Stålhane, T. S. Biffl, T. Moser, and W. Sunindyo, “A domain ontology building process for guiding requirements elicitation, International working conference on requirements engineering: foundation for software quality.” Springer, pp. 188–202, 2010.

[44] M. G. Ilieva, and O. Ormandjieva, “Automatic transition of natural language software requirements specification into formal presentation” in, Lecture Notes in Computer Science, pp. 392–397, 2005.

[45] L. Mich, “NL-OOPS: from natural language to object oriented requirements using the natural language processing system LOLITA,” Natural Language Engineering, vol. 2, pp. 161–187, 1996.

[46] K. Li, R. G. Dewar, and R. J. Pooley, 2004, Requirements Capture in Natural Language Problem Statements. Heriot-Watt Technical Report HW-MACS-TR-0023.

[47] D. K. Deeptimahanti, and M. A. Babar, “An automated tool for generating UML models from natural language requirements” in Proc. 2009 IEEE/ACM intl conf. on autom. softw. eng., pp. 680–682, 2009.

[48] M. Ibrahim, and R. Ahmad, “Class diagram extraction from textual requirements using natural language processing (NLP) techniques”. Second International IEEE Conf. on Comput. Res. and Dev., pp. 200–204, 2010.

[49] W. Maalej, M. Nayebi, T. Johann, and G. Ruhe, “Toward data-driven requirements engineering,” IEEE Software, vol. 33, pp. 48–54, 2015.

[50] E. C. Groen, J. Doerr, and S. Adam, “Towards crowd-based requirements engineering a research preview,” Lecture Notes in Computer Science, pp. 247–253, 2015.

[51] M. Robeer, G. Lucassen, J. M. E. M. van der Werf, F. Dalpiaz, and S. Brinkkemper, “Automated extraction of conceptual models from user stories via NLP. IEEE” 24th International Requirements Engineering Conf. (RE), pp. 196–205, 2016.

[52] L. H. Bordignon, T. S. Thom, V. Silva, S. Dani, M. Fantinato, and R. C. B. Ferreira, “Natural language processing in business process identification and modeling: A Systematic Literature Review,” Proc. XIV Braz. Symp. on Information Systems, pp. 1-8, 2018.

[53] G. Lucassen, M. Robeer, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, “Extracting conceptual models from user stories with Visual Narrator,” Requirements Engineering, vol. 22, pp. 339-358, 2017.

[54] C. Hollis, and T. Bhowmik, “Automated support to capture verbal justin-time requirements in agile development: A practitioner view” in 25th International Requirements Engineering Conf. Workshops (REW). IEEE, pp. 419–422, Sept 2017.

[55] E. C. Groen, S. Kopczyńska, M. P. Hauer, T. D. Krafft, and J. Doerr, “Users—the hidden software product quality experts?: A study on how app users report quality aspects in online reviews” in 25th International

Requirements Engineering Conf. (RE), vol. 2017. IEEE, pp. 80-89. IEEE, 2017, Sept.

[56] B. D. Cruz, “Detecting vague words & phrases in requirements documents in a multilingual environment. IEEE” 25th International Requirements Engineering Conf., pp. 233–242, 2017.

[57] C. S. R. K. Surana, D. B. Gupta, and S. P. Shankar, “Intelligent chatbot for requirements elicitation and classification” in 4th Intl Conf. on Recent Trends on Electron., Information, Communication & Technol. (RTEICT), vol. 2019. IEEE, pp. 866-870, 2019, May.

[58] F. S. E. Friesen, F. S. Bäumer, M., and M. Geierhos, “CORDULA: software requirements extraction utilizing chatbot as communication interface,” REFSQ Workshops, vol. 2075, 2018.

[59] J. Slankas, and L. Williams, “Automated extraction of non-functional requirements in available documentation.” 1st Intl Workshop on Nat. Lang. Anal. in Softw. Eng. (NaturaLiSE), pp. 9-16, May 2013.

[60] R. Veleda, and L. M. Cysneiros, “Towards an ontology-based approach for eliciting possible solutions to non-functional requirements” in, Lecture Notes in Computer Science Intl Conf. on Adv. Inf. Syst. Eng. Cham: Springer, (pp. 145-161), 2019, Jun.

[61] H. Yang, and P. Liang, “Identification and classification of requirements from app user reviews. EASE,”, Proc. 21st Intl. Conf. on Evaluation and Assessment in Software Engineering, June (2017), pp. 344–353, '17.

[62] M. Lu, and P. Liang, “Automatic classification of non-functional requirements from augmented app user reviews, ”, Proc. 21st Intl. Conf. on Evaluation and Assessment in Software Engineering, pp. 344-353, 2017.

[63] M. S. Murtazina, and T. V. Avdeenko, “Ontology-based approach to the requirements engineering in agile environment” in International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE), vol. 2018 XIV. IEEE, pp. 496-501, 2018, Oct.

[64] S. Gnesi, and G. Trentanni, “QuARS: a NLP tool for requirements analysis” in REFSQ Workshops, 2019.

[65] M. Mezghani, Juyeon Kang, and F. Sèdes, “Using k-Means for Redundancy and Inconsistency Detection: application to Industrial Requirements ” in Natural Language Processing and Information Systems, M. Silberztein et al, Eds. Cham: Springer International Publishing. ISBN: 978-3-319-91947-8, pp. 501–508, 2018.

[66] F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, “Natural language processing for requirements engineering: the best is yet to come,” IEEE Software, vol. 35, pp. 115-119, 2018a.

[67] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, “End-to-end neural coreference resolution” in Proc. 2017 Conf. on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp. 188–197, 2017.

[68] H. Van der Aa, J. Carmona Vargas, H. Leopold, J. Mendling, and L. Padró, “Challenges and opportunities of applying natural language processing in business process management. In COLING 2018: the,”, Proc. Conf.: August 20-26, 2018 Santa Fe, New Mexico, USA 27th Intl Conf. on Comp. Linguist. Association for Computational Linguistics, pp. 2791-2801, 2018.

[69] F. A. Acheampong, C. Wenyu, and H. Nunoo - Mensah, “Text-based emotion detection: advances, challenges, and opportunities,” Engineering Reports, vol. 2, p. e12189, 2020.

[70] C. Guo, J. Cao, X. Zhang, K. Shu, and H. Liu, 2019, Dean: Learning Dual Emotion for Fake News Detection on Social Media. arXiv Preprint ArXiv:1903.01728.

[71] P. Zhong, D. Wang, and C. Miao, “Knowledge-enriched transformer for emotion detection in textual conversations, CoRR abs/1909.10681 (2019). URL”: http://arxiv, org/abs/1909.10681.

[72] A. Seyeditabari, N. Tabari, S. Gholizadeh, and W. Zadrozny, 2019, Emotion Detection in Text: Focusing on Latent Representation. arXiv Preprint ArXiv:1907.09369.

[73] M. Abdul-Mageed, and L. Ungar, “Emonet: fine-grained emotion detection with gated recurrent neural networks” in Proc. 55th Annu. Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 718-728, 2017, Jul.

[74] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, “Towards empathetic opendomain conversation models: a new benchmark and dataset” in Proc. 57th Annu. Meeting of the Association for Computational Linguistics, Florence, Italy. Association for Computational Linguistics, pp. 5370–5381, Jul. 2019.

[75] Z. Kaddari, Y. Mellah, J. Berrich, M. G. Belkasmi, and T. Bouchentouf, “Natural language processing: challenges and future directions” in Intl Conf. on Artif. Intell. & Ind. Appl. Cham: Springer, pp. 236-246, 2020, Mar.

[76] B. E. Stein, T. R. Stanford, and B. A. Rowland, “The neural basis of multisensory integration in the midbrain: its organization and maturation,” Hearing Research, vol. 258, pp. 4–15, 2009.

[77] G. Schirner, D. Erdogmus, K. Chowdhury, and T. Padir, “The future of human-in- the-loop cyber-physical systems,” Computer, vol. 46, pp. 36–45, 2013.

[78] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, 2019, Visual Question Answering Using Deep Learning: A Survey and Performance Analysis. arXiv Preprint ArXiv:1909.01860.

[79] P. Kordjamshidi, T. Rahgooy, and U. Manzoor, “Spatial language understanding with multimodal graphs using declarative learning based programming” in Proc. 2nd Workshop on Structured Prediction for Natural Language Processing, Copenhagen, Denmark. Association for Computational Linguistics, pp. 33–43, Sept. 2017.

[80] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, 2018, How2: a Large-Scale Dataset for Multimodal Language Understanding. arXiv Preprint ArXiv:1811.00347.

[81] M. S. Amac, S. Yagcioglu, A. Erdem, and E. Erdem, “Procedural reasoning networks for understanding multimodal procedures,” Arxiv e-Prints, arXiv-1909, 2019.

[82] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, 2019, Multimodal Unified Attention Networks for Vision-and-Language Interactions. arXiv Preprint ArXiv:1908.04107.

[83] H. van der Aa, H. Leopold, and H. A. Reijers, “Dealing with behavioral ambiguity in textual process descriptions,”, Lecture Notes in Computer Science. Intl Conf. on Bus. Process Manag. Springer International Publishing, pp. 271–288, 2016.

[84] F. Dalpiaz, I. Van der Schalk, and G. Lucassen, “Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and NLP” in, Lect Notes Comput Sci. Cham: Springer, (pp. 119-135), 2018b, Mar.

[85] T. Iqbal, P. Elahidoost, and L. Lucio, “A bird's eye view on requirements engineering and machine learning” in 25th Asia-Pacific Software Engineering Conf. (APSEC), vol. 2018. IEEE, pp. 11-20, 2018, Dec.

[86] N. Madhavji, L. Pasquale, A. Ferrari, and S. Gnesi, Eds., “Requirements engineering: foundation for software quality”: 26th International Working Conf., REFSQ 2020, Pisa, Italy, March 24–27, 2020, 2020, Proceedings (Vol. 12045). Springer Nature.

[87] H. Femmer, “Requirements quality defect detection with the Qualicen requirements scout” in REFSQ Workshops, 2018.

[88] A. Al-Hunaiyyan, R. Alhajri, A. Alzayed, B. Alraqqas. “Towards an effective Distance Learning Model: Implementation Framework for Arab Universities”. International Journal of Computer Applications. Vol. 6, Issue 5, 2016.

[89] A, Bimba, N. Idris, R. Mahmud, A. Al-Hunaiyyan. “A Cognitive Knowledge-based Framework for Adaptive Feedback”. In S. Phon-Amnuaisuk, T. W. Au & S. Omar (Eds.), Computational Intelligence in Information Systems; Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2016). Pp. 245-255. Cham: springer International Publishing.

[90] R. Alhajri, S. Al-Sharhan. A. Al-Hunaiyyan, T. Al-Othman. “Design of educational multimedia interfaces: individual differences of learners”. Proceedings of (KCESS ’11) the Second Kuwait e-Services and e-Systems Conference. April 5-7, 2011. Kuwait.

An Evaluation of Automatic Text Summarization of News Articles: The Case of Three Online Arabic Text Summary Generators

Fahad M. Alliheibi[1]
Department of Arabic Language, Faculty of Arts
King Abdulaziz University, Jeddah
Saudi Arabia

Abdulfattah Omar[2]
Department of English, College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts
Port Said University, Egypt

Nasser Al-Horais[3]
Department of Arabic Language and Arts
College of Arabic Language and Social Studies
Qassim University
Buraidah, Saudi Arabia

***Abstract*—Digital news platforms and online newspapers have multiplied at an unprecedented speed, making it difficult for users to read and follow all news articles on important, relevant topics. Numerous automatic text summarization systems have thus been developed to address the increasing needs of users around the world for summaries that reduce reading and processing time. Various automatic summarization systems have been developed and/or adapted in Arabic. The evaluation of automatic summarization performance is as important as the summarization process itself. Despite the importance of assessing summarization systems to identify potential limitations and improve their performance, very little has been done in this respect on systems in Arabic. Therefore, this study evaluated three text summarizers AlSummarizer, LAKHASLY, and RESOOMER using a corpus built of 40 news articles. Only articles written in Modern Standard Arabic (MSA) were selected as this is the formal and working language of Arab newspapers and news networks. Three expert examiners generated manual summaries and examined the linguistic consistency and relevance of the automatic summaries to the original news articles by comparing the automatic summaries to the manual (human) summaries. The scores for the three automatic summarizers were very similar and indicated that their performance was not satisfactory. In particular, the automatic summaries had serious problems with sentence relevance, which has negative implications for the reliability of such systems. The poor performance of Arabic summarizers can mainly be attributed to the unique morphological and syntactic characteristics of Arabic, which differ in many ways from English and other Western languages (the original language/s of automatic summarizers), and are critical in building sentence relevance and coherence in Arabic. Thus, summarization systems should be trained to identify discourse markers within the texts and use these in the generation of automatic summaries. This will have a positive impact on the quality and reliability of text summarization systems. Arabic summarization systems need to incorporate semantic approaches to improve performance and construct more coherent and meaningful summaries. This study was limited to news articles in MSA. However, the findings of the study and their implications can be extended to other genres, including academic articles.**

Keywords*—AlSummarizer; Arabic; automatic summarization; discourse markers; extraction; LAKHASLY; news articles; RESOOMER; sentence relevance

I. Introduction

The recent unprecedented growth of digital news platforms and online newspapers has resulted in considerable changes in terms of news production and audience reception. Compared to traditional newspapers, digital news networks and online newspapers are extremely fast and easily accessible. Different reports indicate that online newspapers replaced traditional newspapers causing them to lose much of their audience [1]. That is why, almost all daily and weekly newspapers in the West and in the Arab world run their own websites and publish electronic editions [2-4]. Furthermore, news websites have now proliferated in an unprecedented manner, and without any of the traditional restrictions. There is no need for licenses, offices or even employees and correspondents. Anyone can create a news website in the same way used to create a personal website. This has the effect of producing fast and prolific news in an unprecedented manner.

The extensive number and popularity of online newspapers have changed the way people consume newspapers and magazines in many ways. According to Watson [5], by 2020, more than two-thirds of people in the United Kingdom were reading or downloading online news, newspapers, or magazines. In comparison to 2007, the number of online readers had tripled. As of February 2019, the Guardian and Mail Online were the second and third most popular websites in the UK, respectively, as shown in Fig. 1.

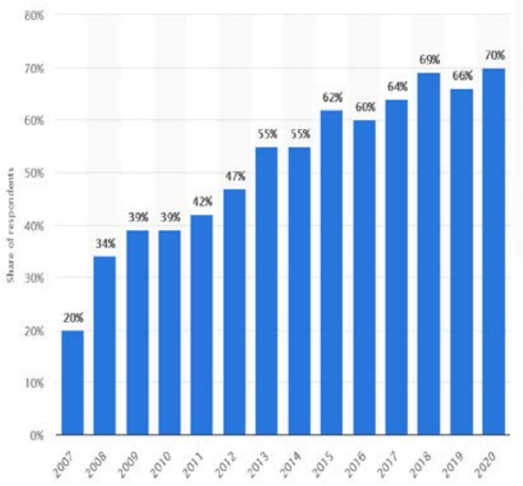


Fig. 1. Share of Individuals Reading or Downloading Online News, Newspapers or Magazines in Great Britain from 2007 to 2020 (Statista.com).

In the same study, Thurman [6] asserts that over recent years online newspapers and news websites have been gaining massive popularity and increasing in an unprecedented manner. In the face of these developments, it is impossible for a normal audience to follow all that is written on important and relevant topics. It is a challenging task for individuals to read this huge content in a limited time span, as the news changes daily if not hourly.

In response, numerous automatic summarization systems have been developed over recent years to generate meaningful summaries that can reduce reading and processing times. Researchers have developed summarization tools and methods that automatically summarize the content of news articles in effective ways. It is even argued that summarization has become an integral part of everyday life [7]. This can be seen in the rapid developments of numerous applications and websites including Summarize Bot, Resoomer, SMMRY, Inshorts, and Text Summarization API (Rapid API) that provide summarization services to news articles to millions of users around the world.

Despite the availability of text summarizers in different languages including Chinese, English, French, and Spanish that provide good summarization services for millions of global users, automatic summarization in Arabic is still very limited. This can be attributed to the unique linguistic system of Arabic where multilingual text summarizers cannot be used with Arabic texts. It is also true that the morphological and syntactic properties of Arabic still pose serious challenges for different Natural Language Processing (NLP) applications, including information retrieval, localization, machine translation, and automatic text summarization [8, 9].

Another reason for this limitation is the lack of evaluation studies of automatic Arabic text summarization. Evaluation is an integral part in the automatic summarization process [10-16]. According to Al Qassem, et al. [17], the evaluation process is one of the main challenges that have adverse impacts on the availability and reliability of automatic Arabic text summarization systems. This can be attributed to the lack of gold standard summaries for Arabic. They add that automatic evaluation of Arabic summarization is more complicated due to the lack of Arabic benchmark corpora, lexica, and machine-readable dictionaries.

To address this limitation, an evaluation of three text summarizers AlSummarizer, LAKHASLY, and RESOOMER is carried out hereunder. A corpus of 40 news articles was built. The articles were randomly selected from the most popular and digital news networks and online newspapers in the Arab world. It was, however, considered that the selected news articles cover different themes and subjects including politics, business, sports, and entertainment. Only articles written in Modern Standard Arabic were selected. The rationale being that MSA is still the formal and working language of the Arab newspapers and news networks.

The remainder of this article is organized as follows. Section 2 is a brief survey of automatic summarization literature in general and automatic Arabic text summarization systems in particular. Section 3 defines the research methods and procedures. In this part, the selected summarizers and articles are defined. Procedures of carrying out the study are also defined and established. Section 4 reports the results. It evaluates the performance of the selected summarizers regarding the news articles. Section 5 concludes the study.

II. Previous Work

Previously, text summarizations were carried out using non-computational methods, using philological methods where experts and professionals produced their summaries based on their own evaluation of the most important concepts in the texts under investigation. With the development of digital technologies, computational approaches have been integrated into text summarization for generating automatic summaries of different text genres. It can be obviously seen that the recent years have witnessed an increasing rate in the development of automatic text summarizers. These have been essentially developed to address the increasing needs of users all over the world. Gambhir and Gupta [18] assert that it has now become impossible for traditional or non-computational classification methods to deal with the large amounts of data available today. They indicate that conventional or non-computational summarization methods are no longer effective or reliable to deal with the prolific size of digital texts and archives available on the internet. According to Cheng and Lapata [19], the need to access and digest large amounts of textual data has provided strong impetus to develop automatic summarization systems, aiming to create shorter versions of one or more documents, whilst preserving their information content. Much effort in automatic summarization has been devoted to sentence extraction, where a summary is created by identifying and subsequently concatenating the most salient text units in a document.

Soni, et al. [20] agree that due to massive rate of rising data at a on the Internet, automatic text summarization tools have a powerful effect on today's world. The entire material is very difficult for a person to describe and ingest. It is a very difficult task to manually convert or summarize, hence, automation is required. Using artificial intelligence methods, automatic text summarization can be accomplished.

Nenkova and McKeown [21] add that objectivity and consistency have always been main considerations in automatic summarizations. The argument is that automatic summarization is imperative for addressing the inconsistencies and lack of objectivity that were associated with conventional or non-computational methods of text summarization.

The first attempt of computer-based text summarization is attributed to Hans Peter Luhn in 1958 [22-25]. In his article 'The Automatic Creation of Literature Abstracts', Luhn [26] proposed an algorithm to facilitate quick and accurate identification of the topic of published papers in a way that saves prospective readers time and effort in finding useful information in a given article or report. The underlying principle of Luhn's approach was that the salient points of an author's argument can be identified through the statistical analysis of the most frequent words and phrases occurring in texts. The hypothesis was that authors generally use important words more frequently throughout a paper and this can be conveniently used as a predictor for selecting the sentences with more repetition of the keywords and extracting them to generate an automatic summary [27-29].

Despite the development of different approaches to automatic text summarization, extractive methods remain the most popular summarization methods. In such methods, automatic summarizers are trained to identify the most important phrases and sentences, usually using statistical methods, and generate automatic summaries based on the extraction process of these sentences and phrases. Extractive-based summarization approaches are based on identifying and selecting only the most important phrases and sentences in texts under consideration. To generate an automatic summary, automatic summarizers then incorporate all the important phrases and sentences. In this case, therefore, every line and word of the summary actually belongs to the original summarized text [30].

In Arabic, over recent years, a very limited number of automatic text summarization systems have been developed, compared to other languages including English, Spanish, and Chinese [16, 17, 31]. Al-Saleh and Menai [10] comment that despite the long history of text summarization, studies of the Arabic language in this area have only recently emerged, and they have been negatively influenced by the lack of Arabic gold standard summaries.

The literature indicates that automatic text summarization systems have been largely based on extractive methods. These extractive summarization systems are mainly based on numerical and statistical measures [32-36]. The main hypothesis in these approaches is the ability of training the machine to identify the most important sentences and phrases for building the summaries. This is usually carried out through using different statistical measures including Principal Components Analysis (PCA) and Term-Frequency Adverse Document-Frequency (TD-IDF). To simplify, these weighting methods are used as indicators for retaining only the important information within texts and discarding information of secondary or minor importance. Successful implementation of these mechanisms is thus a critical factor for the success of summarization systems [37].

Weighting methods are usually combined with other methods that support the identification of the most important sentences, clauses, and phrases in the texts. One popular method is the position or location of sentences. The premise is that the location of a sentence in a document is related to the amount of information it contains. Other techniques involve grouping and summarizing related texts in a process known as multi-document summarization.

The argument is that the majority of automatic Arabic summarization systems are based on statistical methods. One limitation with these techniques is that automatic summaries are based only on those sentences with the highest scores based on statistical measures and techniques. In response, symbolic approaches have been adopted [38-41]. Unlike the numerical and statistical approaches, symbolic-based extractive summarization systems are based solely on linguistic information and indicators for identifying the most important sentences and expressions within texts. The premise of such approaches is that summaries should be built on a rhetorical structure that considers the rhetorical relations within texts. Work on symbolic-based extractive summarization systems is still very limited.

Over recent years, work on/with automatic Arabic summarization systems has reflected on the development of online summarizers that provide summarization services for users in an easy and accessible way. These summarizers are based on the developments in automatic Arabic summarization research and industry. One major problem, however, is the lack of evaluation studies that can determine the readability and reliability of these summarizers. This study addresses this gap in the literature through an evaluation of three Arabic summarizers, namely AlSummarizer, LAKHASLY, and RESOOMER.

III. Methods and Results

This study is based on a corpus of 40 newspaper articles. The articles were selected from nine newspapers issued published in different Arab countries, as shown in Table 1.

The selected articles and opinions represent different topics including politics, business, and sports as shown in Table 2.

TABLE I. THE SELECTED NEWSPAPERS

Selected newspapers		
Newspaper Title	***Country***	***Number of articles***
Al-Ahram	Egypt	8
Al-Bayan	Emirates	4
Al-Itihad	Emirates	3
Alrai	Jordan	3
Al-Watan	Kuwait	4
Annahar	Lebanon	4
Assabah	Morocco	4
Al-Jazeera	Saudi Arabia	5
Asharq Al-Awsat	Saudi Arabia	5

TABLE II. TOPICS OF THE SELECTED ARTICLES

Articles topics	
Category	***Number of articles***
Politics	8
Business	8
Sports	8
Entertainment/Culture/ Family (ETF)	8
Science/Technology	8

For convenience, the selected articles were coded 01-40, as shown in Table 3. The full information pertaining to the selected articles, including transliteration and English translations of the headlines of the selected articles is given in Appendix No. 1.

TABLE III. NEWSPAPER ARTICLES AND CODES

Newspaper articles and codes			
Title	***Code***	***Category***	***Newspaper***
تحرير الأسعار ... وقود الحرائق	01	Business	*Assabah*
عتز صدقي: صناعة السياحة على مستوى العالم «هشة»	02	Business	*Al-Ahram*
تسهيلات جمركية للسيارات الكهربائية المستعملة	03	Business	*Al-Ahram*
تدهور تاريخي لليرة السورية أمام الدولار	04	Business	*Al-Watan*
انتاج" تناقش التحديات والفرص لتنمية الصادرات " الأردنية لتكنولوجيا المعلومات	05	Business	*Alrai*
المنتدى الاقتصادي الأردني يطالب بإعادة النظر في مفهوم القطاع العام وأهدافه وطريقة المساءلة	06	Business	*Alrai*
أير أسيا: نعمل على إطلاق خدمة التاكسي الطائر	07	Business	*Al-Itihad*
وزارة التجارة السعودية: منح 100 ألف سجل تجاري لسيدات خلال 2020	08	Business	*Al-Bayan*
إدانة ساركوزي بتهم الفساد	09	Politics	*Assabah*
«النقد الدولي» يدعو الحكومات للنظر بعين الاعتبار للنساء عند إعداد ميزانياتها	10	Politics	*Al-Ahram*
وزير الإعلام اليمني: ميليشيا الحوثي تواصل المراوغة والتلاعب بملف ناقلة النفط صافر	11	Politics	*Al-Ahram*
زيارة بابا الفاتيكان للعراق استدعاء لقيم العيش المشترك	12	Politics	*Al-Ahram*
قلق سعودي من تنامي وتيرة خطاب الكراهية ضد المسلمين	13	Politics	*Al-Jazeera*
الرياض تطالب مجلس الأمن بمحاسبة الحوثيين على إرهابهم	14	Politics	*Asharq Al-Awsat*
إرادة الشعب والمراهقة السياسية	15	Politics	*Al-Watan*
آلاف الجزائريين يتظاهرون في الاسبوع الثاني من استئناف مسيرات الحراك	16	Politics	Alrai
كاف” تطلق عصبة الأبطال النسوية“	17	Sports	*Assabah*
كلوب: الهزيمة أمام تشيلسي ضربة قوية	18	Sports	*Asharq Al-Awsat*
الإفلاس يضرب 16 نادياً في الدوري الصيني	19	Sports	*Al-Jazeera*
الأسطورة بوفون يكشف موعد اعتزاله	20	Sports	*Al-Watan*
"دربي" مانشستر في الواجهة... وليفربول لاستعادة "سمعته"	21	Sports	*Annahar*
الكرة المصرية.. مواهب «أتلفها الهوى	22	Sports	*Al-Bayan*
برشلونة يترقب رئيسه الجديد وسط جائحة كورونا وعاصفة من الأزمات	23	Sports	*Al-Bayan*
إبراهيموفيتش.. «العودة وشيكة» إلى منتخب السويد	24	Sports	*Al-Itihad*
دراسة طبية: توفير الثقة والدعم للمراهقين يساعدهم في التغلب على الاكتئاب	25	ETF	*Al-Ahram*
«الجارديان» تختار مسجد السلطان برقوق كأحد أفضل 10 زيارات افتراضية في العالم	26	ETF	*Al-Ahram*
أفلام مخرجين صنعت السينما	27	ETF	*Asharq Al-Awsat*
سنوات السينما	28	ETF	*Asharq Al-Awsat*
السينما بعيون أفلاطون	29	ETF	*Al-Jazeera*
تحول الصحف الورقية إلى إلكترونية يسير وتحتاج إلى الدعم..	30	ETF	*Al-Jazeera*
وثائقي جديد يكشف الوجه القبيح للمخرج الأمريكي وودي ألن	31	ETF	*Al-Watan*
الحب المتبادل ينمي ذكاء الأطفال	32	ETF	*Assabah*
علماء يبحثون عن آثار حياة على بعد 26 سنة ضوئية فقط	33	S & T	*Al-Itihad*
تويتر" يتيح ميزة جديدة كان يطالب بها " المستخدمون	34	S & T	*Al-Bayan*
واتساب تطلق المكالمات الصوتية والمرئية لتطبيق الويب قريباً	35	S & T	*Annahar*
الشبيهة بفيديوهات Fast Laughsنتفليكس تطلق "تيك توك"	36	S & T	*Annahar*
"بايت دانس" تطور تطبيقًا شبيهاً بـ "كلوب هاوس"	37	S & T	*Annahar*
أداة ذكية تدمج بالهاتف لقياس الحرارة	38	S & T	*Al-Jazeera*
هياكل روبوتية خارجية تدعم قوة العمل البشرية	39	S & T	*Asharq Al-Awsat*
حسام موافي يكشف عن سبب دقات القلب العالية	40	S & T	*Al-Ahram*

The selected articles were summarized using three Arabic summarizers: AlSummarizer, LAKHASLY, and RESOOMER. These are currently the most popular Arabic summarizers. All three summarizers are based on extractive summarization methods. AlSummarizer is multilingual software that provides summarization solutions in different languages including Arabic, Dutch, English, Farsi, and Turkish. LAKHASLY is an Online Summarization tool. It provides automatic summaries for Arabic and English texts. It is widely used all over the world for users interested in Arabic summaries. Finally, RESOOMER is online summarizers which provide summarization services in different languages including Arabic, English, French, German, Italian, Polish, and Spanish.

For evaluating the performance of the selected summarizers, manual (human) evaluation methods were used. Three examiners were selected to generate manual summaries and to examine the linguistic consistency and relevance of the automatic summaries to the original news articles.

IV. ANALYSIS AND DISCUSSIONS

To evaluate the performance of the three selected Arabic summarizers, the automatic summaries were compared to the manual (human) summaries produced by the experts who participated in the study. Results are shown in Table 4, Table 5, and Table 6.

Comparisons to the manual summaries indicate that the scores of the three summarizers are very similar. The scores also indicate that the performance of the three summarizers is not satisfactory. This has negative impacts on the reliability of these summarizers. Another problem with the three summarizers is the lack of sentence relevance and coherence.

This can be seen in the following example taken from LAKHASLY. There is a problem with the sentence relevance. Sentences are not well connected to one another, which has adverse impacts on the readability and understanding what the original text about, as shown in Box 1.

TABLE IV. SENTENCE OVERLAP IN THE AUTOMATIC SUMMARIES GENERATED BY ALSUMMARIZER

Document Code	Percentage of Overlap	Document Code	Percentage of Overlap
01	33	21	47
02	37	22	34
03	36	23	36
04	42	24	49
05	33	25	35
06	38	26	48
07	49	27	36
08	51	28	45
09	47	29	52
10	39	30	48
11	42	31	36
12	41	32	31
13	35	33	33
14	42	34	37
15	44	35	43
16	49	36	45
17	37	37	37
18	38	38	36
19	33	39	44
20	43	40	47

TABLE V. SENTENCE OVERLAP IN THE AUTOMATIC SUMMARIES GENERATED BY LAKHASLY

Document Code	Percentage of Overlap	Document Code	Percentage of Overlap
01	37	21	47
02	36	22	37
03	33	23	33
04	37	24	48
05	36	25	37
06	48	26	33
07	37	27	43
08	40	28	45
09	41	29	39
10	35	30	35
11	41	31	46
12	51	32	38
13	47	33	35
14	39	34	44
15	44	35	42
16	41	36	47
17	53	37	38
18	33	38	42
19	37	39	40
20	36	40	52

TABLE VI. SENTENCE OVERLAP IN THE AUTOMATIC SUMMARIES GENERATED BY RESOOMER

Document Code	Percentage of Overlap	Document Code	Percentage of Overlap
01	36	21	41
02	31	22	53
03	33	23	33
04	37	24	37
05	43	25	36
06	45	26	38
07	37	27	49
08	36	28	51
09	44	29	47
10	47	30	39
11	37	31	42
12	36	32	41
13	33	33	35
14	37	34	42
15	36	35	44
16	48	36	49
17	37	37	37
18	40	38	38
19	41	39	33
20	35	40	43

TABLE VII. BOX NO. 1. AN AUTOMATIC SUMMARY GENERATED BY LAKHASLY

Original Text
أعربت المملكة أمس عن القلق الشديد من تنامي وتيرة خطاب الكراهية والتعصب ضد المسلمين حول العالم، حيث أصبح التعصب والإسلاموفوبيا خطرًا واضحًا يهدد أمن المجتمعات المستقرة. وقال مندوب المملكة الدائم لدى الأمم المتحدة في جنيف الدكتور عبدالعزيز الواصل في كلمة المملكة أمام مجلس حقوق الإنسان المنعقد في جنيف: إن المملكة العربية السعودية تعرب عن القلق من استهداف الأقليات الدينية بصورة متزايدة بسبب معتقداتهم ومن تنامي خطاب الكراهية داخل الأطياف السياسية التي تستخدم مصطلحات الإقصاء والتهميش وغيرها في ظل وجود منصات التواصل الاجتماعي التي سهلت بث التطرف والكراهية وتعزيز الصور النمطية السلبية ووصم المسلمين عبر وسائل عابرة للحدود. وأكد الواصل خلال الحوار التفاعلي مع المقرر الأممي المعني بحرية الدين والمعتقد، أهمية ترسيخ قيم التفاهم والتسامح والحوار والتعددية والتعارف بين الشعوب والتقارب بين الثقافات ومحاربة كل أيدلوجية تدعو للكراهية وتحرض على العنف وتسوّغ الجرائم الإرهابية التي لا يمكن قبولها في أي دين أو قانون، داعيًا إلى أهمية وجود إرادة دولية وشاملة وذلك من خلال مكافحة الخطاب المتطرف وخطاب الكراهية وعدّ الدين جزءًا من الحل وليس سبب المشكلة. ولفت الانتباه إلى أن المملكة تولي اهتماماً خاصًا بهذا الموضوع حيث أسست مركز الملك عبدالله بن عبدالعزيز العالمي للحوار بين أتباع الأديان والثقافات، وتبنَّت عدة مبادرات للاعتدال ومكافحة التطرف، منها تأسيس مركز اعتدال في المملكة العربية السعودية الذي أطلق مؤخرًا مشروعًا جديدًا تحت عنوان «تفنيد «للتصدي للأفكار المتطرّفة عبر فضاءات «الإنترنت» ومواقع التواصل الاجتماعي، كما استضافت المملكة مؤتمر رابطة العالم الإسلامي الذي عقد في مكة المكرمة في 2019 حيث أقر 1200 شخصية إسلامية من 139 دولة يمثلون 27 مكونًا إسلاميًا من مختلف المذاهب والطوائف، «وثيقة مكة المكرمة» التي تتضمن 29 بندًا لإرساء قيم التعايش بين أتباع الأديان والثقافات والأعراق والمذاهب في البلدان الإسلامية وتحقيق السلم والوئام بين مكونات المجتمع الإنساني كافة.
Summary
وقال مندوب المملكة الدائم لدى الأمم المتحدة في جنيف الدكتور عبدالعزيز الواصل في كلمة المملكة أمام مجلس حقوق الإنسان المنعقد في جنيف: إن المملكة العربية السعودية تعرب عن القلق من استهداف الأقليات الدينية بصورة متزايدة بسبب معتقداتهم ومن تنامي خطاب الكراهية داخل الأطياف السياسية التي تستخدم مصطلحات الإقصاء والتهميش وغيرها في ظل وجود منصات التواصل الاجتماعي التي سهلت بث التطرف والكراهية وتعزيز الصور النمطية السلبية ووصم المسلمين عبر وسائل عابرة للحدود .أهمية ترسيخ قيم التفاهم والتسامح والحوار والتعددية والتعارف بين الشعوب والتقارب بين الثقافات ومحاربة كل أيدلوجية تدعو للكراهية وتحرض على العنف وتسوّغ الجرائم الإرهابية التي لا يمكن قبولها في أي دين أو قانون، وتبنَّت عدة مبادرات للاعتدال ومكافحة التطرف،

Omar [42] explains that if sentences and clauses in the automatic summaries are not connected, the overall argumentative structure of the text is not supported and the thematic significance of the original texts is lost. He adds that in many cases automatic summaries generated in this fashion are misleading for readers and users. Alami, et al. [43] agree that one main reason for the low performance of text summarizers in Arabic is that sentences in the extracted or generated summaries are not relevant; therefore, the main point of the original texts is not clear. The lack of sentence relevance thus has negative impacts on the user’s ability to grasp the meaning of original texts. In such a case, automatic summaries do not provide the users with concise and relevant information that helps them determine and assess the importance of texts without having to read all of the texts.

The lack of sentence relevance and coherence in automatic summarization in Arabic can be attributed to the multilingual nature of automatic text summarizers. Many of the summarizers are usually offered in different languages without considering language- specificity. It is almost agreed upon, that the unique linguistic properties of Arabic are always associated with the low performance in different natural language processing (NLP) applications including automatic summarization [10, 17, 44-46]. To put it into context, the morpho-syntactic system of Arabic is different in many ways from English and other Western languages (the original language/s of automatic summarizers). The morphological and syntactic properties of Arabic are indispensable in building sentence relevance and coherence in Arabic. According to Al Qassem, et al. [17], the main challenge in Arabic text summarization is in the complexity of the Arabic language itself: 1) the meaning of a text is highly dependent on the context; 2) there are more inherent variations within Arabic than any other language; 3) the diacritics are usually absent in the texts of news articles and any online content.

To improve the sentence relevance in the automatic summarization of news articles in Arabic, this study proposes that summarization systems should be trained to identify the discourse markers within the texts and furthermore to use these discourse markers in the generation of automatic summaries. The hypothesis is that discourse markers can be gainfully used to create cohesive texts with sentences that are linked together and relevant to one another. In other words, discourse markers can be used to build cohesive and coherent texts through interrelated sentences which will have positive impacts on the quality and reliability of text summarization systems. Arabic summarization systems need to integrate semantic-based methods for improving the quality of summarization performance and generating more coherent and meaningful summaries.

V. Conclusion

Digital news platforms and online newspapers have multiplied today at an unprecedented speed, making it difficult for users to read and follow all news articles on important relevant topics. Numerous automatic text summarization systems have thus been developed to address the increasing needs of users around the world to access summaries that reduce reading and processing time. In Arabic, different automatic summarization systems have been developed and/or adapted in order to address the increasing need for automatic Arabic summaries. Evaluation of automatic summarization performance is as important as automatic summarization itself.

Despite the importance of the evaluation and assessment of automatic summarization systems for identifying the limitations and for improving the summarization performance, very little has been done on the evaluation systems of automatic text summarization in Arabic. Therefore, an evaluation of three text summarizers AlSummarizer, LAKHASLY, and RESOOMER was carried out. A corpus of forty news articles was built. Only articles written in Modern Standard Arabic (MSA) were selected. The rationale being that MSA is still the formal and working language of Arab newspapers and news networks. For evaluating the performance of the selected summarizers, manual (human) evaluation methods were used. Three examiners were selected to generate manual summaries and to examine the linguistic consistency and relevance of the automatic summaries to the original news articles. To evaluate the performance of the three selected Arabic summarizers, the automatic summaries were compared to the manual (human) summaries produced by the experts who participated in the study. Results indicated that the scores of the three summarizers were very similar. The scores also indicate that the performance of the three summarizers is not satisfactory. Furthermore, the automatic summaries have a serious problem with sentence relevance that has adverse impacts on the reliability of such systems.

It can be concluded that the poor performance of Arabic summarizers can be mainly attributed to the unique morphological and syntactic characteristics of Arabic. The morphological-syntactic system of Arabic is different in many ways from English and other Western languages (the original language/s of automatic summarizers). The morphological and syntactic properties of Arabic are indispensable in building sentence relevance and coherence in Arabic. Thus, this study proposes that summarization systems should be trained to identify the discourse markers within the texts and furthermore to use these discourse markers in the generation of automatic summaries. This will have positive impacts on the quality and reliability of text summarization systems. Arabic summarization systems need to incorporate semantic approaches for improving the quality of summarization performance and building more coherent and meaningful summaries.

This study was limited to the news articles in MSA. However, the findings of the study and their implications can be extended to other genres including academic articles. Further research, however, is recommended to address the performance of automatic Arabic text summarization in social media language and colloquial dialects in Arabic.

Acknowledgment

The corresponding author would like to acknowledge the support provided by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia.

References

[1] E. Siapera and A. Veglis, The Handbook of Global Online Journalism. New York: Wiley, 2012.

[2] W. A. Rugh, Arab mass media: Newspapers, radio, and television in Arab politics. Greenwood publishing group, 2004.

[3] B. Garrison, "Online newspapers," Online news and the public, pp. 3-46, 2005.

[4] M. Gasher and S. Gabriele, "Increasing circulation? a comparative news-flow study of the Montreal Gazette's hard-copy and on-line editions," Journalism Studies, vol. 5, no. 3, pp. 311-323, 2004.

[5] A. Watson. Online news, newspaper and magazine consumption in Great Britain 2007-2020 [Online]. Available: https://www.statista.com/statistics/286210/online-news-newspapers-and-magazine-consumption-in-great-britain/

[6] N. Thurman, "Newspaper consumption in the digital age: Measuring multi-channel audience attention and brand popularity," Digital Journalism, vol. 2, no. 2, pp. 156-178, 2014.

[7] U. Hahn and I. Mani, "The Challenges of Automatic Summarization," Computer, vol. 33, no. 11, pp. 29-36, 2000.

[8] A. Omar, "An Evaluation of the Localization Quality of the Arabic Versions of Learning Management Systems," International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 443-449, 2021.

[9] A. Omar, "Ambiguity Resolution in Arabic Localization: The Case of Learning Management Systems," Applied Linguistics Research Journal, vol. 5, no. 1, pp. 1-6, 2021.

[10] A. B. Al-Saleh and M. E. B. Menai, "Automatic Arabic text summarization: a survey," Artificial Intelligence Review, vol. 45, no. 2, pp. 203-234, 2016.

[11] F. Kiyoumarsi, "Evaluation of Automatic Text Summarizations Based On Human Summaries," Procedia - Social and Behavioral Sciences, vol. 192, pp. 83-91, 2015.

[12] M. Lapata and R. Barzilay, "Automatic Evaluation of Text Coherence: Models and Representations," Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1085-1090, 2005.

[13] L. D. Robert, W. D. Kevin, and A. M. Laura, "A comparison of rankings produced by summarization evaluation measures," presented at the Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, Seattle, Washington, 2000.

[14] L. Scanlon, S. Zhang, X. Zhang, and M. Sanderson, "Evaluation of Cross Domain Text Summarization," in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1853-1856.

[15] J. Steinberger and K. Jezek, "Evaluation Measures for Text Summarization," Computing and Informatics, vol. 28, no. 2, pp. 251-275, 2009.

[16] L. Belguith, M. Ellouze, M. Maaloul, M. Jaoua, F. Jaoua, and P. Blache, "Automatic summarization," in Natural language processing of semitic languages, theory and applications of natural language processing, I. Zitouni, Ed. Berlin: Springer, 2014, pp. 371–408.

[17] L. M. Al Qassem, D. Wang, Z. Al Mahmoud, H. Barada, A. Al-Rubaie, and N. I. Almoosa, "Automatic Arabic summarization: a survey of methodologies and systems," Procedia Computer Science, vol. 117, pp. 10-18, 2017.

[18] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," Artificial Intelligence Review, vol. 47, no. 1, pp. 1-66, 2017.

[19] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," arXiv preprint arXiv:1603.07252, 2016.

[20] V. Soni, L. Kumar, A. K. Singh, and M. Kumar, "Text Summarization: An Extractive Approach," in Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol. 1154, M. Pant, K. Sharma, R. Arya, B. Sahana, and H. Zolfagharinia, Eds. Singapore: Springer, 2020.

[21] A. Nenkova and K. McKeown, Automatic summarization. Now Publishers Inc, 2011.

[22] I. Mani, Automatic Summarization. Amsterdam: John Benjamins Publishing, 2001.

[23] A. Fiori, Trends and Applications of Text Summarization Techniques. IGI Global, 2019.

[24] Torres-Moreno and Juan-Manuel, Automatic text summarization. Hoboken, NJ: John Wiley & Sons, 2014.

[25] R. C. Balabantaray, D. Sahoo, B. Sahoo, and M. Swain, "Text summarization using term weights," International Journal of Computer Applications, vol. 38, no. 1, pp. 10-14, 2012.

[26] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, no. April, pp. 159-165, 1958.

[27] A. Fiori, Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding: Revolutionizing Knowledge Understanding. Hershey, Pennsylvania: IGI Global, 2014.

[28] J. Madhuri and R. G. Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), pp. 1-3, 2019.

[29] Y. K. Meena and D. Gopalani, "Analysis of sentence scoring methods for extractive automatic text summarization," Proceedings of the 2014 international conference on information and communication technology for competitive strategies, pp. 1-6, 2014.

[30] D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda, "Extractive text summarization system to aid data extraction from full text in systematic review development," Journal of Biomedical Informatics, vol. 64, pp. 265-272, 2016/12/01/ 2016.

[31] M. Boudabous, M. Maaloul, and L. Belguith, "Digital learning for summarizing arabic documents," in Advances in natural language processing, lecture notes in computer science, vol. 6233, H. Loftsson, E. Rgnvaldsson, and S. Helgadttir, Eds. Berlin: Springer, 2010, pp. 79–84.

[32] F. Douzidia and G. Lapalme, "Lakhas, an arabic summarization system," DUC 2004, pp. 128–135, 2004.

[33] F. El-Ghannam and T. El-Shishtawy, "Multi-topic multi-document summarizer," Int J Comput Sci Inf Technol, vol. 5, no. 6, pp. 77–90, 2013.

[34] M. El-Haj, U. Kruschwitz, and C. Fox, "Exploring clustering for multi-document arabic summarisation," in Information retrieval technology, vol. 7097, M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, and H. Khelalfa, Eds. (lecture notes in computer science, Berlin: Springer, 2011, pp. 550–561.

[35] N. El-Fishawy, A. Hamouda, G. Attiya, and M. Atef, "Arabic summarization in twitter social network," Ain Shams Engineering Journal, vol. 5, no. 2, pp. 411–420, 2013.

[36] R. Belkebir and A. Guessoum, "A supervised approach to arabic text summarization using adaboost," in New contributions in information systems and technologies, advances in intelligent systems and computing, vol. 353, A. Rocha, A. Correia, S. Costanzo, and L. Reis, Eds., 2015, pp. 227–236.

[37] K. Nandhini and S. R. Balasundaram, "Improving readability through extractive summarization for learners with reading difficulties," Egyptian Informatics Journal, vol. 14, no. 3, pp. 195-204, 2013/11/01/ 2013.

[38] W. Al-Sanie, "Towards an infrastructure for arabic text summarization using rhetorical structure theory," Master’s thesis, King Saud University, Riyadh, 2005.

[39] F. AL-Khawaldeh and V. Samawi, "Lexical cohesion and entailment based segmentation for arabic text summarization (lceas)," World Comput Sci Inf Technol J, vol. 5, no. 3, pp. 51–60, 2015.

[40] H. Fejer and N. Omar, "Automatic arabic text summarization using clustering and keyphrase extraction," 2014 International Conference on information technology and multimedia (ICIMU), pp. 293–298, 2014.

[41] A. Ibrahim and T. Elghazaly, "Improve the automatic summarization of arabic text depending on rhetorical structure theory," 12th Mexican International Conference on Artificial Intelligence (MICAI), pp. 223–227, 2013.

[42] A. Omar, "Addressing the Problem of Coherence in Automatic Text Summarization: A Latent Semantic Analysis Approach," International Journal of English Linguistics, vol. 7, no. 4, pp. 33-44, 2017.

[43] N. Alami, N. En-nahnahi, S. A. Ouatik, and M. Meknassi, "Using unsupervised deep learning for automatic summarization of Arabic documents," Arabian Journal for Science and Engineering, vol. 43, no. 12, pp. 7803-7815, 2018.

[44] A. Omar and W. I. Hamouda, "The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic," International Journal of

Advanced Computer Science and Applications, vol. 11, no. 1, pp. 116-121, 2020.

[45] A. Omar and M. Aldawsari, "Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines," International Journal of English Linguistics, vol. 10, no. 3, pp. 219-228, 2020.

[46] A. Farghaly and K. Shaalan, "Arabic natural language processing: challenges and solutions " ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 8, no. 4, pp. 14:1–14:22, 2009.

AUTHORS' PROFILE

Fahad Alliheibi is a Full Professor of Arabic Language and Linguistics in the Department of Arabic Language, College of Arts, King Abdulaziz University (KSA). Prof. Alliheibi received his PhD degree in Linguistics in 1999 from Durham University, UK. His research interests include Arabic Linguistics, Pragmatics, Text Linguistics, and Translation.

ORCID: 0000-0002-2233-5165

Abdulfattah Omar is an Associate Professor of English Language and Linguistics in the Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University (KSA). Also, he is a standing lecturer of English Language and Linguistics in the Department of English, Faculty of Arts, Port Said University, Egypt. Dr Omar received his PhD degree in computational linguistics in 2010 from Newcastle University, UK. His research interests include computational linguistics, digital humanities, discourse analysis, and translation studies.

ORCID: 0000-0002-3618-1750

Nasser Al-Horais is a Full Professor of Arabic Language and Linguistics in the Department of Arabic Language and its Arts, College of Arabic Language & Social Studies, Qassim University (KSA). Prof. Al-Horais received his PhD degree in Linguistics in 2009 from Newcastle University, UK. His research interests include Arabic Syntax, Generative Linguistics (Minimalist Program), Discourse Analysis, and Comparative Syntax.

ORCID: 0000-0002-2511-9791

P.O.Box 6611, Buraidah 51425, Saudi Arabia

APPENDIX NO. 1: SELECTED NEWSPAPER ARTICLES AND CODES

Title	***Code***	***Category***	***Newspaper***	***Country***
تحرير الأسعار ... وقود الحرائق	01	Business	*Assabah*	*Morocco*
tahrir al'asear ... waqud alharayiq				
Price liberalization…Adding fuel to the fire				
معتز صدقي: صناعة السياحة على مستوى العالم «هشة»	02	Business	*Al-Ahram*	*Egypt*
muetaz sdqy: sinaeat alsiyahat ealaa mustawaa alealam ‘hashah’				
Moataz Sidqi: The global tourism industry is 'fragile'				
تسهيلات جمركية للسيارات الكهربائية المستعملة	03	Business	*Al-Ahram*	*Egypt*
tashilat jumrukiat lilsayarat alkahrabayiyat almustaemala				
Egypt approves new customs tax breaks for used electric cars				
تدهور تاريخي لليرة السورية أمام الدولار	04	Business	*Al-Watan*	*Kuwait*
tadahwur tarikhi liliarat alsuwriat 'amam alduwlar				
Historical deterioration of the Syrian pound against the dollar				
انتاج" تناقش التحديات والفرص لتنمية الصادرات الأردنية لتكنولوجيا المعلومات"	05	Business	*Alrai*	*Jordan*
"Intaj" tunaqash altahadiyat walfuras litanmiat alssadirat al'urduniyat litiknulujia almaelumat				
"Intaj" discusses the challenges and opportunities for developing Jordanian exports of information technology				
المنتدى الاقتصادي الأردني يطالب بإعادة النظر في مفهوم القطاع العام وأهدافه وطريقة المساءلة	06	Business	*Alrai*	*Jordan*
almuntadaa alaiqtisadiu al'urduniyu yutalib bi'iieadat alnazar fi mafhum alqitae aleami wa'ahdafah watariqat almusa'ala				
The Jordanian Economic Forum calls for a review of the concept of the public sector, its objectives, and the methods of accountability				
أير أسيا: نعمل على إطلاق خدمة التاكسي الطائر	07	Business	*Al-Itihad*	*Emirates*
'ayr 'asya: naemal ealaa 'iitlaq khidmat alttakisii alttayir				
AirAsia: We are working on launching the Flying Taxi service				
وزارة التجارة السعودية: منح 100 ألف سجل تجاري لسيدات خلال 2020	08	Business	*Al-Bayan*	*Emirates*
wizarat altijarat alsaeudiati: manh 100 'alf sajal tijariun lasaydat khilal 2020				
Saudi Arabia issued over 100,000 commercial registrations for women in 2020				
إدانة ساركوزي بتهم الفساد	09	Politics	*Assabah*	*Morocco*
'iidanat sarkwzy bituham alfasad				

Former French President Nicolas Sarkozy Found Guilty Of Corruption				
«النقد الدولي» يدعو الحكومات للنظر بعين الاعتبار للنساء عند إعداد ميزانياتها	10	Politics	*Al-Ahram*	*Egypt*
“alnaqd alduwly” yadeu alhukumat lilnazar bieayn alaietibar lilnisa' eind 'iiedad mizaniatiha				
The International Monetary Fund calls on governments to consider women when preparing their budgets				
وزير الإعلام اليمني: ميليشيا الحوثي تواصل المراوغة والتلاعب بملف ناقلة النفط صافر	11	Politics	*Al-Ahram*	*Egypt*
wazir al'iielam alyamani: milishia alhawthayi tuasil almurawaghat waltalaeub bimalf naqilat alnaft safir				
Houthis spreading misleading information regarding Safer tanker: The Yemeni Information Minister				
زيارة بابا الفاتيكان للعراق استدعاء لقيم العيش المشترك	12	Politics	*Al-Ahram*	*Egypt*
ziarat PaPa alvatikan lileiraq aistidea' liqim aleaysh almushtara				
Pope's visit to Iraq sends message of coexistence				
قلق سعودي من تنامي وتيرة خطاب الكراهية ضد المسلمين	13	Politics	*Al-Jazeera*	*Saudi Arabia*
qalaq saeudiun min tanami watirat khitab alkirahiat dida almuslimin				
Saudi concern for the growing hate speech against Muslims				
الرياض تطالب مجلس الأمن بمحاسبة الحوثيين على إرهابهم	14	Politics	*Asharq Al-Awsat*	*Saudi Arabia*
alriyad tutalib majlis al'amn bimuhasabat alhuthiayn ealaa 'iirhabihim				
Saudi Arabia calls on UN to hold Houthis accountable for terror attacks				
إرادة الشعب والمراهقة السياسية	15	Politics	*Al-Watan*	*Kuwait*
'iiradat alshaeb walmurahaqat alsiyasia				
The will of the people and the immaturity of the politicians				
آلاف الجزائريين يتظاهرون في الاسبوع الثاني من استئناف مسيرات الحراك	16	Politics	Alrai	*Jordan*
alaf aljazayiriiyn yatazaharun fi al'usbue alththani min aistinaf masirat alhirak				
Thousands of Algerians demonstrate in the second week of Hirak’s recovery				
كاف” تطلق عصبة الأبطال النسوية“	17	Sports	*Assabah*	*Morocco*
“kaf" tutliq eusbat al'abtal alnaswia				
CAF launches Women’s Champions League				
كلوب: الهزيمة أمام تشيلسي ضربة قوية	18	Sports	*Asharq Al-Awsat*	*Saudi Arabia*
klwb: alhazimat 'amam tshylsy darbatan qawia				
Jurgen Klopp admitted Liverpool's defeat against Chelsea was a massive blow				
الإفلاس يضرب 16 نادياً في الدوري الصيني	19	Sports	*Al-Jazeera*	*Saudi Arabia*
al'iiflas yadrib 16 nadyaan fi aldawrii alsiynii				
Bankruptcy hits 16 clubs in the Chinese Football League				
الأسطورة بوفون يكشف موعد اعتزاله	20	Sports	*Al-Watan*	*Kuwait*
al'usturat bufun yakshif maweid aietizalah				
Buffon reveals when he wants to retire				
"دربي" مانشستر في الواجهة... وليفربول لاستعادة "سمعته	21	Sports	*Annahar*	*Lebanon*
"drabi" manshistar fi alwajihta... walifarbul liaistieada "smieatih				
"Derby" Manchester in the forefront ... and Liverpool to restore "its reputation				
الكرة المصرية.. مواهب «أتلفها الهوى	22	Sports	*Al-Bayan*	*Emirates*
alkurat almisriata.. mawahib iatalifaha alhuaa				

The Egyptian Football: Wasted football talents				
برشلونة يترقب رئيسه الجديد وسط جائحة كورونا وعاصفة من الأزمات	23	Sports	*Al-Bayan*	*Emirates*
barshilunat yataraqab rayiysih aljadid wasat jayihat kwrwna waeasifatan min al'azamat				
Barcelona elects new president amid Pandemic COVID-19 and a storm of crises				
إبراهيموفيتش.. «العودة وشيكة» إلى منتخب السويد	24	Sports	*Al-Itihad*	*Emirates*
Iibrahymwfytsh: aleawdat wshy 'iilaa muntakhab alsuwid				
Ibrahimovic’s return to the Swedish national team is imminent				
دراسة طبية: توفير الثقة والدعم للمراهقين يساعدهم في التغلب على الاكتئاب	25	ETF	*Al-Ahram*	*Egypt*
dirasat tbyt: tawfir althiqat waldem lilmurahiqin yusaeiduhum fi altaghalub ealaa alaiktiab				
Clinical Study: Providing teens with confidence and support helps them overcome depression				
«الجارديان» تختار مسجد السلطان برقوق كأحد أفضل 10 زيارات افتراضية في العالم	26	ETF	*Al-Ahram*	*Egypt*
“aljardyan” takhtar masjid alsultan biruquq kahd 'afdal 10 ziarat aiftiradiatan fi alealam				
The Guardian chose the Mosque-Madrassa of Sultan Barquq among the 10 virtual tours of spectacular buildings around the world				
أفلام مخرجين صنعت السينما	27	ETF	*Asharq Al-Awsat*	*Saudi Arabia*
'aflam mukhrijin sunieat alsiynama				
Debut Films of Famous Directors				
سنوات السينما	28	ETF	*Asharq Al-Awsat*	*Saudi Arabia*
sanawat alsiynama				
Cinema years				
السينما بعيون أفلاطون	29	ETF	*Al-Jazeera*	*Saudi Arabia*
alsiynama bieuyun 'aflatun				
Cinema through Plato’s Eyes				
..تحول الصحف الورقية إلى إلكترونية يسير وتحتاج إلى الدعم	30	ETF	*Al-Jazeera*	*Saudi Arabia*
tahul alsuhuf alwarqiat 'iilaa 'iiliktruniat yasir watahtaj 'iilaa alduem.				
Converting traditional newspapers to electronic is easy and needs support				
وثائقي جديد يكشف الوجه القبيح للمخرج الأمريكي وودي ألن	31	ETF	*Al-Watan*	*Kuwait*
wathayiqiin jadid yakshif alwajh alqubih lilmukhrij al'amrikii wawdi 'alan				
A new documentary reveals the ugly face of American filmmaker Woody Allen				
الحب المتبادل ينمي ذكاء الأطفال	32	ETF	*Assabah*	*Morocco*
alhabu almutabadal yanmi dhaka' al'atfal				
Mutual love develops children's intelligence				
علماء يبحثون عن آثار حياة على بعد 26 سنة ضوئية فقط	33	S & T	*Al-Itihad*	*Emirates*
eulama' yabhathun ean athar hayat ealaa bued 26 sanatan dawyiyatan faqa				
Scientists find Earth-like planet with an atmosphere 26 light years away				
تويتر" يتيح ميزة جديدة كان يطالب بها المستخدمون"	34	S & T	*Al-Bayan*	*Emirates*
"Twitter" yutih myzt jadidatan kan yutalib biha almustakhdimun				
"Twitter" provides a new feature that was demanded by users				
واتساب تطلق المكالمات الصوتية والمرئية لتطبيق الويب قريباً	35	S & T	*Annahar*	*Lebanon*

Whatsapp tutliq almukalamat alsawtiat walmaryiyat litatbiq alwayb qrybaan				
WhatsApp is launching audio and video calls for the web app soon				
الشبيهة بفيديوات "تيك توك"Fast Laughs"نتفليكس تطلق	36	S & T	*Annahar*	*Lebanon*
natiflikis tutliq FAST LAUGHS alshbyht bfydywat "tyk twk"				
Netflix launches short-form video feature similar to TikTok				
"بايت دانس" تطور تطبيقًا شبيهاً بـ "كلوب هاوس"	37	S & T	*Annahar*	*Lebanon*
"byte dance" tatawur ttbyqana shbyhaan b "klub haws"				
Byte Dance develops a clubhouse-like app				
أداة ذكية تدمج بالهاتف لقياس الحرارة	38	S & T	*Al-Jazeera*	*Saudi Arabia*
'adat dhakiat tadmaj bialhatif liqias alharara				
A smart tool integrated into the phone to measure the temperature				
هياكل روبوتية خارجية تدعم قوة العمل البشرية	39	S & T	*Asharq Al-Awsat*	*Saudi Arabia*
hiakl rubutiat kharijiat tadeam quat aleamal albasharia				
Robotic exoskeletons support the human workforce				
حسام موافي يكشف عن سبب دقات القلب العالية	40	S & T	*Al-Ahram*	*Egypt*
husam mawafi yukshaf ean sbb daqqat alqalb alealia				
Hussam Mowafi reveals the cause of the high heart rate				

Online Parameter Estimation of DC-DC Converter through OPC Communication Channel

Mohammad A Obeidat[1]
Department of Electrical Power and Mechatronics Engineering
Tafila Technical University
Tafila, Jordan

Malek Al Anani[2]
Instrumentation and Control Engineering Department
SEPCO (Samra Electric Power Co.)
Amman, Jordan

Ayman M Mansour[3]
Department of Communication Electronics and Computer Engineering, Tafila Technical University, Tafila, Jordan

***Abstract*—System identification is a very powerful tool for determining the system model and parameters from sets of observable input and output data. Once the system parameters are obtained, the system dynamic behavior, including all the system characteristics (time constant, overshoot, settling time, etc.) can be accessed and evaluated. Despite the difficulty and communication channel lag, online parameter estimation outperforms offline system identification due to the ability to remotely monitor and control the system as well as improve the system's controller, making it more accurate and reliable. With the extreme development in technology, the importance of combining wireless networks with closed automatic control systems has emerged. This connection facilitates communication processes between the different units in the control for remotely controlled of the output. However, there are some errors affecting such system resulted from communication channel, A/D and D/A conversion process, identification process, or the existence of adaptive weight Gaussian noise. In this paper, the errors were investigated using real system, and then a suitable controller was tuned and optimized in order to reduce and eliminate various errors. The results show excellent dynamic behavior of the system under transmitting and receiving process.**

***Keywords*—Online parameters estimation; Open Platform Communication; OPC; communication channel; ARMAX model (autoregressive-moving average with exogenous terms); DC-DC Converter; chopper circuit**

I. Introduction

Since online system identification has a number of benefits, such as estimating parameters in real time while the system is operating as well as gaining the ability to monitor and operate the system remotely, it also has a number of challenges. These challenges include (1) analog system discretization with the proper sampling time depending on the system dynamics (2) use a communication channel to send and receive data (3) using the regression method to find the new approximate parameters based on the nature of the measured data (4) deploying the identification process algorithm to determine the new estimated parameters (5) fine-tune an appropriate controller to reduce or remove all identification process errors, this paper focuses on the online parameters estimation for a DC-DC chopper circuit by applying the measured input and output data to the identification algorithm via OPC (Open Platform Communication) after being modeled in the form of ARMAX model(autoregressive-moving average with exogenous terms).

II. DC-DC Chopper Circuit Models

The buck-boost type of the DC-DC Converter will be modeled in both continues and discrete domains.

Fig. 1 shows Buck/Boost converter circuit diagram.

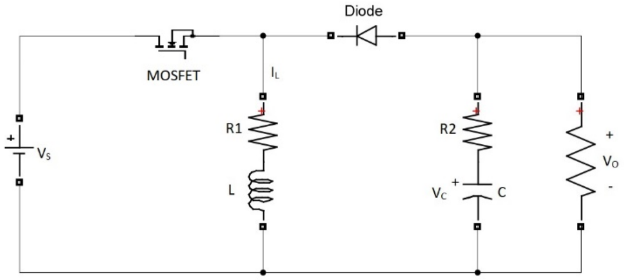


Fig. 1. Buck-Boost Converter Circuit.

Equation 1 describes the relation between input and output voltages which depends on the duty cycle d.

$$V_O = \frac{d}{1-d} * V_S$$

Mode 1: the switch is short circuit, and the diode is open circuit.

Mode 2: the switch is open circuit, and the diode is short circuit. Assume that the switch and the diode are ideal devices without delay [1, 2]. The state space representation matrices for mode 1 and 2 are:

- Mode 1

$\dot{x} = A_1 x + B_1 u$ and $y = C_1 x + D_1$

$$\mathbf{A_1} = \begin{bmatrix} -\frac{R_a}{L} & 0 \\ 0 & -\frac{1}{C} * \left(\frac{1}{R_O + R_b}\right) \end{bmatrix} \qquad \mathbf{B_1} = \begin{bmatrix} 1/L \\ 0 \end{bmatrix}$$

$$\mathbf{C_1} = \begin{bmatrix} 0 & 1 - \frac{R_b}{R_O + R_b} \end{bmatrix} \qquad \mathbf{D_1} = \mathbf{0}$$

- Mode 2

$\dot{x} = A_2 x + B_2 u$ and $y = C_2 x + D_2 u$

$$\mathbf{A_2} = \begin{bmatrix} -(\frac{R_a}{L} + \frac{R_O * R_b}{L*(R_O+R_b)}) & -\frac{R_b}{L*(R_O+R_b)} + \frac{1}{L} \\ -\frac{1}{C} * (\frac{R_O}{R_O+R_b}) & -\frac{1}{C} * (\frac{1}{R_O+R_b}) \end{bmatrix} \qquad \mathbf{B_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{C_2} = [-\frac{(R_O * R_b)}{R_O+R_b} \quad 1 - \frac{R_b}{R_O+R_b}] \qquad \mathbf{D_2 = 0}$$

Mode 1 for Buck/Boost converter is valid for dT_S and mode 2 is valid for $(1-d)T_S$, where T_s is the sampling period. After combining both modes equations using averaging technique and divide them by sampling period T_S, the system block diagram for the Buck/Boost circuit is shown in Fig. 2 [3,4].

In order to study the effect of disturbance on the system output, then MISO (Multi-input single output) system can be presented as two transfer functions, first one describes the relation between first input which is the disturbance and the output voltage, and the other transfer function describes the relation between the second input which is the reference voltage, and the output voltage [5-14]. Fig. 3 shows the MISO block diagram.

Laplace Transform is a useful tool for solving differential equations and analyzing analog systems. It enables us to examine stability through the use of simple pole-zero plots and to describe frequency response of systems through factorization. The Z Transform is a similar tool that can be used with digital signals. It saves us a lot of time manipulating difference equations. It will assist you in understanding the behavior and stability of a digital system.

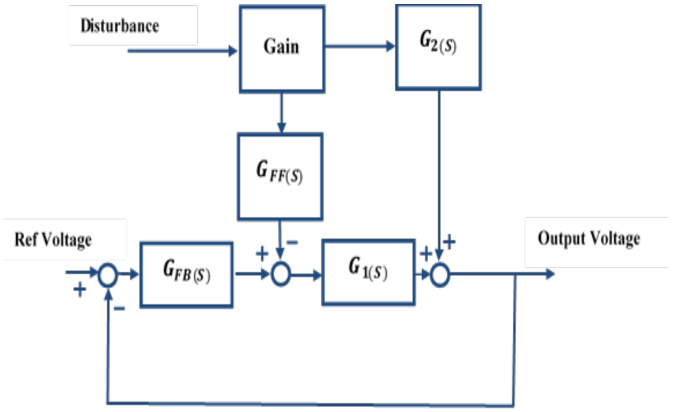


Fig. 2. Buck/Boost Circuit Block Diagram.

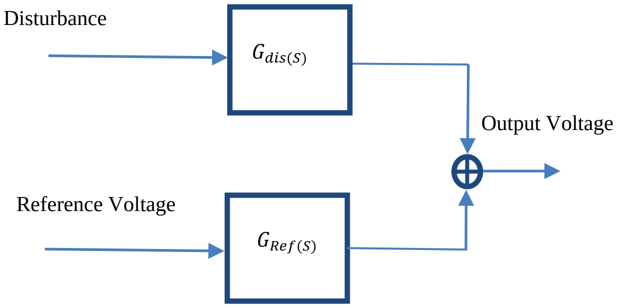


Fig. 3. MISO Block Diagram.

The unilateral Z-Transform of a digital sequence xn is given by.

$$Z(x_n) = X(z) = \sum_{n=0}^{\infty} x_n z^{-n}$$

The discrete state space model is:

$$x(k+1) = Gx(k) + Hu(k)$$

$$y(k) = Cx(k) + Du(k)$$

Table I shows the transformation from continuous to discrete.

TABLE I. TRANSFORMATION FROM CONTINUOUS TO DISCRETE

Continuous	discrete
$Ax(t)$	$Gx(k)$
$Bu(t)$	$Hu(k)$
C	C
D	D

The pulse transfer function will be obtained as following:

$$\frac{\mathrm{Y(z)}}{\mathrm{U(z)}} = \mathrm{C(zI-G)^{-1}H + D}$$

The structure of the Auto Regression Exogenous ARX model can be written as follows:

$$A(q)y(t) = B(q)u(t-nk) + e(t)$$

Where A(q) and B(q) are defined by:

$$A(q) = 1 + a_1 q^{-1} + \cdots + a_{nq} q^{-na}$$

$$B(q) = b_1 + b_2 q^{-1} + \cdots + b_{nb} q^{-nb+1}$$

where, $u(t)$ and $y(t)$ are respectively, the input and the output of the system, "t" is time unit, and q-1 represent the delay operator, $[qu(k) = u(k-1)]$.

III. Hardware and Software Configuration through OPC Communication Channels

This section illustrates experiment and hardware as well as the software configuration, the components of the real experiment are: Tow computers with network interface card (NIC), Ethernet cable cat5 with Rj45 connecters in both side And an Ethernet switch. Tow computers have been used here to make the identification process more realistic, the data are being transferred from one to the other and then received again.

OPC is a compatibility framework for transferring data securely and reliably in the industrial automation and other industries. It is platform that irrelevant and ensures a smooth flow of data between devices from various manufacturers, and OPC; it is based on the client server model as shown in Fig. 4. [15-18].

The OPC software used in this paper is KEPSERVER, this platform's design enables users to connect, manage, monitor, and control various automation devices and software applications via a single user interface [19].

OPC Client
OPC Server

Fig. 4. OPC Client and Server.

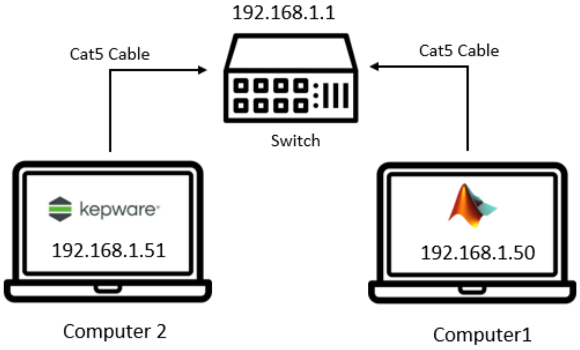


Fig. 5. Software and Hardware Topology.

Each one of the computers must be given a specific IP address through the Ethernet switch, one of the computers will be the OPC server "KRPSERVER" [20] and the other computer will be the OPC client "MATLAB" Fig. 5 shows the hardware and software topology [21].

As you can see from Fig. 5, both computers must be in same subnet mask OPC Client - computer 1 is configured with "192.168.1.50" IP address and OPC Server – computer 2 is configured with "192.168.1.51" IP address.

IV. Results and Discussion

The effect of each identification factor, including sampling time T_s, regression model type, number of parameters in A(q) and B(q), OPC communication channel effect, and the effect of adding noise to the input disturbance signal, will be investigated. The values of the DC-DC Buck/Boost converter circuit parameters are shown in the Table II.

The continuous transfer functions of the block diagram shown in Fig. 2 become:

G_1(s) between disturbance and Output voltage.

$$G_1(s) = \frac{0.74527\ (s + 7.5e04)(s - 1616)}{(s^2\ +\ 227.2s\ +\ 2.085e05)}$$

G_2(s) between $\hat{u}$ and $\hat{y}$

$$G_2(s) = \frac{-6.995\ (s + 5.757e04)}{(s^2\ +\ 227.2s\ +\ 2.085e05)}$$

The feedback controller G_{FB}(s):

$$G_{FB}(s) = \frac{S + 1789}{S + 7.757e04}$$

The feedforward controller G_{FF}(s):

$$G_{FF}(s) = \frac{-0.33}{S}$$

TABLE II. Buck-Boost Circuit Parameters

The parameter	The value
$\overline{V_{IN}}$	40 V
L	2.85*10^{-3} H
C	193*10^{-6} F
R_1	0.09 ohm
R_2	0.09 ohm
R_O	27.9 ohm
$\bar{d}$	2/3
Fs	60 HZ

All transfer functions will be converted to discrete form depending on the sampling time Ts. Each value of Ts gives a specific pulse transfer function, assuming Ts = 0.001second such that:

G_1(z) between $\hat{d}$ as input and $\hat{y}$ as the output in Z domain:

$$G_1(z) = \frac{0.74527(z + 11.52)(z - 7.534)}{(z\verb|^|2\ +\ 1.613z\ +\ 0.7968)}$$

G_2(z) between $\hat{u}$ as input and $\hat{y}$ as the output in Z domain:

$$G_2(z) = \frac{-0.1898(z + 0.8653)}{(z\verb|^|2\ +\ 1.613z\ +\ 0.7968)}$$

The feedback controller pulse transfer function G_{FB}(z):

$$G_{FB}(z) = \frac{Z - 0.9689}{z\ -\ 9.941e - 26}$$

The feedforward controller pulse transfer function G_{FF}(z):

$$G_{FF}(z) = \frac{-0.00033}{z - 1}$$

Fig. 6 shows the online identification process with all its steps, the blue dashed line is the OPC communication channel, while the black line indicates the MATLB system signal.

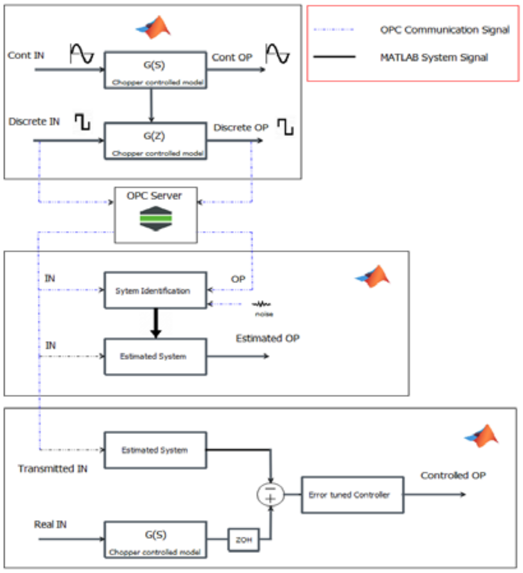


Fig. 6. Identification Process System Block Diagram.

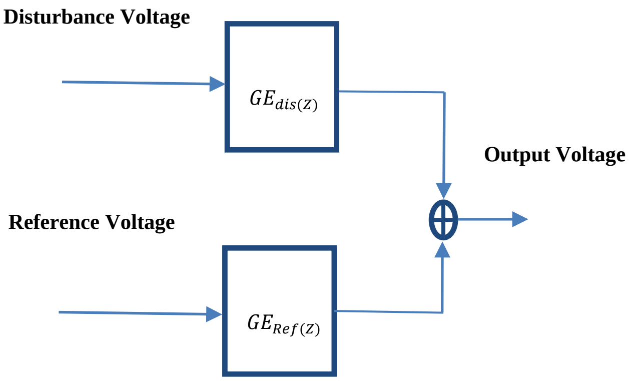


Fig. 7. Estemated Transfer Functions.

After the parameter's estimation of $G_{dis(S)}$ and $G_{Ref(S)}$ new estimated transfer functions are obtained $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ as shown in Fig. 7. Where, $GE_{dis(Z)}$ describe the relation between first input (disturbance) and output, and $GE_{Ref(Z)}$, describe the relation between second input (disturbance) and output.

A. Sampling Time Ts

In this section, sampling time will be investigated, and the effect of changing Ts on the identification process will be determined, the parameters value for the DC-DC Converter will be as in the Table II. The identification parameters are shown in Table III.

TABLE III. CONSTANT IDENTIFICATION FACTOR RELATED TO VARYING TS

The parameter	The value
Regression type	ARMAX
Regression parameter	A(q)=3 B(q) = 3
OPC	Scan Rate = same as Ts Communication remotely via ethernet cable
Noise	No noise

But sampling time Ts will be varying, the examples below will explain this methodology.

a) Example 1

In this example the sampling time will assumed to be 0.1 second, Fig. 8 shows the continues OP (output) and Discrete OP (output).

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table IV.

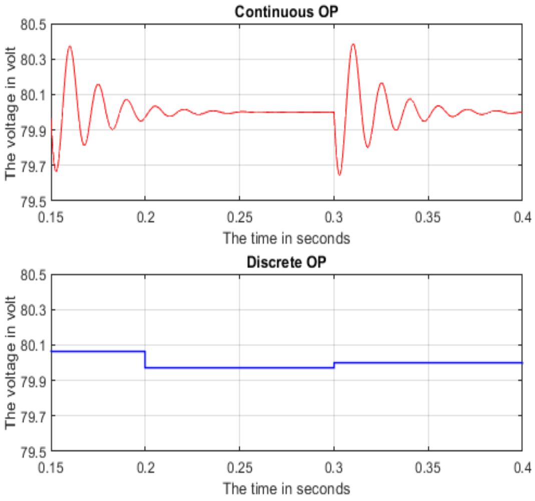


Fig. 8. Continuous and Discrete Outputs related to 0.1 Second Sampling Time.

Fig. 9 shows the Estimated OP(Output).

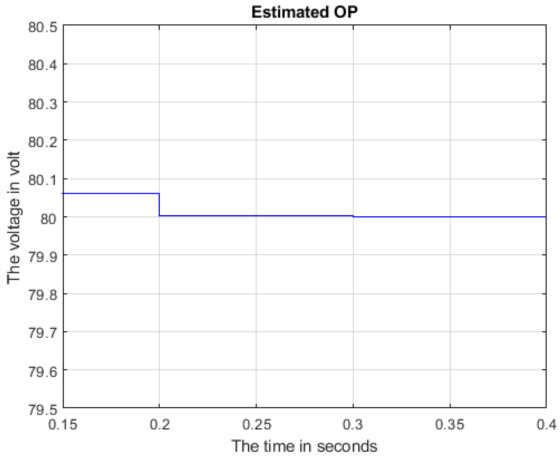


Fig. 9. Estimated Output related to 0.1 Second Sampling Time.

TABLE IV. ESTIMATED PARAMETERS RELATED TO 0.1 SECOND SAMPLING TIME

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-9.64E-08	-2.37E-05	3.6E-04	1.001
q^{-2}	2.2E-016	2.2E-016	2.6E-05	-3.6E-4
q^{-3}	2.2E-016	2.2E-016	2.2E-16	-2.6E-05

b) Example 2

In this example the sampling time will assumed to be 0.01 second, Fig. 10 shows the continues OP and Discrete OP.

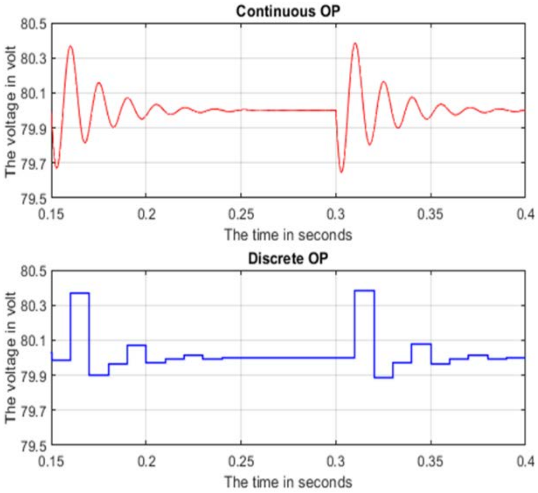


Fig. 10. Continuous and Discrete Outputs related to 0.01 second Sampling Time.

Fig. 11 shows the Estimated OP.

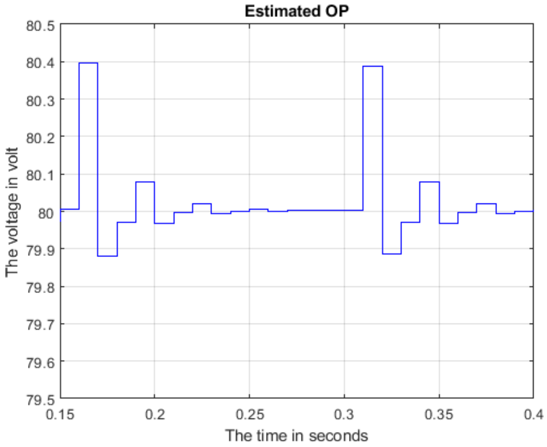


Fig. 11. Estimated Output related to 0.01 Second Sampling Time.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table V.

TABLE V. ESTIMATED PARAMETERS RELATED TO 0.01 SECOND SAMPLING TIME

q-operator	Estimated parameters in both transfer functions.			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	0.3443	0.048189	0.3356	0.8474
q^{-2}	0.1861	-0.04588	0.2025	0.2453
q^{-3}	-0.1129	-0.002297	-0.1114	0.334

c) Example 3

In this example the sampling time will assumed to be .001 second Fig. 12 shows the continues OP and Discrete OP:

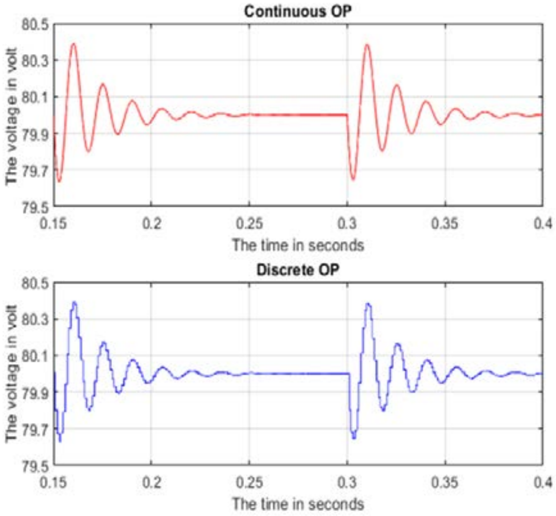


Fig. 12. Continuous and Discrete Outputs related to 0.001 Second Sampling Time.

Fig.13 shows the Estimated OP.

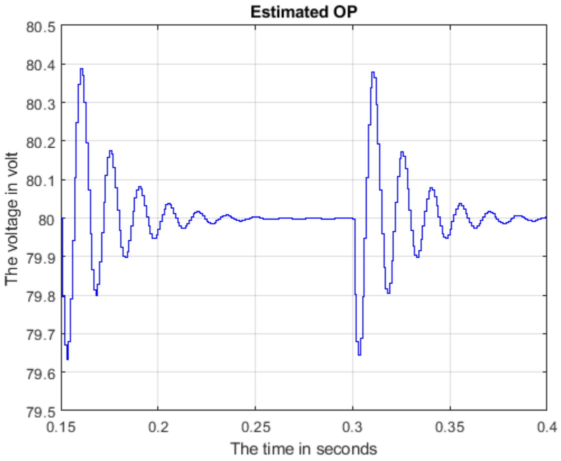


Fig. 13. Estimated Output related to 0.001 second sampling Time.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table VI.

TABLE VI. ESTIMATED PARAMETERS RELATED TO 0.001 SECOND SAMPLING TIME

q-operator	Estimated parameters in both transfer functions.			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-2.657	-0.0244	-2.617	-0.0029
q^{-2}	2.505	0.04952	2.434	0.01425
q^{-3}	-0.8339	-0.02513	-0.7972	0.00864

- Discussion

The conclusion is that as the sampling time increases, the identification error decreases and the identification process becomes more accurate, because as the sample time increases, more data from the system is analyzed and the identification loop has more data to estimate the parameter. In the case of Ts =0.001, we have 401 sets of values, which is sufficient to accurately estimate the parameter, whereas when Ts =0.01, we only have 41 sets of values, which is insufficient to estimate the parameters. On the other hand, choosing Ts to be very low will take a long time to complete the identification process, In the worst-case scenario, Ts=0.1 second leads to poor identification because the system completes its dynamic behavior and reaches steady state in 0.15 second, because when Ts=0.1 second, only two values are generated to describe the complete system's behavior, which is insufficient. Assuming Ts = 0.001, we have two advantages.

1) 1. provides a sufficient set of data to enable accurate identification.

2) It takes an acceptable amount of time, not too long but also not too fast.

Conclusion is that assuming Ts =.001 is better for this system, considering the response time and dynamic behavior for our system (DC Converter circuit), Table VII compares the sampling time with the system characteristic (overshoot, response time... etc.), delay time, and the identification error in general, also the ability to tune a controller to overcome the identification error.

Table VIII shows that as much as the sampling time increase high delay is added, extremely high identification error and the system characteristic and be measured or obtained.

B. *Number of Parameters*

The effect of changing the number of parameters will be investigated in this section, and thus the effect of changing the number of parameters on the identification process will be determined; the parameters value for the DC-DC Converter will be as shown in Table II.

TABLE VII. SAMPLING TIME EFFECT TO THE IDENTIFICATION PROCESS

TS(Second)	System characteristics	Identification error	Controller tuning	Delay time
0.1	Cannot be obtained	Extremely high	No controller	Very Hight more than 0.2 second
0.01	Can be obtained with high error	Hight	Controller can be tuned but error will still high	Hight around 0.02 second
0.001	Can be obtained highly accurate	Low	Controller can be tuned, and the error will be reduced	Low around 0.002 second

TABLE VIII. CONSTANT IDENTIFICATION FACTOR RELATED TO VARYING PARAMETERS NUMBER

The parameter	The value
Regression type	ARMAX
Sampling Time	Ts = 0.001
OPC	Scan Rate = same as Ts Communication remotely via ethernet cable
Noise	No noise

The identification parameters are shown in Table VIII.

But the parameters number will be varying, the examples below will explain this methodology.

a) Example 4

In this example the parameters number will assumed to be two in both A(q) and B(q), referring to Fig. 12 the continues OP and Discrete OP are shown.

Fig. 14 shows the Estimated OP.

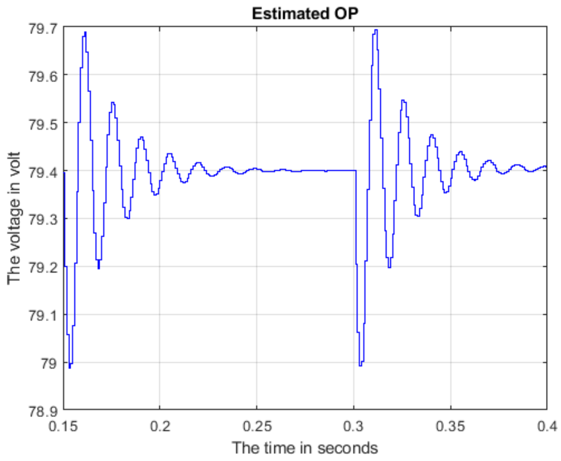


Fig. 14. Estimated OP related to Two Parameters.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table IX.

TABLE IX. ESTIMATED PARAMETERS RELATED TO TWO PARAMETERS

q-operator	Estimated parameters in both transfer functions.			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-1.732	-0.02445	-1.786	-0.0029
q^{-2}	0.9079	0.02445	0.8227	0.03975

b) Example 5

In this example the parameters number will assumed to be three in both A(q) and B(q), referring to Fig. 12 the continues OP and Discrete OP are shown.

Fig. 15 shows the Estimated OP.

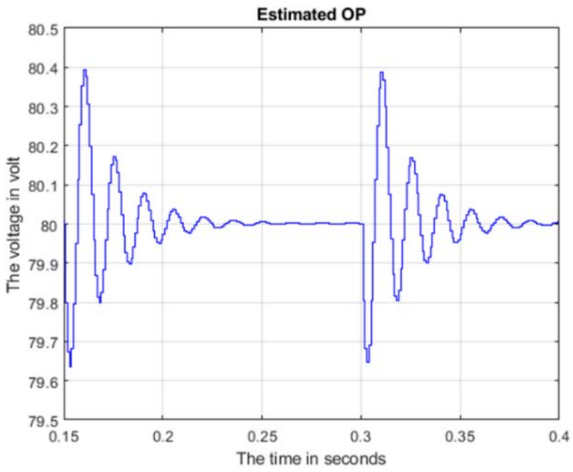


Fig. 15. Estimated OP related to Three Parameters.

Table X shows the estimated parameters in both transfer functions.

TABLE X. ESTIMATED PARAMETERS RELATED TO THREE PARAMETERS

q-operator	Estimated parameters in both transfer functions.			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-2.657	-0.0244	-2.617	-0.0029
q^{-2}	2.505	0.04952	2.434	0.01425
q^{-3}	-0.8339	-0.02513	-0.7972	0.00864

- Discussion

A conclusion is reached, which is that as the number of parameters increases, the identification error decreases and the identification process becomes more accurate, because as the number of parameters increases, more of the system's properties and characteristics can be measured. When the number of parameters is assumed to be two, thigh error is introduced into the system, and these two parameters cannot describe the system dynamic. However, when the number of parameters is three, the induced error is very low and can be ignored, resulting in good identification with all the original system properties being obtained well. On the other hand, as the number of parameters increases, more complexity is added to the system, making tuning a controller to reduce identification errors more difficult, in this case, selecting three parameters is sufficient, and the induced error is very low. Conclusion is that assuming number of parameters to be three is the better choice as it leads to good identification and low complexity.

Table XI compares the number of parameters with the system characteristic (overshoot, response time … etc.), Degree of complexity, the identification error in general and the ability to tune a controller to reduce or eliminate the identification errors.

TABLE XI. NUMBER OF PARAMETERS EFFECT TO THE IDENTIFICATION PROCESS

Number of parameters	System characteristics	Identification error	Controller tuning	Degree of complexity
2	Can be obtained with high error	high	No controller	Low complexity
3	Can be obtained with low error	Low	Controller can be tuned, and the error will be reduced	Low complexity

Table XI shows that as the number of parameters increases, good identification is obtained, but the system becomes very complex, so choosing three parameters results in an acceptable error and a low degree of complexity. Making a system with a low degree of complexity is preferable to making a system with a high degree of complexity as long as the error remains within acceptable limits.

C. Regression Model Type

In this section, the effect of changing the regression model will be investigated, and the effect of changing the model type either ARMAX or ARX to the identification process will be determined, the parameters value for the DC-DC Converter will be as in the Table II, the identification parameters is shown in Table XII.

TABLE XII. CONSTANT IDENTIFICATION FACTOR RELATED TO VARYING PARAMETERS NUMBER

The parameter	The value
Regression parameter	A(q)=3 B(q) = 3
Sampling Time	Ts = 0.001
OPC	Scan Rate = same as Ts Communication remotely via ethernet cable
Noise	No noise

But the regression model type will be varying, the examples below will explain this methodology.

a) Example 6

This example will compare the effect of both regression models ARMAX and ARX and the other factors as shown in Table XIV will be constant, Fig. 12 shows the continues OP and Discrete OP:

Fig. 16 shows the Estimated OP for ARMAX & ARX.

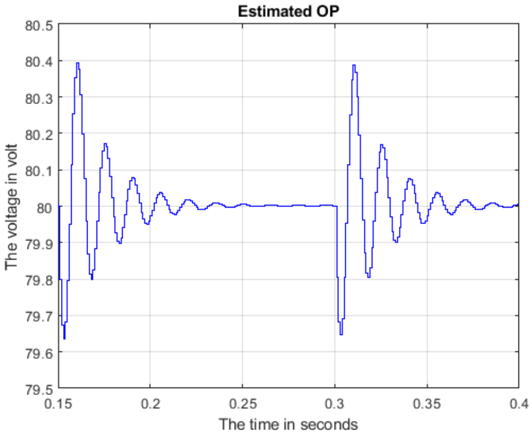


Fig. 16. Estimated OP for Both ARX and ARMAX.

After applying the identification algorithm for ARMAX and ARX, both models give the same parameters in both transfer functions $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ as shown in Table XIII.

TABLE XIII. ESTIMATED PARAMETERS IN BOTH ARX AND ARMAX

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-2.657	-0.0244	-2.617	-0.0029
q^{-2}	2.505	0.04952	2.434	0.01425
q^{-3}	-0.8339	-0.02513	-0.7972	0.00864

- Discussion

After comparing the results of ARX and ARMAX, the results were similar; both models give the same estimated OP as well as the same value of the estimated parameters as shown in Table XIII.

Results above show that either ARX or ARMAX can work properly with low identification error, so choosing either of them has no effect on the identification process; however, from ARMAX concept it would work more efficiently if the system became more complex with a large variation in the input or output signals because it combines the advantages of both MA (moving average) and AR (autoregressive).

D. *OPC Communication Channel*

In this section, the effect of the OPC communication channel will be investigated, the identification process will be working in two modes:

1) Local mode, where no communication channel is used and all the data analyzing will be locally in the same computer and in the same environment which is MATLAB 2020.

2) Remote mode, where the data is transferred and received from one computer to the other in the private network through the OPC communication channel via ethernet cable cat5.

The parameters value for the DC-DC Converter will be as in the Table II, the identification parameters are shown in Table XIV.

TABLE XIV. CONSTANT IDENTIFICATION FACTOR RELATED TO LOCAL AND REMOTE MODES

The parameter	The value
Regression parameter	A(q)=3 B(q) = 3
Sampling Time	Ts = 0.001
Regression Model	ARMAX
Noise	No noise

The identification error will be studied in both remote and local mode; also the communication channel effect will be obtained.

a) Example 7

The local mode identification process will be studied, Fig. 12 shows the continues OP and Discrete OP, Fig. 17 shows the comparison between the Discrete OP and Estimated OP in local mode.

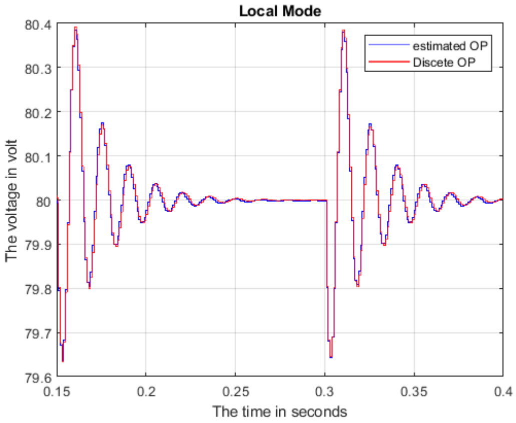


Fig. 17. Estimated OP in Local Mode.

As shown in Fig. 17, the estimated OP is very close to the discrete OP with a very low amount of error and an efficient identification, which is caused by processing the data very quickly with no delays because the local mode does not require sending or receiving data from one computer to another, and no hardwires (Ethernet cable cat5) are used.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table XV.

TABLE XV. ESTIMATED PARAMETERS RELATED TO LOCAL MODE

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-2.65837542	-0.0244	-2.6151791	-0.0029257
q^{-2}	2.508143613	0.0495739	2.43228732	0.01426002
q^{-3}	-0.8365384	-0.025169	-0.7969769	0.008797002

b) Example2

The remote identification process will be studied, Fig. 12 shows the continues OP and Discrete OP, Fig. 18 shows the comparison between the Discrete OP and Estimated OP in remote mode.

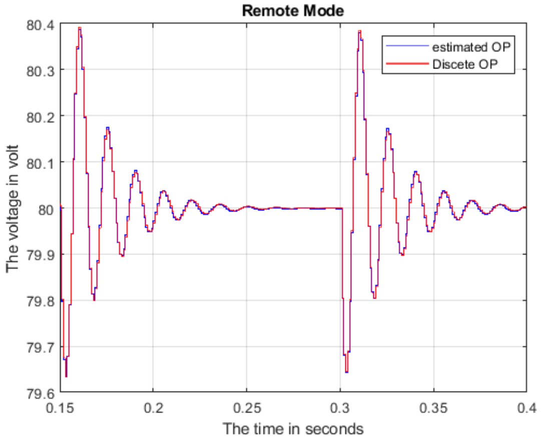


Fig. 18. Estimated OP in Remote Mode.

As shown in Fig. 18, the estimated OP is very close to the discrete OP with very amount error and an efficient identification. Using an OPC communication channel has no effect on the identification process because of the low error, and the communication channel was very fast and accurate due to the Ethernet cat5 high speed rate, high accuracy of the OPC communication channel and high-speed processing units.

The OPC communication channel error is caused by the delay time, as the OPC communication channel transmits the data as a query, it collects the data and convert it to a query then it sends it as packets over the Ethernet Cat5 cable, Due to high processing units and using cable with very high transition speed rate the communication channel does not add high error but rather a very small amount of error caused by the delay time.

Comparing the Discrete output with the received output is the key to determining the communication channel delay error. Fig. 19 shows MATLAB block for this comparison.

Fig. 19 shows MATLAB blocks for the comparing between received OP and Discrete OP, Fig. 20 shows the comparing results.

Fig. 20 shows the communication error, it clears from this figure that there is an error caused by the OPC communication channel.

Because the communication channel error is a delay time error, the best way to test the communication channel's performance is to add the same amount of delay to the discrete OP and then compare these signals again, as shown in Fig. 21.

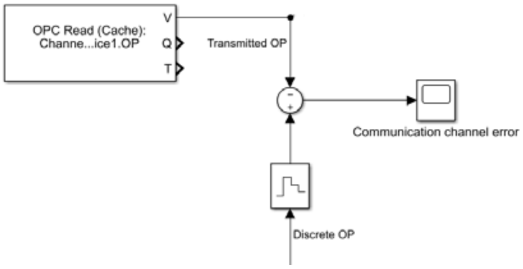


Fig. 19. Discrete and Received Outputs Comparison.

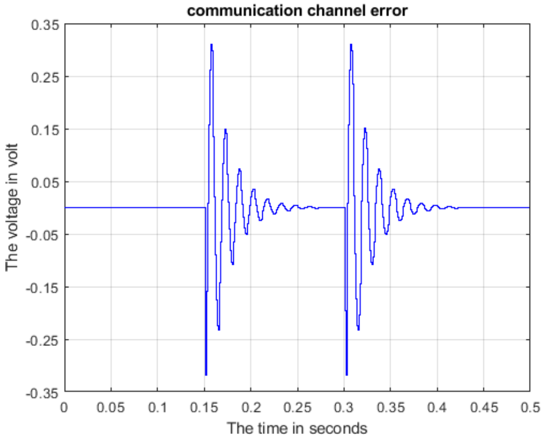


Fig. 20. Communication Error.

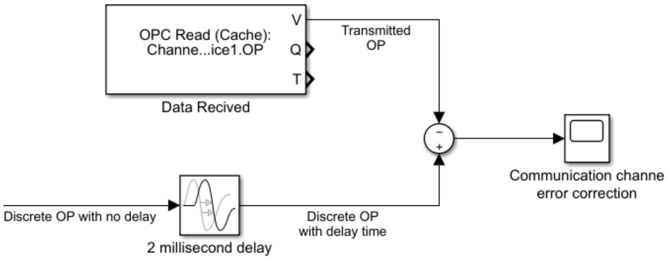


Fig. 21. Example of a Figure Caption.

Fig. 21 shows the MATLAB block ordering to compare the delayed Discrete OP signal with the received OP, Fig. 22 below shows the comparison result.

Fig. 22 shows that the communication channel error has been eliminated by adding 2 millisecond delay time to the discrete OP, as a result of this comparison, the OPC communication channel error is just 2 millisecond delay with neglectable noise, the noise as shown in the figure above are two pulses with very low amplitude the value of the noise amplitude is a power of -4 ($\times 10^{-4}$) which makes it neglectable and does not affect the identification process, the hardwire cable is the main cause for this noise.

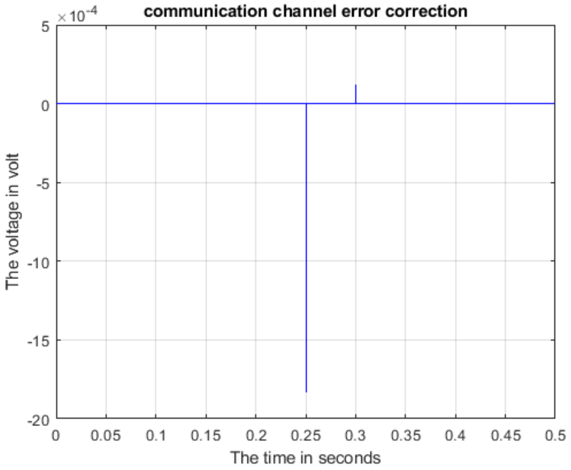


Fig. 22. Delay Time Error Cancellation.

TABLE XVI. ESTIMATED PARAMETERS RELATED TO REMOTE MODE

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-2.6572643	-0.024396	-2.616792	-0.0029133
q^{-2}	2.50481665	0.0495228	2.4338879	0.01424647
q^{-3}	-0.833855	-0.025127	-0.7971178	0.0086445

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table XVI.

- Discussion

The remote mode, which uses an OPC communication channel via a hardwire Ethernet cable cat5, results in a good identification process with very low error, as long as the transmitted signal has a short delay time, as a result of using a high-speed rate cable with a high computer processing unit performance, the communication channel error will be very small, and in some cases can be ignored.

Because the remote mode provides low communication channel error, the advantage of using it instead of the local mode is high, due to the remote-control ability that does not exist in the local mode. The conclusion is that using remote mode is better than using the local mode because of the low error presence as well as the ability to remotely control and observe the system via the network.

E. *Noisy Input Disturbance Signal*

The effect of changing the noise amplitude and frequency at the input disturbance signal will be investigated in this section, the output signal will be noise free, as will the reference input signal, and thus the effect of changing the noise configuration to the identification process will be determined, The DC-DC Converter parameters will be as shown in Table II.

The identification parameters are shown in Table XVII.

TABLE XVII. CONSTANT IDENTIFICATION FACTOR RELATED NOISY INPUT SIGNAL

The parameter	The value
Regression parameter	A(q)=3 B(q) = 3
Sampling Time	Ts = 0.001
OPC	Scan Rate = same as Ts Communication remotely via ethernet cable
Regression model type	ARMAX

Fig. 23 shows the MATLAB blocks for the noisy input signal.

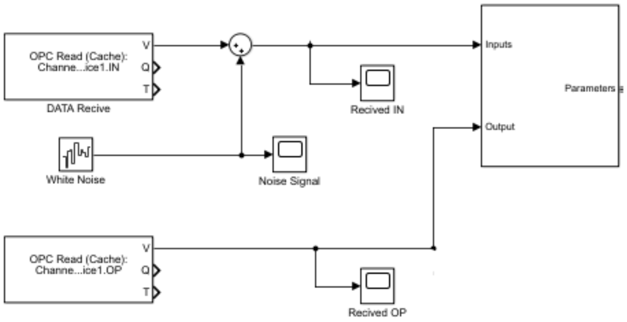


Fig. 23. MATLAB Ordering Blocks for Noisy Input Signal.

As illustrated in Fig. 23 the input signal with adding noise is applied to the identification algorithm, on the other hand the output signal and the reference input signal still noise free signals.

But the noise configuration will be varying, the examples below will explain this methodology.

a) Example 8

In this example noise will assumed to be 0.01 Power spectral density with 100 Hz, the following Fig. 24 shows the Received communicated noise input signal with 0.01 PSD and 100 HZ frequency compared with the (not communicated signal) MATLAB input signal, the input signal here is the disturbance signal to the converter circuit and the other input which is the reference signal will be noise free.

Fig. 25 shows the Estimated OP and the Discrete OP.

It is clear from Fig. 25 that the estimated OP deviates greatly from the discrete OP, so that as the noise PSD and frequency increase, the identification process efficiency decreases, and high error is added to the system. From Fig. 25, the system characteristics (settling time, overshoot, oscillations number.... etc.) can be measured but with high error.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table XVIII.

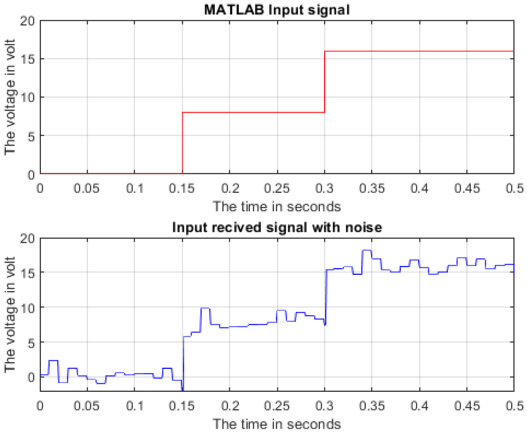


Fig. 24. Noisy Input related to 0.01 PSD and 100 HZ Frequncy.

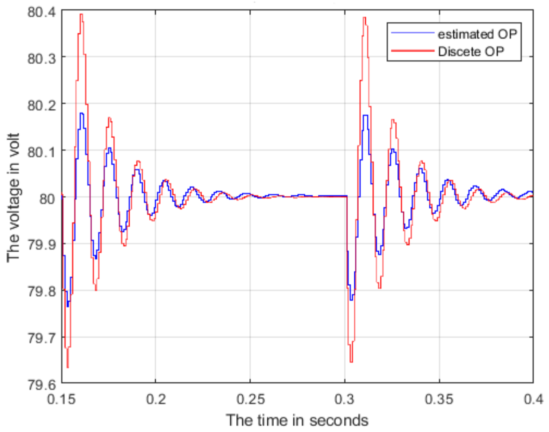


Fig. 25. Estimated and Discrete OP related to 0.01 PSD and 100 HZ Frequncy.

TABLE XVIII. ESTIMATED PARAMETERS RELATED TO 0.01 PSD AND 100HZ FREQUENCY

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-1.80	-0.014765	-2.6169	-0.0029
q^{-2}	1.0196	0.0173581	2.4341	0.0142
q^{-3}	-0.0501	-0.00252	-0.7972	0.0086

b) Example 9

In this example noise will assumed to be 0.03 Power spectral density with 200 Hz, the flowing Fig. 26 shows the Received communicated noise input signal with 0.03 PSD and 200 HZ frequency compared with the (not communicated signal) MATLAB input signal, the input signal here is the disturbance signal to the converter circuit and the other input which is the reference signal will be noise free.

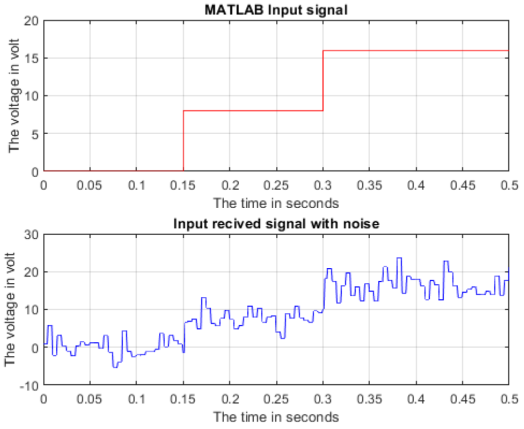


Fig. 26. Noisy Input related to 0.03 PSD and 200 HZ Frequncy.

Fig. 27 shows the Estimated OP and the Discrete OP.

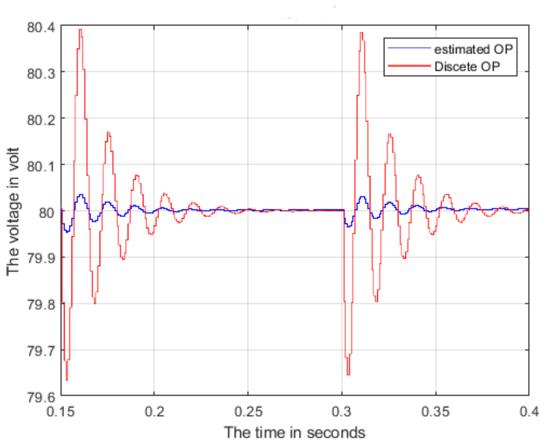


Fig. 27. Estimated and Discrete OP related to 0.03 PSD and 200 HZ Frequncy.

It is clear from Fig. 27 that the estimated OP deviates greatly from the discrete OP, so that as the noise PSD and frequency increase, the identification process efficiency decreases, and high error is added to the system. From Fig. 36, the system characteristics (settling time, overshoot, oscillations number, etc.) can be measured but with even higher error from previous example.

The estimated parameters for both transfer function $GE_{dis(Z)}$ and $GE_{Ref(Z)}$ are shown in Table XIX.

TABLE XIX. ESTIMATED PARAMETERS RELATED TO 0.03 PSD AND 200HZ FREQUENCY

q-operator	Estimated parameters in both transfer functions			
	$GE_{dis(q)}$		$GE_{Ref(q)}$	
	$A(q)$	$B(q)$	$A(q)$	$B(q)$
q^0	1	0	1	0
q^{-1}	-1.7041	-0.0023	-2.6169	-0.0029
q^{-2}	0.8616	0.0024	2.4341	0.0142
q^{-3}	0.02905	-8.1281e05	-0.7972	0.0086

- Discussion

From the previous examples, it is clear that as the noise amplitude and frequency increase, so does the identification error, the identification error caused by noise can be seen as compression in the Y-axis; as the noise amplitude and frequency increase, the response will compress more; however, the system dynamic remains unchanged, so this error can be reduced or eliminated by tuning a proper controller that makes the identification process efficient.

Also, as shown in the Table XIX it clears that the parameters in both A(q)&B(q) change only in $GE_{dis(q)}$ but the $GE_{Ref(q)}$ remains the same in all example, the reasons of that is coming from adding the noise to the disturbance signal only making only the transfer $GE_{dis(q)}$ changes as this transfer function describe the relation between the disturbance and the output, so that it is logical to change every time the disturbance noisy signal changes.

On the other hand, $GE_{Ref(q)}$ which describe the relation between the reference input signal and the output voltage does not change as the input reference signal still the same in all examples with no noise added.

Future work can be conducted in many directions and applications. The use of OPC communication channel would have interesting future directions in many applications such as power [22-25], health [26-31], communication [32-37], AI applications [38-43], and optimization [44,45].

V. Conclusion

The identification process is highly dependent on the error associated with each factor. To determine whether the identification process is efficient or not, the effects or error associated with each factor are investigated to determine whether the identification process can truly provide a good estimate or not. As a result, before proceeding to the online detection, these factors must be optimized to produce the least amount of error possible, ensuring that the best parameters have been estimated.

Acknowledgment

The authors would like to express their gratitude to the support of Tafila Technical University as well as the support of the technical and administrative team at Samra Electric Power Company SEPCO, especially Eng. Mohammad Shiab and Eng. Yousef Balasya.

References

[1] M. A. Obeidat, L. Y. Wang and Feng Lin, "On-line parameter estimation of PMDC motors using binary-valued speed measurements," 2012 IEEE Power and Energy Conference at Illinois, 2012, pp. 1-5, doi: 10.1109/PECI.2012.6184608.

[2] Obeidat, Mohammad, and Ali Hamad. "Applying two controller schemes to improve input tracking and noise reduction in DC-DC converters." Przegląd Elektrotechniczny 95 (2019).

[3] M. A. Obeidat, L. Y. Wang and F. Lin, "Online parameter estimation of PMDC motors using quantized output observations," 2012 IEEE Transportation Electrification Conference and Expo (ITEC), 2012, pp. 1-6, doi: 10.1109/ITEC.2012.6243416.

[4] M. A. Obeidat, L. Y. Wang and F. Lin, "Real-Time Parameter Estimation of PMDC Motors Using Quantized Sensors," in IEEE Transactions on Vehicular Technology, vol. 62, no. 7, pp. 2977-2986, Sept. 2013, doi: 10.1109/TVT.2013.2251431.

[5] Lopa, Shafinaz A., et al. "Design and simulation of DC-DC converters." International Research Journal of Engineering and Technology (IRJET) 3.01 (2016): 63-70.

[6] Raja MAZ and Chaudhary NI. Two-stage fractional least mean square identification algorithm for parameter estimation of CARMA systems. Signal Process 2015; 107: 327–339.

[7] Xu, Ling. "The parameter estimation algorithms based on the dynamical response measurement data." *Advances in Mechanical Engineering* 9.11 (2017): 1687814017730003.

[8] Diversi, Roberto, Roberto Guidorzi, and Umberto Soverini. "Identification of ARMAX models with noisy input and output." IFAC Proceedings Volumes 44.1 (2011): 13121-13126.

[9] Rachad, Sofia, Hicham Fouraiji, and Bahloul Bensassi. "Identification approach for a production system using ARX model." 2014 International Conference on Logistics Operations Management. IEEE, 2014.

[10] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379–423 and 623–656, (July and October, 1948)

[11] Ali ibraheem , "DC Chopper Circuits Modeling and Behaviour using Feedback and Feedforward Control",(2019)Tafila technical university

[12] linear approximation of a nonlinear function,(2021),x-engineer,retrieved from x-engineer.org

[13] Al-Alaoui, Mohamad Adnan. "Novel stable higher order s-to-z transforms." *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 48.11 (2001): 1326-1329.

[14] Analytic Sciences Corporation. Technical Staff. (1974). Applied optimal estimation. Gelb, Arthur, 1937-. Cambridge, Mass.: M.I.T. Press. pp. 121. ISBN 0-262-20027-9. OCLC 960061

[15] R.H. Middleton & G.C. Goodwin (1990). Digital control and estimation: a unified approach. p. 33f. ISBN 978-0132116657.

[16] OPC Foundation and MTConnect Institute Announce a Memorandum of Understanding". OPC Foundation. 2010-10-21. Archived from the opcfoundation.org on 2011-06-16. Retrieved 2010-10-26.

[17] Nicola, Marcel, Claudiu-Ionel Nicola, and M. Duta. "SCADA Systems Architecture Based on OPC and Web Servers and Integration of Applications for Industrial Process Control." International Journal of Control Science and Engineering 8.1 (2018): 13-21.

[18] Ayatollahi, Iman, et al. "Prototype OPC UA server for remote control of machine tools." Proceedings of International Conference on Innovative Technologies. 2013.

[19] Rohani, Mohd Faiz, et al. "OPC Protocol Application for Real-Time Carbon Monitoring System for Industrial Environment." International Journal of Electrical & Computer Engineering (2088-8708) 7.2 (2017).

[20] KEPSERER OPC Server software,(2021),Kepware, Retrieved from www.kepware.com

[21] OPC Toolbox OPC Toolbox Read and write data from OPC servers and data historians,(2021),Mathworks ,Retrieved from /www.mathworks.com

[22] Mansour, A.M., Abdallah, J., Obeidat, M.A.,” An efficient intelligent power detection method for photovoltaic system,” International Journal of Circuits, Systems and Signal Processing, vol. 14, pp. 686–699, 2020.

[23] M. A. Obeidat, M. Qawaqneh, A. M. Mansour and J. Abdallah, "Smart Distribution System using Fuzzy Logic Control," 2021 12th International Renewable Engineering Conference (IREC), Amman, Jordan, 2021, pp. 1-5.

[24] M. A. Obeidat, A. M. Mansour, B. Al Omaireen, J. Abdallah, F. Khazalah and M. Alaqtash, "A Deep Review and Analysis of Artificial Neural Network Use in Power Application with Further Recommendation and Future Direction," 2021 12th International Renewable Engineering Conference (IREC), Amman, Jordan, 2021, pp. 1-5.

[25] A. M. Mansour, M. A. Obeidat and J. Abdallah, "A Novel Multi-agent Mechanism for Managing Electrical Power Smart Grids," 2021 12th International Renewable Engineering Conference (IREC), Amman, Jordan, 2021, pp. 1-6.

[26] Ayman M. Mansour, Murad M. Alaqtash, Mohammad Obeidat “Intelligent Classifiers of EEG Signals for Epilepsy Detection,” WSEAS Transactions on Signal Processing, vol. 15, 2019.

[27] Murad Alaqtash, Ayman M Mansour, Mohammad Obeidat, “Fuzzy Assessment Model for Functional Impairments in Human Locomotion”. IOSR-JECE, vol. 14, no. 1, Jan-Feb 2019.

[28] Ayman M. Mansour, “Intelligent E-Health System for Patient and Elderly People Monitoring Using Multi Agents System,” Jordan Journal of Electrical Engineering, vol. 4, no. 1, 2018.

[29] Ayman M. Mansour, “Decision Tree-Based Expert System for Adverse Drug Reaction Detection using Fuzzy Logic and Genetic Algorithm,” International Journal of Advanced Computer Research (IJACR), vol. 8, no. 36, 2018.

[30] Mohammad A. Obeidat and Ayman M. Mansour, “EEG Based Epilepsy Diagnosis System using Reconstruction Phase Space and Naïve Bayes Classifier,” WSEAS Transactions on Circuits and Systems, vol. 17, 2018.

[31] Mansour, A.M., Obaidat, M.A. and Hawashin, B. Elderly people health monitoring system using fuzzy rule based approach. International Journal of Advanced Computer Research, vol. 4, no. 4, p.904. 2014.

[32] Ayman M. Mansour, “GSM based Vehicle-to-Vehicle Communication using Multi-Agent Intelligent System,” WSEAS Transactions on Electronics, vol. 10, 2019.

[33] Ayman M Mansour, “ Cooperative Multi-Agent Vehicle-to-Vehicle Wireless Network in a Noisy Environment,” International Journal of Circuits, Systems and Signal Processing, vol. 15, 2021.

[34] Obeidat, M. A. Real-Time DC Servomotor Identification and Control of Mechanical Braking System for Vehicle to Vehicle Communication. International Journal of Computer Applications, 975, 8887.

[35] A. Mansour, H. Ying, P. Dews, Y. Ji and R. M. Massanari, "Identifying adverse drug reaction signal pairs by a multi-agent intelligent system with fuzzy decision model," 2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), Berkeley, CA, USA, 2012, pp. 1-6.

[36] A. Mansour et al., "Finding similar patients in a multi-agent environment," 2011 Annual Meeting of the North American Fuzzy Information Processing Society, El Paso, TX, USA, 2011, pp. 1-6.

[37] A. Mansour et al., "A multi-agent system for detecting adverse drug reactions," 2010 Annual Meeting of the North American Fuzzy Information Processing Society, Toronto, ON, Canada, 2010, pp. 1-6.

[38] Ayman M. Mansour, “Texture Classification using Naïve Bayes Classifier,” International Journal of Computer Science and Network Security (IJCSNS), vol. 18, no. 1, January 2018

[39] D.A. Al Nadi and Ayman Mansour, "Independent Component Analysis (ICA) for texture classification", 5th International Multi-Conference on Signals and Devices, IEEE SSD, 2008.

[40] B. Hawashin et al., "Efficient Texture Classification Using Independent Component Analysis," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 544-547, 2019.

[41] Hawashin, B., Fotouhi, F. and Grosky, W., 2010, December. Diffusion maps: A superior semantic method to improve similarity join performance. In 2010 IEEE International Conference on Data Mining Workshops (pp. 9-16). IEEE.

[42] Hawashin, B., Aqel, D., AlZu’bi, S. and Jararweh, Y., 2019, June. Novel weighted interest similarity measurement for recommender systems using rating timestamp. In 2019 Sixth International Conference on Software Defined Systems (SDS) (pp. 166-170). IEEE.

[43] Hawashin, B., Aqel, D., Alzubi, S. and Elbes, M., 2020. Improving recommender systems using co-appearing and semantically correlated user interests. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 13(2), pp.240-247.

[44] Suhail Sharadqah, Ayman M Mansour, Mohammad A Obeidat, Ramiro Marbello and Soraya Mercedes Perez, “Nonlinear Rainfall Yearly Prediction based on Autoregressive Artificial Neural Networks Model in Central Jordan using Data Records: 1938-2018” International Journal of Advanced Computer Science and Applications(IJACSA), 12(2), 2021.

[45] Jafar.Abu Khait, Ayman M Mansour and Mohammad Obeidat, “Classification based on Gaussian-kernel Support Vector Machine with Adaptive Fuzzy Inference System,” Przegląd Elektrotechniczny., vol 5, pp 16-24, 2018.

Image Contrast Optimization using Local Color Correction and Fuzzy Intensification

Avadhesh Kumar Dixit[1]
Research Scholar, Deptt. of CSE
IFTM University, Moradabad, India

Rakesh Kumar Yadav[2]
Assistant Professor, Deptt. of CSE
IFTM University, Moradabad, India

Ramapati Mishra[3]
Professor, Deptt. of ECE
IET, Dr. RMLAU, Ayodhya, India

***Abstract*—Global image enhancement techniques are used to enhance contrast in images but these techniques are found to be under-enhanced or over-enhanced in differently illuminated regions of the image. Local color correction methods work on local pixel regions to optimize the color contrast enhancement but they also have been found to show a lag while covering pixel regions which are overexposed, compared to those which are underexposed causing local artifacts. In this work, we overcome the shortcomings of both the local color correction and global color correction. This method uses local color correction in the Hue Saturation Luminance (HSL) domain, and fuzzy intensification operators are used to control the color fidelity of the local color corrected images. Thus, is able to sort out the problem of overexposed and underexposed regions and provide optimized contrast enhancement in colored images. Several experiments have been performed to analyze the performance of the proposed method and feasibility as compared to existing techniques. Performance parameters such as Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measurement (SSIM) and Naturalness Image Quality Evaluator (NIQE) is evaluated and the comparison with some existing techniques of contrast enhancement of color images is performed. The obtained result have good contrast and approve the better performance of the proposed method in support of the quantitative measure of perceptual appearance of the processed images and low computational time.**

Keywords—Contrast enhancement; local color correction; fuzzy operators; optimization

I. Introduction

Image enhancement is one of the most important steps in any image processing algorithm whether it is pattern detection, image classification or biometric recognition. Most of the computer vision based algorithms consist of image quality enhancement steps at some stages. The image quality is also strongly affected by the image acquisition device's quality. A higher-resolution camera is needed for better image quality, which can significantly increase the system's cost. Several methods of image enhancement exist which can improve the characteristics of the images to acceptable level even if they have been acquired with low quality cameras. Considering the factors like domain, environment, quality of images etc. different mathematical arbitrations can be applied on the images to improve their characteristics. Image correction may include various steps such as saturation, sharpness, denotation, tonal shift, tonal equilibrium and contrast correction. Even from the perspective of humans, the perceptibility of images is dependent on the contrast of the images. Contrast generally refers to the difference between different pixel levels based on their intensity. Compared to the other objects and the context, the difference between visual properties makes the parts of images distinctive. The sensitivity to human contrast depends on the spatial frequency; thus, when determining the contrast the spatial content of the picture should be considered. Local contrast is defined in accordance with the local luminance and background lights as a function of every point in the image as well as in every frequency band [1]. The variation in color and brightness of the element and other items within the same field of view is calculated to provide the visual perception of the real world.

The techniques of contrast enhancement can usually be divided into two categories: direct and indirect. Direct methods define a function based on the contrast whereas indirect methods don’t define a specific contrast function. The direct and indirect methods are categorized further as spatial methods in which pixel level operations are performed and frequency domain methods operating in transformation of images [1]. Color image enhancement poses more challenges as compared to a grayscale image enhancement because of the presence of three color intensity levels viz. red, green and blue as compared to a single intensity level in case of grayscale images. Thus techniques like histogram equalization (HE) which gives a fast image enhancement with a grayscale image cannot be directly used for color image enhancement. Histogram Equalization, however, is not widely used in consumer electronics because it can lead to irregularities such as changes in excessive brightness, vibration and saturation of light [2]. One of the popular techniques from the family of general histogram modification techniques are gamma correction techniques accomplished by using a particular adaptive parameter “γ”, defined as below [3]:

$$T(l) = l_{max}(l/l_{max})^{\gamma} \tag{1}$$

where “l_{max}” is the maximum intensity of the input.

The gamma curves illustrated in Fig. 1 show that “ γ >1” has exactly the opposite effect as those generated with “γ < 1”. Global correction methods like gamma correction suffer from this problem of uneven distribution of pixel color intensity.

Many algorithms concentrate on improving the image quality, but usually lead to an unnatural appearance such as confusion and objects with light sources. Many others therefore seek to reduce over-improvement at the cost of accuracy [4]. Since traditional histogram algorithms can result in over-enhancement, several algorithms have been suggested, including limitation of brightness and contrast. Maintaining efficiency in applications requires the preservation of light. The preservation of luminosity in areas of diminished intensity, such as dark regions, is disadvantageous to detailed enhancement in uniform illumination images [4]. An improvement over the traditional HE is the brightness preserving dynamic histogram equalization (BPDHE) algorithm, which maintains mean brightness of the image by taking the same mean brightness of the output image as that of the input image [5].

In local enhancement techniques, transformation of image pixels is dependent on neighboring pixel information. A small window is moved over every pixel of the image and only those blocks of images are enhanced, whose pixel falls under the window. Local information is preserved and brightness is improved for future use. Therefore contrast ratio can be smoothly improved in all parts of the image. Global knowledge about luminosity is missing and local objects can be created. Compared to the global enhancement methods, code complexities of these methods are high. Hybrid strategies provide global and local strategies for development. Here, adjacent pixels and global image information are taken into account during the transformation [6].

The need to improve the image visual quality that is produced under intense lighting conditions resulted in the development of an overexposure and underexposure fusion of blurred enhanced images. This can be done with the Gaussian membership function by means of a smooth image creation process. For further optimization, Discrete Wavelet Transform (DWT) based fusion of these fuzzy enhanced images is possible. The improved Fuzzy algorithm reduces noise, preserves information and improves image contrast [7]. Nonlinear and knowledge-based fuzzy techniques are feasible, creating a new technique for the enhancement of contrast.

In a dehazing based enhancement model suggested by Dong et. al [8], a photometric negative of the input image resembling hazy image has been obtained. The enhancement is then achieved by dehazing the hazy image effects, but this approach had issues with blocking artefacts, and the images produced seemed artificial or sketchy.

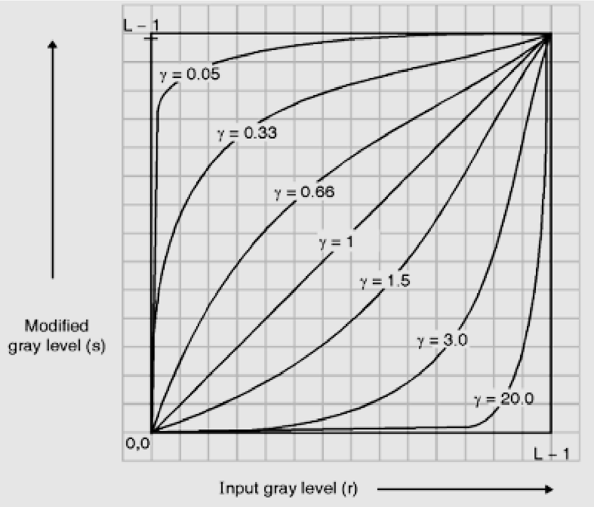


Fig. 1. Gamma Variation.

Retinex based methods have been developed which model human vision of color and luminance. Adaptive Multi Scale Retinex (AMSR) achieves better illumination and tone fluency by adding weights on the single scale retinex of the input image [9] over the various retinex methods such as Single Scale Retinex and, multi scale retinex.

Differential grey level histogram for Color images (DHECI) [10], generates two differential grey level histograms one for the intensity and the other for saturation level. The enhanced image is obtained as a weighted sum of these two levels controlled by a human color perception parameter.

Retinex based methods do not consider the noise factor while increasing the luminosity of the images. The joint denoising and enhancement(JED) technique [11], is an integrated approach providing both enhancement and denoising by combining both the luminance map estimation and the reflectance map in sequential steps, thus able to remove noise and obtain better enhancement.

In this work, a novel framework for an optimized image enhancement has been presented, which is modeled to include both the properties of local enhancement as well as global enhancement. Local color correction provides the local contrast enhancement and then fuzzy intensification operators are used to provide global color enhancement. The upcoming sections consist of literature review of previous work done in the field of color image enhancement in Section 2, description of proposed methodology and algorithms is presented in Section 3, results and discussion is presented in Section 4 and conclusion is presented in Section 5.

II. Literature Review

Chun-Tsai et al. [12] suggested to use the fuzzy logic concept and illumination analysis to correct color images of face. The images are divided by a fuzzy logic classification scheme into back lit, normal lit and front lit images. Lighting has been evaluated for the distributions of image illumination, over the input image. There are a growing array of peaks and valleys. The properties of this piecewise linear transformation are these peaks and valleys. The results showed that in comparison with other available methods, the method is efficient and effective.

The automated self-contained method for determining contrast gain function for automatic contrast improvement was suggested by Iyad F. Jafar et al. [13]. Contrast gain functions abbreviated by FACE(fuzzy adaptive contrast enhancement) adapted and modified smoothly in conjunction with the neighbourhood characteristics of the pixels, resulting into low, medium, and high activity gain. The Gain Function is defined by local image statistics, Fuzzy C-means (FCM) clustering and a Fuzzy Inference System (FIS) based on rule-based processing on a wide spectrum of pictures. In comparison with other methods, the process produced significantly improved contrast images and less artifacts.

A novel fusion-based technique was introduced by Amina Saleem et al. [14], to improve both grayscale and color images. The results show that the proposed algorithm is effective in strengthening local and global contrasts, reducing saturation and over-improvement while retaining the original imagery. The fusion-based enhancement method is ideal for non-real - time image processing applications that require high-quality images.

Mohan Liu et al [15] introduced a new conceptual contrast metric which follows the traditional background and foreground energy model and the energy used to calculate the technological luminance contrast instead of the maximum, minimum, and average luminance intensities. The authors also demonstrate that the chrominance information can also be used to assess the contrast quality of images. The technical measure of the luminance contrast is mapped onto a perceptual-based measure using a human attention model. This method yields massive enhancements relative to the state-of-the art, but at the expense of increased complexity.

A new improvement approach to both images and videos was introduced by Shih-Chia Huang et al [16]. In addition to probability and statistical inferences, the histogram analysis gives the special knowledge of a single image. The weighting distribution in the second step is used to smooth the fluctuating tendency and thereby avoid the creation of objects which are unfavorable. The picture contrast automatically improved by a smoothing curve in the third step. For multiple frames in a video set, writers used temporal information to reduce the computation time. The entropy model was used to decide if the transformation curve would be modified or not depending on the variation in the information material.

V. Magudeeswaran et al. [17] proposed the Fuzzy logical histogram equalization method which provided, the enhanced image contrast with the inaccuracy of gray-level values by the use of numerically produced fuzzy data. It is an efficient way to improve performance in comparison with classic flat histograms. The color quality, brightness and image contrast can also be enhanced automated. The approach suggested, efficiently removes the washed-out and adverse objects caused by many current approaches.

Jeyong Shin et al. [18] proposed a new histogram-based approach of enhancing contrast, to maintain the position of the initial histogram and to improve global contrast. The question of optimization is introduced where the histogram incorporates possible local conditions in order to increase the contrast to the location. The method makes a smooth change in comparison to different histogram profile images by creating clearer comparing images of local histograms.

Chin Yeow Wong et al. [19] used optimized chosen hyperbolic tangent profiles to eliminate unnecessary artifacts. A series of experiments were conducted to improve color image quality using qualitative evaluations and quantifiable measures such as entropy, gradients, colorfulness and saturation. The pipeline approach has been stressed as increasing image quality, enhancing contrast and sharpening color images, without losing color and saturation. The method is easy to use and sufficient for systematic color image preprocessing.

Zohair Al Ameen [20] developed a system using tuned blurred intensification operators to rapidly filter bad photographs taken in a dusty weather. The processing of images was effective in producing appropriate colors and fine details, as has been showed by tests on different images.

A new method to enhance image contrast based on local histogram equalization was implemented by M. Shakeri et al [21]. Originally, the image was divided into several sub-images with a clustering of the brightness values, where clusters are calculated using an automatic method based on histogram analysis of the two images. Every cluster of the image brilliance levels represented a spectrum in the histogram which was added to each histogram segment with the transfer feature introduced; the improved image obtained across the segments.

Zohair Al-Ameen [22] introduced a new, flexible stretching technique based on the concept of linear contrast to increase the color image contrast with few calculations. The method is tested with a variety of real low-contrast color images compared to four technical improvement techniques and three prominent IQA (Image quality Analysis) measures, measure the quality of the results collected. Technique received satisfactory results by obtaining results, as it produced natural contrasting pictures without visible features and by rating the highest precision, it outperformed comparative techniques.

The perceptively controlled enhancement of contrast and color in photographs using JND transform and color continuity was proposed by Long Yu et al. [23]. The JND (Just Noticeable Difference) transformation was done to obtain a JND map representing HVS response. Perceptual GEMs have been paired with JND transform to enhance contrast, color reproduction and images. Experimental results showed that the method effectively improves color and dark-tone low-contrast images.

The new technique of fugitive contrast enhancement, which A. A. Salih et al. [24] suggested, is designed to improve individual regions that use the Gaussian membership function (GMF) as their main tool in their fluxation processes. Their method has successfully improved low-contrast and uniform picture lighting with the highest performing and quantitative picture analysis.

Magudeeswaran Veluchamy et al. [25] used improved adaptive gamma correction and histogram equalization based image enhancement on benchmark test images with subjective and objective measures. The method represents natural colors more accurately and outperformed the other existing enhancement methods in terms of entropy, colorfulness and histogram utilization efficiency and having least annoying artefacts.

Syed Zaheeruddin et al. [26] proposed an improvement in contrast in uniform illumination images by the means of homographic decomposition. The proposed method results in better luminosity, contrast and detail, compared to other common methods.

To reduce over enhancement and detail preservation, Subramani, B., et al. [27] proposed quadrant dynamic clipped

HE method with gamma correction. Author partitioned the histogram of the input image into four sections using its mean value and applied histogram clipping and gamma correction to control the color enhancement rate. All clipped sub histogram is equalized independently and then combined together which produces an enhanced image. Experimental results shows that proposed method performed well in terms of preserving entropy, colorfulness, saturation, and obtaining uniform degree enhancement.

In order to improve the lightness and contrast in colored images G., Hazim et al. [28] introduced the Fuzzy Logic Based-on Sigmoid Membership Function (FLBSMF). The lightness component was covered by the algorithm only using the YIQ color space without changing colors. The results of the experiment showed a better color image and good average entropy value, mean square saturation error, and lightness order error.

Magudeeswaran Veluchamy et. al. [29] presented a Fuzzy Dissimilarity Adaptive Histogram Equalization with Gamma Correction (FDAHE-GC) algorithm in which a Fuzzy Dissimilarity Histogram (FDH) is obtained from the neighbourhood characteristics forming intensity mapping function. Different methods for evaluating the performance of the presented method include entropy, Colorfulness, Hue Deviation Index, Saturating, Contrast Enhancement Factor, and Gradient which show improved color enhancement.

Another intensity based enhancement was presented by Krishnamurthy Mayathevar et al. [30], using fuzzy histogram constructed using intensity of neighbourhood regions and further improved by gamma correction in dark regions. Fading effect is mitigated by limiting maximum saturation range as well.

We find that contrast stretching loses some of the detail information of the images during enhancement, Histogram Equalization and its variations gives better results but it cannot preserve the brightness of the original image. Homomorphic filtering techniques have better response but suffers from bleaching effects and multiplicative noise. Retinex is the advance technique for the color image but it still suffers from Gray level violation problem, washed out appearances and unnatural color rendition.

Proposed model is simple, effective and solve the key shortcomings of current models. The main concept behind the proposed method is local color correction in the Hue Saturation Luminance (HSL) domain and fuzzy intensification. The RGB colorspace is first converted to the Hue, Saturation and Luminance colorspace. Local color correction is applied to the luminance component to optimize the color enhancement based on neighboring pixel. Fuzzy intensification operators are used to control the color fidelity of the local color corrected images and thus is able to sort out the problem of overexposed and underexposed regions and provide optimized contrast enhancement in colored images. Several experiments have been performed to analyze the performance of the proposed method and feasibility as compared to existing techniques.

III. Proposed Method

The proposed framework has been shown in the block diagram in Fig. 2.

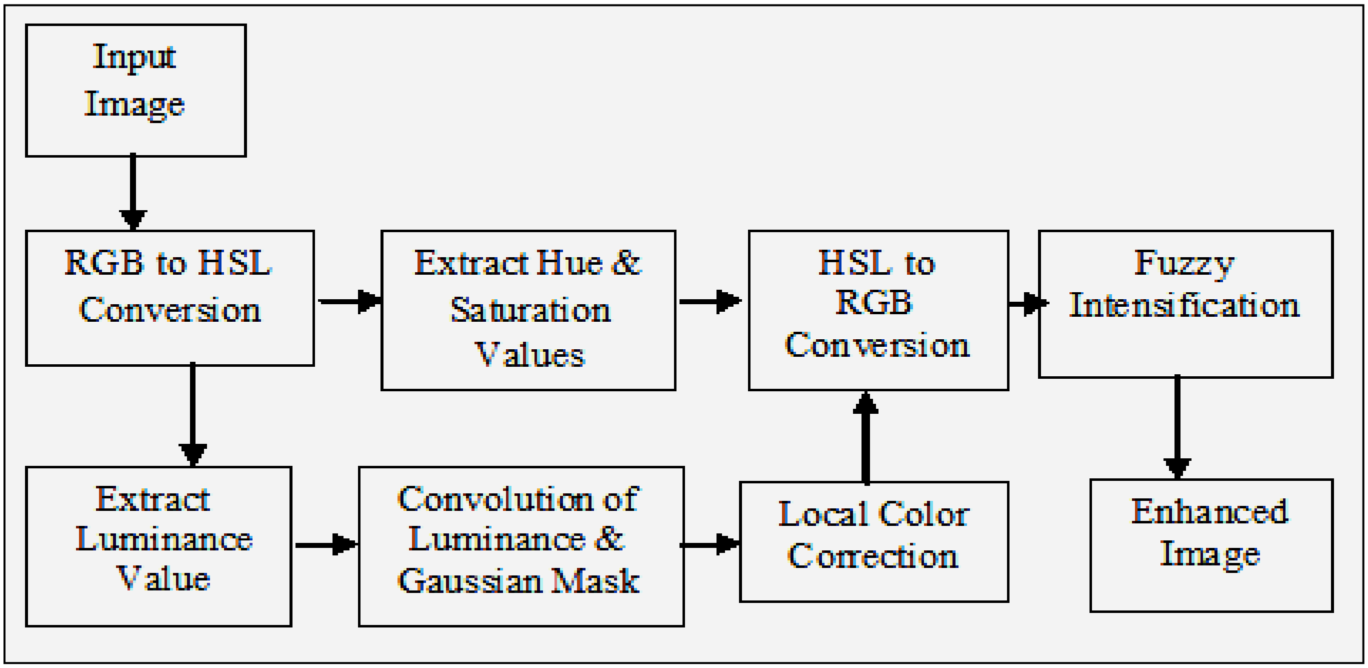


Fig. 2. Block Diagram of Proposed Framework.

Algorithm 1: Color Space Conversion

Step 1: RGB image values is converted to the range 0-1.

Step 2: Obtain the maximum and minimum values from the R,G,B values.

Step 3: Calculate the Luminance (L) value by adding the max and min values and divide by 2.

Step 4: Calculate Saturation(S)

If L< 0.5,

S = $(I_{max}-I_{min})/(I_{max}+I_{min})$

If L>0.5

$$S = (I_{max}-I_{min})/(2.0-I_{max}-I_{min}) \quad (2)$$

where I_{max} & I_{min} are respectively maximum and minimum image pixel value.

Step 5: The Hue(H) formula is depending on what RGB color channel is the max value.

If Red is max, then

H = $(G-B)/(I_{max}-I_{min})$

If Green is max, then

H = $2.0 + (B-R)/(I_{max}-I_{min})$

If Blue is max, then

$$H = 4.0 + (R-G)/(I_{max}-I_{min}) \quad (3)$$

Where *R,G,B* represents the respective color intensity value in the RGB space.

The Hue values are converted to degree values.

The next step is to apply Local Color Correction on the Luminance component. A mask of luminance metric is computed from the original luminance and a Gaussian kernel, which guarantees that image contrast, will not be excessively reduced along the edges is calculated. The resulting mask indicates the corresponding regions of the image which will become dark or light than the original image. The algorithm for color correction is as shown below.

Algorithm 2: Local Color Correction

Step 1: Extract Luminance Component, L $L \leftarrow HSL$

Step 2: Apply Fast Fourier Transform(FFT) on Luminance component, $L_{dft} \leftarrow L$

Step 3: Calculate Gaussian Mask on L, $G_L \leftarrow L$

Gaussian Mask can be defined as:

$$G_L = e^{-2\pi^2\sigma^2(\varepsilon_1^2 + \varepsilon_2^2)} \quad (4)$$

where "ξ" is the frequency of the k^{th} sample given by, $\xi = k/N$, for $k = -N/2,.. N/2-1$, where *N* is the total number of samples and σ is the standard deviation.

Step 4: Apply Discrete Fourier Transform (DFT) on Gaussian Mask (calculated using equation 5) of Luminance, $G_{LDFT} \leftarrow G_L$

$$M'(x,y) = (G_L \text{ x } I)(x,y) \quad (5)$$

Step 5: Perform Convolution of L_{dft}&G_{LDFT}, L_{conv}

Step 6: Apply Inverse Fast Fourier Transform (IFFT) on luminance convolution value L_{conv} and calculate output color corrected luminance, as per equation 6:

$$L_{out}(x,y) = (L_{conv}(x,y))2^{(2M'(x,y)-1)} \quad (6)$$

Equation (6) represents the Output Luminance L_{out} where *x and y* indicate the coordinates.

The color corrected luminance is combined with the original hue and saturation to obtain the color corrected image matrix. The fuzzy intensification operators are then applied on this image to obtain an optimized enhancement.

In order to measure fuzzy intensification operators, two variables are necessary. Firstly, tau(τ) is a parameter that describes the operators threshold limits. The second parameter is a membership function, which is necessary since the default range between zero and one is fixed by pixels of a given channel. Using "τ" allows the operators to filter image pixels.

$$fIr = \frac{[Ir - \min(Ir)]}{\max(Ir) - \min(Ir)} \quad (7)$$

$$fIg = \frac{[Ig - \min(Ig)]}{\max(g) - \min(g)} \quad (8)$$

$$fIb = \frac{[Ib - \min(Ib)]}{\max(Ib) - \min(Ib)} \quad (9)$$

where, *fIr*, *fIg&* *fIb* represent membership function value for each channel respectively.

$kI_R = 2 * (fIr(x,y))^2 \ if \ fIr(x,y) < \tau R$

$$1\text{-}2 * (fIr(x,y))^2 \ otherwise \quad (10)$$

$kI_G = 2 * (fIg(x,y))^2 \ if \ fIg(x,y) < \tau G$

$$1\text{-}2 * (fIg(x,y))^2 \ otherwise \quad (11)$$

$kI_B = 2 * (fIb(x,y))^2 \ if \ fIb(x,y) < \tau B$

$$1\text{-}2 * (fIb(x,y))^2 \ otherwise \quad (12)$$

where, {τR, τG, τB} are scalars. "kIR, kIG, kIB" are the processed channels by intensification operators. To obtain pixels of the output following functions are utilized.

$$u_R = (kIr)^{\tau R + \xi} \quad (13)$$

$$u_G = (kIg)^{\tau G + \xi} \quad (14)$$

$$u_B = (kIb)^{\tau B + \xi} \quad (15)$$

where "ξ" is the intensification tuning parameter which controls the trueness of colors in the image. The complete RGB image is obrtained by combining outputs.

IV. RESULTS AND DISCUSSION

The proposed framework has been tested on CEED dataset for contrast enhancement. The archive consists of 30 images, which consists of collected images and other popular images used by the other experts[31]. All images in the database are truecolor RGB images with pixel size 512x512 height and width respectively. The images contain various scenes of different illuminations, taken both inside and outside as well several benchmark images like peppers, barbara. Thus this dataset covers an exhaustive range of different image types and the proposed frameworks evaluation is carried on this dataset to cover different image types. For comparison, various existing methods have been selected viz. Histogram Equalization [4], BPDHE [5] DHECI [10], Dong [8], AMSR [9], and JED [11] algorithms. The visual appearnce of the enhanced images gives a qualitative approximation of the proposed framework. For an objective analysis of any image processing technique, quantitative parameters have been used in majority of research works. Some of the common performance measures which give a quantitaive approximation of any image processing methods are Mean Square Error(MSE), Peak Signal Noise Ratio(PSNR), Structural Similarity Index Measurement(SSIM) and Natural Image Quality Evaluation(NIQE).

MSE can be defined as difference between the original image and the output image. This difference must be very low for a better performance.

$$\text{MSE} =\sum_{i=1}^{N}((I(x,y)-I'(x,y))^2 \quad (16)$$

Where *I(x,y) and I'(x,y)* represent input and output image pixels respectively.

PSNR can be defined as the ratio between peak signal power and noise power. The PSNR gives a measure of the quality of reconstruction in the final image with respect to the original image.

$$\text{PSNR}=10\log_{10}\left[\frac{255^2}{MSE}\right] \quad (17)$$

Another parameter to measure perceived quality is SSIM, which considers image degradation as perceived change in "structural information" in contrast to MSE and PSNR which estimate the absolute errors only ehich is an important perceptual phenomena which include both contrast and luminance masking terms.

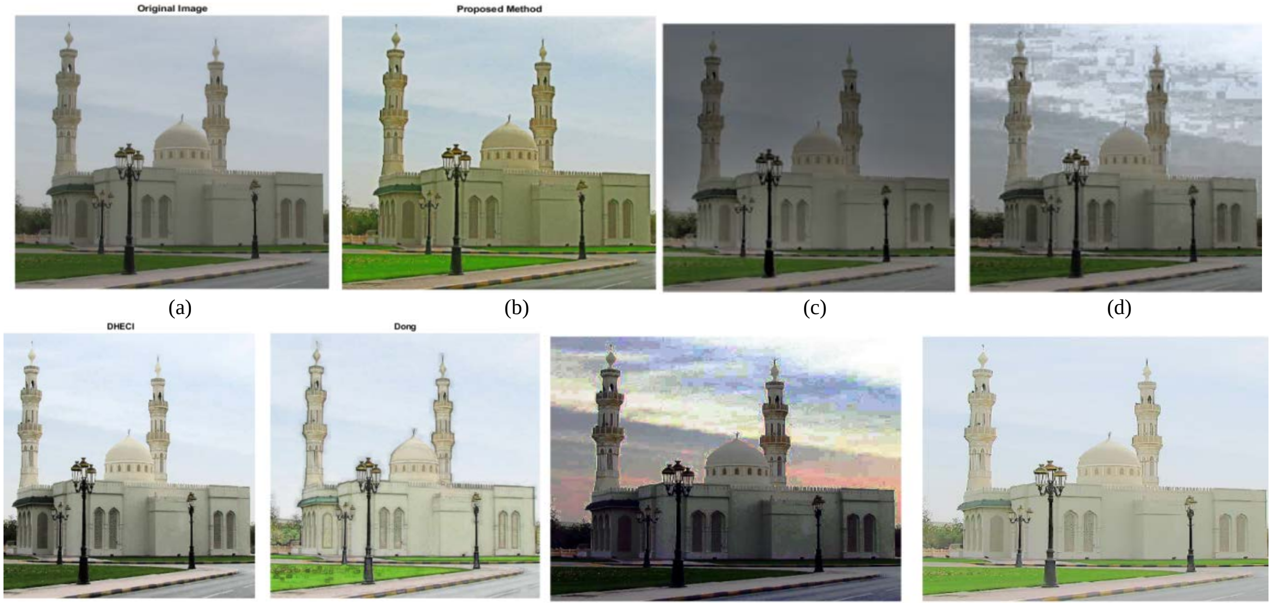
$$\text{SSIM}=\frac{(2\mu_x\mu_y+c1)((2\sigma_{xy}+c2)}{(\mu_x2+\mu_y2+c1)+(\sigma_x2+\sigma_y2+c2)} \quad (18)$$

where "μ_x" is mean of x," μ_y" is average of y, σ_x2 are respective variance &σ_y2 ,$\sigma_{xy}$is covariance. "c1" & "c2", two variables to stabilize the division with weak denominator.

NIQE is an image quality assessment method which uses measurable deviations from statistical regularities observed in natural images.

The experiment were carried out in MATLAB R2018a on original images dataset obtained form CEED 2016 databse of of images. Fig. 3 shows the result images which have obtained by implementing the various existing methods and proposed method for 'img1' image of CEED dataset.

Table 1 shows the experimental results of comparision of different techniques on the basis of MSE, PSNR, SSIM and NIQE parameter and the related graphs in Fig. 4(a), 4(b), 4(c) and 4(c), respectively.



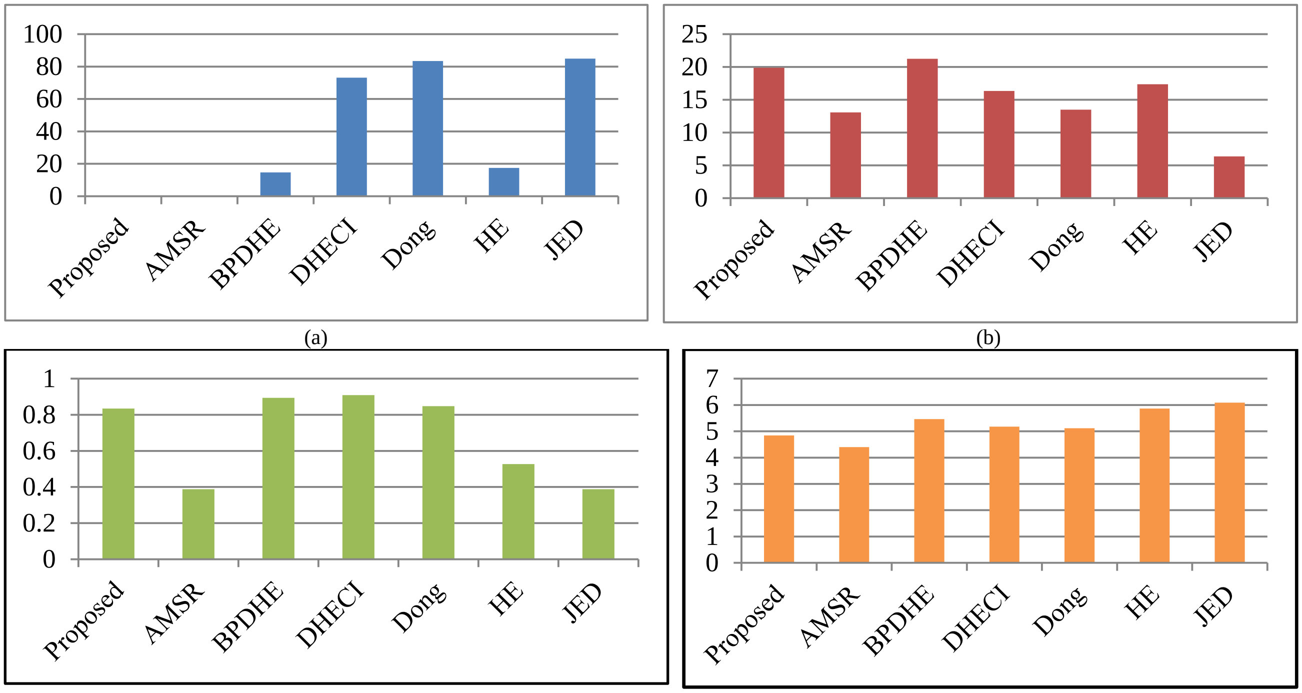
(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. Results for Img1 of CEED Dataset (a) Original Image (b) Proposed Method (c) AMSR (d)BPDHE (e) DHECI (f) Dong (g) HE (h) JED.

TABLE I. COMPARISON OF DIFFERENT TECHNIQUES WITH PROPOSED METHOD

Method	**MSE**	**PSNR**	**SSIM**	**NIQE**
Proposed	0.0034	19.8736	0.8349	4.8460
AMSR	0.0165	13.0662	0.3880	4.3994
BPDHE	14.7533	21.2600	0.8938	5.4617
DHECI	73.1251	16.3550	0.9094	5.1752
Dong	83.4262	13.4936	0.8479	5.1147
HE	17.4953	17.3790	0.5273	5.8644
JED	84.9617	6.3707	0.3880	6.0888



(a)

(b)

(c)

(d)

Fig. 4. Graphical Representation of Comparative Values of (a) MSE (b) PSNR (c) SSIM (d) NIQE.

The proposed method has been compared with the existing techniques in terms of various quality metrics as shown in Table I. The Mean Square Error values for the proposed technique is lowest indicating higher correlation with the original image. PSNR values for the img1 is also better as compared to all other methods except the BPDHE [5], which gives better PSNR value than the proposed method. SSIM values obtained for the proposed method is better than AMSR [9], HE [4] and JED [11], while lower than BPDHE [5], DHECI [10], Dong [8] and JED [11]. NIQE values obtained are better than all other methods except the AMSR [9]. Though, if we analyze the visual perceptibility as a qualitative measure, we can see that the methods having better quantitative values i.e. PSNR, SSIM or NIQE doesn't have a good visual perceptibility as compared to the proposed method. Thus, it can be said that there is a tradeoff between the performance measures and qualitative visual perceptibility as well, as is evident from this research. Thus, based on this, it can be said that the proposed technique achieves better performance in terms of both qualitative measures and quantitative metrics. Fig. 5 to Fig. 7 shows some more results for different images.

Table II shows the computational cost by proposed and existing methods, in terms of execution time. The execution time of the proposed method is slightly more as compared to AMSE, Dong and BPDHE but it is many times lower than DHECI and ZED methods. Although HE takes much less time to execute but its many drawbacks force to ignore the method. The proposed algorithm thus gives an acceptable performance in contrast optimization and performs well in terms of execution time as well.

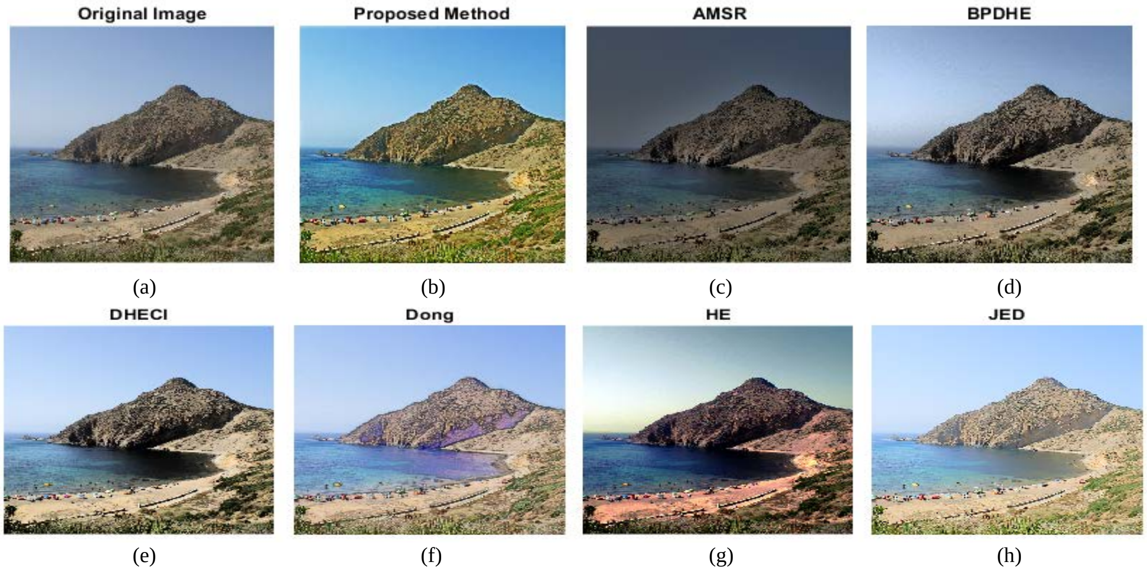


Fig. 5. Results for Img2 of CEED Dataset (a) Original Image (b) Proposed Method (c) AMSR (d)BPDHE (e) DHECI (f) Dong (g) HE (h) JED.

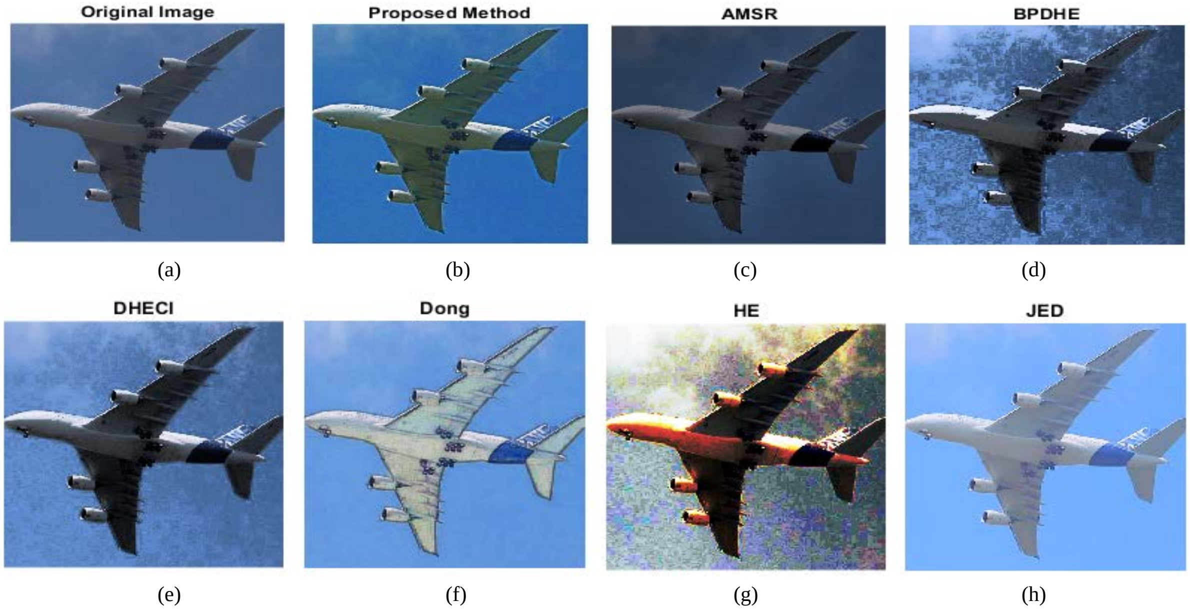
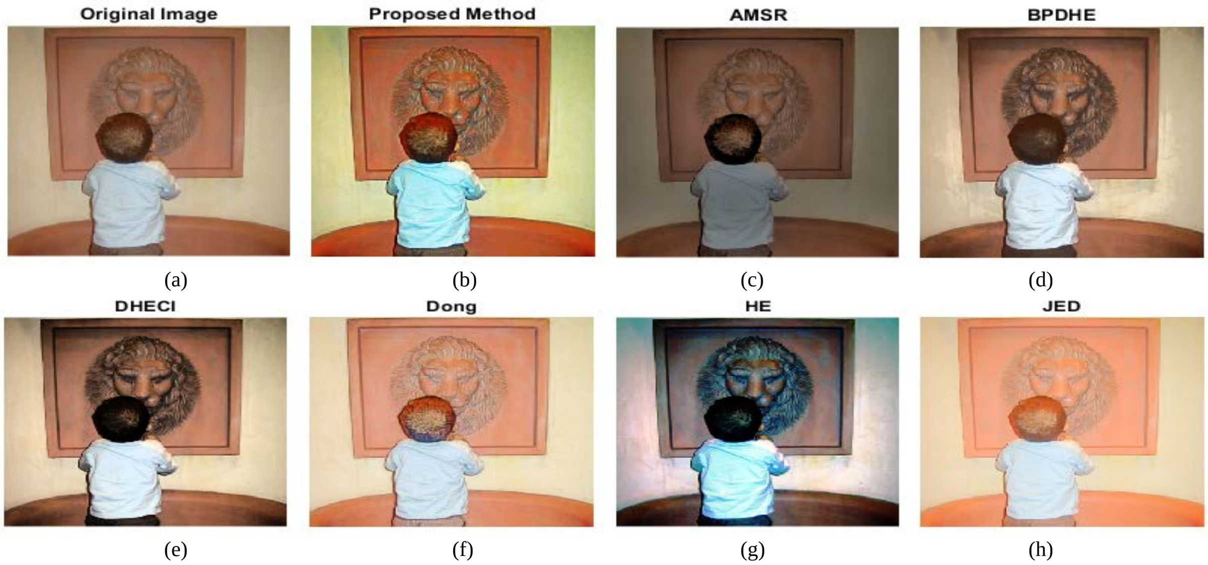


Fig. 6. Results for Img3 of CEED Dataset (a) Original Image (b) Proposed Method (c) AMSR (d) BPDHE (e) DHECI (f) Dong (g) HE (h) JED.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7. Results for Img10 of CEED Dataset (a) Original Image (b) Proposed Method (c) AMSR (d)BPDHE (e) DHECI (f) Dong (g) HE (h) JED.

TABLE II. COMPARISON OF EXECUTION TIME

Image	**Proposed**	**AMSR**	**BPDHE**	**DHECI**	**Dong**	**HE**	**JED**
	1.2950	0.5717	0.4933	15.2758	0.5279	0.01295	10.2032
	1.3472	0.6810	0.5736	15.1493	0.5869	0.01961	10.0039
	1.1541	0.6063	0.4195	15.1940	0.4540	0.0317	9.9660
	3.1674	37.8317	1.1424	16.1688	0.8648	0.5012	24.2718
	1.4195	0.7262	0.4696	14.1218	0.5392	0.0347	9.9046
	1.2339	0.6893	0.5416	14.5583	0.4748	0.0113	11.4457

V. CONCLUSION

In this research, a technique is proposed which takes into account both the local minima and global maxima problems. To overcome the problems and to present an optimal enancment approach, the RGB colorspace is first converted to the Hue, Saturation and Luminance colorspace. Local color correction is applied to the HSL colorspace to optimize the contrast of image based on neighboring pixel. Fuzzy intensification methods is used to control the color fidelity of the local color corrected pixels and thus the overexposed and underexposed regions have been optimally illuminated and obtained images are having good contrast. The methods were implemented and tested on CEED database images, along with the existing techniques. Proposed method provides satisfactory results as it produced natural contrast images with no artifacts and outperformed the other existing contrast enhancement methods in terms of MSE, PSNR, SSIM and NIQE image evaluation parameters.

ACKNOWLEDGMENT

I express my sincere gratitude towards Dr. Rakesh Kumar Yadav, Assistant Prof., Department of Computer Science and Engineering IFTM University Moradabad and Professor Ramapati Mishra, Department of Electronics and Communication Engineering, Dr. RML Avadh University Ayodhya for his valuable suggestions and guidance. I also pay thanks to Head of Department of Computer Science and Engineering IFTM University, Moradabad for providing us necessary facilities for the implementation of this research work.

REFERENCES

[1] Saleem, A., Beghdadi, A. & Boashash, B. Image fusion-based contrast enhancement. J Image Video Proc 2012, 10 (2012). https://doi.org/10.1186/1687-5281-2012-10.

[2] Soong-Der Chen, “A new image quality measure for assessment of histogram equalization-based contrast enhancement techniques”, Digital Signal Process, 22 (2012) 640–647.

[3] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu, “Efficient Contrast Enhancement Using Adaptive Gamma Correction With Weighting Distribution”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 3, MARCH 2013.

[4] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li, “Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images”, IEEE Transaction on Image Processing, VOL. 22, NO. 9, SEPTEMBER 2013.

[5] H. Ibrahim and N. S. Pik Kong, "Brightness Preserving Dynamic Histogram Equalization for Image Contrast Enhancement," in IEEE Transactions on Consumer Electronics, vol. 53, no. 4, pp. 1752-1758, Nov. 2007, doi: 10.1109/TCE.2007.4429280.

[6] Rahman, S., Rahman, M.M., Abdullah-Al-Wadud, M. et al. An adaptive gamma correction for image enhancement. J Image Video Proc. 2016, 35 (2016).

[7] Naina Dhingra, Amita Nandal, Meenu Manchanda, Deepak Gambhir, Fusion of Fuzzy Enhanced Overexposed and Underexposed Images, Proceedia Computer Science, Volume 54,2015,Pages 738-745.

[8] Dong, X.; Pang, Y.A.; Wen, J.G. Fast efficient algorithm for enhancement of low lighting video. In Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 11–15 July 2010.

[9] C. Lee, J. Shih, C. Lien and C. Han, "Adaptive Multiscale Retinex for Image Contrast Enhancement," 2013 International Conference on Signal-Image Technology & Internet-Based Systems, Kyoto, Japan, 2013, pp. 43-50, doi: 10.1109/SITIS.2013.19.

[10] Nakai, K., Hoshi, Y., Taguchi, A.: Color image contrast enhancement method based on differential intensity/saturation gray-levels histograms. In: Intelligent Signal Processing and Communications Systems (ISPACS)(2013).

[11] X. Ren, M. Li, W. Cheng and J. Liu, "Joint Enhancement and Denoising Method via Sequential Decomposition," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351427.

[12] Chun-Ming Tsai and Zong-Mu Yeh, “Contrast Compensation by Fuzzy Classification and Image Illumination Analysis for Back-lit and Front-lit Color Face Images”, IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, August 2010.

[13] Iyad F. Jafar, Khalid A. Darabkh, Ghazi M. Al-Sukkar, A Rule-Based Fuzzy Inference System for Adaptive Image Contrast Enhancement, The Computer Journal, Volume 55, Issue 9, September 2012, Pages 1041–1057, https://doi.org/10.1093/comjnl/bxr120.

[14] Saleem, A., Beghdadi, A. & Boashash, B. Image fusion-based contrast enhancement. J Image Video Proc 2012, 10 (2012). https://doi.org/10.1186/1687-5281-2012-10.

[15] M. Liu and P. Ndjiki-Nya, "A new perceptual-based no-reference contrast metric for natural images based on human attention and image dynamic," 2012 Fourth International Workshop on Quality of Multimedia Experience, Yarra Valley, VIC, 2012, pp. 254-259, doi: 10.1109/QoMEX.2012.6263887.

[16] Huang SC, Cheng FC, Chiu YS. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE Trans Image Process. 2013 Mar;22(3):1032-41. doi: 10.1109/TIP.2012. 2226047.

[17] V. Magudeeswaran and C. G. Ravichandran, “Fuzzy Logic-Based Histogram Equalization for Image Contrast Enhancement”, Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 891864, 10 pages.

[18] Jeyong Shin, Student Member, IEEE, and Rae-Hong Park, Senior Member, IEEE, “Histogram-Based Locality-Preserving Contrast Enhancement”, IEEE SIGNAL PROCESSING LETTERS, VOL. 22, NO. 9, SEPTEMBER 2015.

[19] Chin Yeow Wong, Guannan Jiang, Md Arifur Rahman, Shilong Liu, Stephen Ching-Feng Lin, Ngaiming Kwok, Haiyan Shi, Ying-Hao Yu, Tonghai Wu, Histogram equalization and optimal profile compression based approach for color image enhancement, Journal of Visual Communication and Image Representation, Volume 38,2016,Pages 802-813,ISSN 1047-3203.

[20] Zohair Al-Ameen, “Visibility Enhancement for Images Captured in Dusty Weather via Tuned Tri-threshold Fuzzy Intensification Operators”, I.J. Intelligent Systems and Applications, 2016, 8, 10-17.

[21] M. Shakeri, M.H. Dezfoulian, H. Khotanlou, A.H. Barati, Y. Masoumi, Image contrast enhancement using fuzzy clustering with adaptive cluster parameter and sub histogram equalization, Digital Signal Processing, Volume 62,2017,Pages 224-237,ISSN 1051-2004.

[22] Zohair Al-Ameen, “CONTRAST ENHANCEMENT FOR COLOR IMAGES USING AN ADJUSTABLE CONTRAST STRETCHING TECHNIQUE”, International Journal of Computing, 17(2) 2018, 74-80.

[23] L. Yu, H. Su and C. Jung, "Perceptually Optimized Enhancement of Contrast and Color in Images," in IEEE Access, vol. 6, pp. 36132-36142, 2018, doi: 10.1109/ACCESS.2018.2848671.

[24] A. A. Mohammed Salih, K. Hasikin and N. A. M. Isa, "Adaptive Fuzzy Exposure Local Contrast Enhancement," in IEEE Access, vol. 6, pp. 58794-58806, 2018, doi: 10.1109/ACCESS.2018.2872116.

[25] Magudeeswaran Veluchamy, Bharath Subramani , “Image contrast and color enhancement using adaptive gamma correction and histogram equalization”, Optik, Volume 183,2019,Pages 329-337.

[26] Syed Zaheeruddin, K. Suganthi, “Image Contrast Enhancement by Homomorphic Filtering based Parametric Fuzzy Transform”, Proceedia Computer Science, Volume 165, 2019, Pages 166-172.

[27] Subramani, B., & Veluchamy, M. (2020). “Quadrant dynamic clipped histogram equalization with gamma correction for color image enhancement”, Color Research and Application, 45(4), 644–655. https://doi.org/10.1002/col.22502.

[28] G., Hazim & Daway, Esraa & Kareem, Hana. (2020). Colour Image Enhancement by Fuzzy Logic Based on Sigmoid Membership Function. International Journal of Intelligent Engineering and Systems. 13. 238-246. 10.22266/ijies2020.1031.21.

[29] Magudeeswaran Veluchamy, Bharath Subramani, “Fuzzy dissimilarity color histogram equalization for contrast enhancement and color correction”, Applied Soft Computing, Volume 89, 2020, 106077, https://doi.org/10.1016/j.asoc.2020.106077.

[30] Krishnamurthy Mayathevar, Magudeeswaran Veluchamy, Bharath Subramani, “Fuzzy color histogram equalization with weighted distribution for image enhancement”, Optik, Volume 216, 2020, 164927, https://doi.org/10.1016/j.ijleo.2020.164927.

[31] Qureshi, Muhammad Ali; Sdiri, Bilel; Deriche, Mohamed; Alaya-Cheikh, Faouzi; Beghdadi, Azeddine (2017), “Contrast Enhancement Evaluation Database (CEED2016)”, Mendeley Data.

Exploring Factors Associated with the Social Discrimination Experience of Children from Multicultural Families in South Korea by using Stacking with Non-linear Algorithm

Haewon Byeon
Department of Medical Big Data, College of AI Convergence
Inje University, Gimhae 50834, Republic of Korea

***Abstract*—The number of children from multicultural families is increasing rapidly along with quickly increasing multicultural families. However, there are not enough surveys and basic researches for understanding the characteristics of multicultural children and issues such as social discrimination. This study discovered the machine learning model with the best performance for predicting the social discrimination experience of children from multicultural families by comparing the prediction performance (accuracy) of individual prediction models and stacking ensemble models. This study analyzed 19,431 adolescents (between 19 and 24 years old: 9,835 males and 9,596 females) among the children of marriage immigrants. This study used random forest (RF), rotation forest, artificial neural network (ANN), and support vector machine (SVM) for the base model. Logistic regression algorithm was applied for the meta model. Each machine learning model was built through 5-fold cross-validation. Root-mean-square-error (RMSE), index of agreement (IA), and variance of errors (Ev) were used to evaluate the prediction performance of the developed models. The results of this study indicated that the prediction performance of the rotation forest-logistic regression model had the best performance. The future studies need to explore stacking ensemble models with the best performance through combining a base model and a meta model by using various machine learning algorithms such as clustering and boosting.**

***Keywords*—Stacking ensemble; meta model; root-mean-square-error; index of agreement; rotation forest**

I. Introduction

The number of foreigners residing in South Korea exceeded 2 million as of 2019. This accounts for 3.69% of the South Korean population, which is not a high percentage. However, it is recognized as a noticeable phenomenon in Korean society because the number of immigrants has increased rapidly over the past decade and immigrants are easily distinguishable due to differences in appearance and language [1]. In particular, as this issue has become linked to the marriage to men living in rural areas or men with low-income in urban areas since 2002, the number of multicultural families has reached 900,000 as of 2016 [2]. The number of immigrants will increase more as the population of South Korea will decrease due to the aging and low birth rate [3]. It has drawn more attention because the population composition will be diversified further due to this [3].

The multicultural family means a family made by uniting people with different nationalities or races through international marriage and other methods. South Korea prepared the “Measures to Support the Social Integration for Female Marriage Immigrant Families, Multi-racial People, and Immigrants” in 2006 to help multicultural families settle in South Korea stably. As the Multicultural Families Support Act was enacted in 2008, she strengthened the support for multicultural families at the policy level. As a result, social security and legal status were guaranteed for marriage immigrants.

As the number of foreigners residing in South Korea rapidly increases, the number of children from multicultural families (e.g., international marriage families and foreign workers’ families) is also increasing. Furthermore, as they attend schools, the possibility of conflict due to cultural differences has increased according to the increased personal and cultural contacts.

Nevertheless, in South Korea, social policies for multicultural families have mainly focused on employment or welfare for marriage immigrants and foreign workers [4,5]. Moreover, previous studies on multicultural families [4,6] have been conducted to examine only limited individual aspects such as socioeconomic characteristics, welfare level, human rights discrimination, employment status, and policy analysis. However, there are still not enough studies on the overall social discrimination experiences of children from multicultural families. Children from multicultural families (international marriage families) can be divided into children born in South Korea and those who have entered South Korea after being born in other countries. Since it has been reported that children could not adapt to South Korea well due to the unique characteristics of multicultural families and various changes that they experience during adolescence [7], it is necessary to expand the social foundation that can help them adapt to South Korea well for social integration.

In summary, the number of children from multicultural families is increasing rapidly along with quickly increasing multicultural families. However, there are not enough surveys and basic researches for understanding the characteristics of multicultural children and issues such as social discrimination. Therefore, it is needed to identify the characteristics of multicultural children and seek new policies that reflect them to prepare policies that encompass various problems including the social adaptation of multicultural children in preparation for a rapidly changing multicultural society.

Previous studies [8,9,10,11,12] on the adolescents from multicultural families in South Korea reported the difficulties in peer relations, social support, family support, and language as factors related to discrimination experiences. Most of them used regression analysis for a prediction algorithm. Regression analysis is efficient for detecting individual risk factors, but it is limited in identifying multiple risk factors [13,14]. As a way to overcome this limitation, recent social science studies [15,16] have used predictive modeling based on big data-based machine learning. However, since these prediction studies are based on individual prediction algorithms, the bias existing in each algorithm may be reflected in the prediction results.

This study identified the predictors of social discrimination experiences of children from multicultural families in South Korea by using individual prediction models based on machine learning and reduced the potential bias risk of the models by combining them into a stacking ensemble learning model. Moreover, this study discovered the machine learning model with the best performance for predicting the social discrimination experience of children from multicultural families by comparing the prediction performance (accuracy) of individual prediction models and stacking ensemble models.

II. Methods and Materials

A. Data Source

The data source of this study was the “Study on the National Survey of Multicultural Families [17]” in 2012, which was jointly surveyed by the Ministry of Health, Welfare and Family Affairs, the Ministry of Justice, and the Ministry of Gender Equality and evaluated multicultural families residing in South Korea. The Study on the National Survey of Multicultural Families was conducted to develop policies customized for multicultural families by identifying their living conditions and welfare needs. The survey items consisted of the general characteristics, employment, economic level, marriage, health, and health care of multicultural families. The survey targets for the national survey of multicultural families were 154,333 families, all marriage immigrants. In addition to marriage immigrants, this survey also evaluated the spouses and children of marriage immigrants separately. The target subjects were sampled based on the status of alien residents living in 16 cities and provinces and the basic status of multicultural families collected by the Ministry of Public Administration and Security. Since this survey collected data from all target samples, a sample design was not needed and the survey was conducted from July 20 to October 31, 2012. The multicultural families used for this study were (1) families composed of marriage immigrants and South Korean who obtained South Korean nationality by birth, acknowledgment, or naturalization and (2) families composed of foreigners who obtained South Korean nationality by acknowledgment or naturalization and South Koreans who obtained South Korean nationality by birth, acknowledgment, or naturalization in accordance with the Multicultural Families Support Act. This study analyzed 19,431 adolescents (between 19 and 24 years old: 9,835 males and 9,596 females) among the children of marriage immigrants.

B. Measurements and Definitions of Variables

The target variable (label) was defined as social discrimination experience (yes or no). Features were gender, age, residence (countryside or city), highest level of education (elementary school graduation and below, middle school graduation, high school graduation, or college graduation or higher), Korean reading level (good, average, or poor), Korean speaking level (good, average, or poor), Korean writing level (good, average, or poor), Korean listening level (good, average, or poor), learning support experience (yes or no), economic activity (yes or no), the experience of using a support center for multi-cultural families (yes or no), learning Korean (yes or no), Korean society adaptation training (yes or no), career counseling (yes or no), and social welfare center use (yes or no).

C. Single Machine Learning Algorithm (base model): Support Vector Machine (SVM)

SVM is a machine learning algorithm that finds an optimal decision boundary, which is a linear separation that optimally separates a hyperplane by transforming training data into a high dimension through nonlinear mapping [18]. For example, when A=[a,d] and B=[b,c] are non-linearly separable in 2D, it becomes linearly separable when they are mapped in 3D. Therefore, if an appropriate nonlinear mapping is applied to a sufficiently large dimension, data with two classes can always be separated in a hyperplane. The concept of SVM is presented in Fig. 1.

D. Random Forest

Random forest is an algorithm that randomly learns a number of decision trees. It uses a number of bootstrap samples. After generating a decision tree for each sample, the output value is predicted using the decision tree most frequently used among the generated decision tree when new data is input [20]. The concept of random forest is presented in Fig. 2.

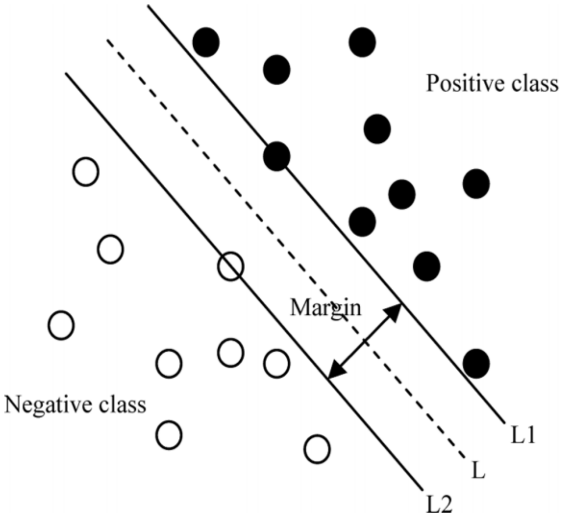


Fig. 1. Concept of SVM [19].

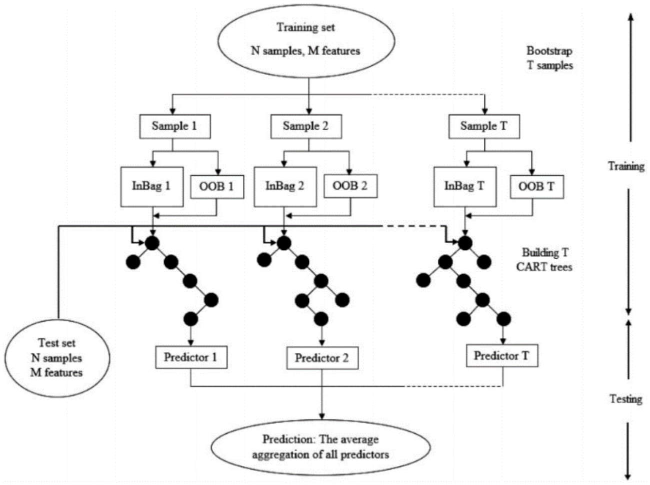


Fig. 2. Structure of Random Forest Algorithm [21].

E. Rotation Forest

Rotation forest is one of the random forest algorithms that performs learning while rotating the data axis by applying principal component analysis (PCA) to the training data. Rotation forest generates classifier ensembles based on feature extraction after excluding random features from the previous feature set used for learning. Principal component analysis (PCA) is performed on randomly divided subsets and training is conducted by rotating the data dimension [22]. Through this process, robust characteristics can be obtained for the input data showing complex distribution [23]. The performance procedure of the rotation forest is presented in Fig. 3.

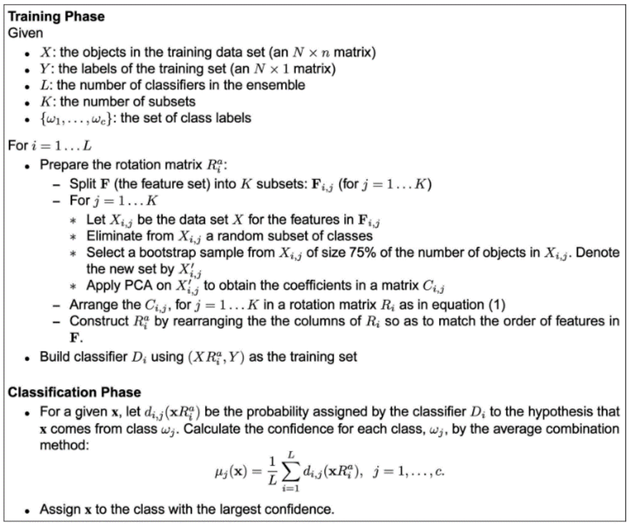
F. ANN (Artificial Neural Network)

ANN is an algorithm created based on the neural network structure of the human brain. It is composed of an input layer that inputs the target data, a hidden layer (or hidden layers) that is an intermediate step, and an output layer that shows the result. Every layer consists of a number of nodes, and only information that exceeds the threshold is passed to the next layer through the activation function. It is possible to predict the result in the output layer after deriving only the necessary information through this. The concept of ANN is presented in Fig. 4.

G. Stacking Ensemble (Meta Model)

This study predicted social discrimination experiences by using stacking ensemble techniques. Stacking ensemble techniques are superior in generalization and robustness to single machine learning models and have been used for classification and prediction in various fields [26,27,28,29]. It is a method of creating a new model by combining different models as if stacking them [30]. It improves the performance of the final model by taking advantage of each model and supplementing its weaknesses while going through the two stages (base and meta) [30].

This study used random forest (RF), rotation forest, ANN, and SVM for the base model. Logistic regression algorithm was applied for the meta model. The regression algorithm is the simplest method to increase the reliability of the base model while maximizing the generality and stability of the model. Feng et al., (2020) [31] reported that it would overfit the training data less probably. Due to this reason, the regression algorithm has been used as a meta model of the stacking ensemble algorithm in many recent publications [31,32], and this study also used it as a meta model for the same reason. The structure of the finally constructed stacking ensemble model is presented in Fig. 5.



```
Training Phase
Given
  • $X$: the objects in the training data set (an $N \times n$ matrix)
  • $Y$: the labels of the training set (an $N \times 1$ matrix)
  • $L$: the number of classifiers in the ensemble
  • $K$: the number of subsets
  • $\{\omega_1, \ldots, \omega_c\}$: the set of class labels

For $i = 1 \ldots L$
  • Prepare the rotation matrix $R_i^a$:
      – Split $\mathbf{F}$ (the feature set) into $K$ subsets: $\mathbf{F}_{i,j}$ (for $j = 1 \ldots K$)
      – For $j = 1 \ldots K$
          * Let $X_{i,j}$ be the data set $X$ for the features in $\mathbf{F}_{i,j}$
          * Eliminate from $X_{i,j}$ a random subset of classes
          * Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote
            the new set by $X'_{i,j}$
          * Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
      – Arrange the $C_{i,j}$, for $j = 1 \ldots K$ in a rotation matrix $R_i$ as in equation (1)
      – Construct $R_i^a$ by rearranging the the columns of $R_i$ so as to match the order of features in
        $\mathbf{F}$.
  • Build classifier $D_i$ using $(XR_i^a, Y)$ as the training set

Classification Phase
  • For a given $\mathbf{x}$, let $d_{i,j}(\mathbf{x}R_i^a)$ be the probability assigned by the classifier $D_i$ to the hypothesis that
    $\mathbf{x}$ comes from class $\omega_j$. Calculate the confidence for each class, $\omega_j$, by the average combination
    method:
        $\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}R_i^a), \;\; j = 1, \ldots, c.$
  • Assign $\mathbf{x}$ to the class with the largest confidence.
```

Fig. 3. Procedure of Rotation Forest [24].

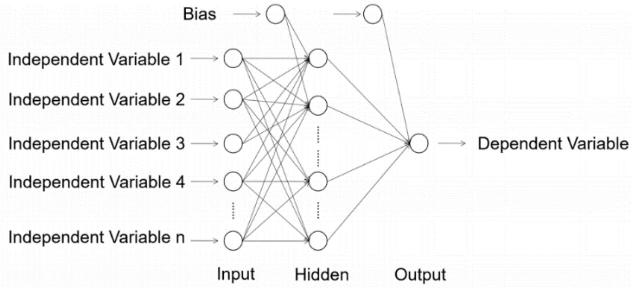


Fig. 4. Algorithmic Structure of a Typical ANN [25].

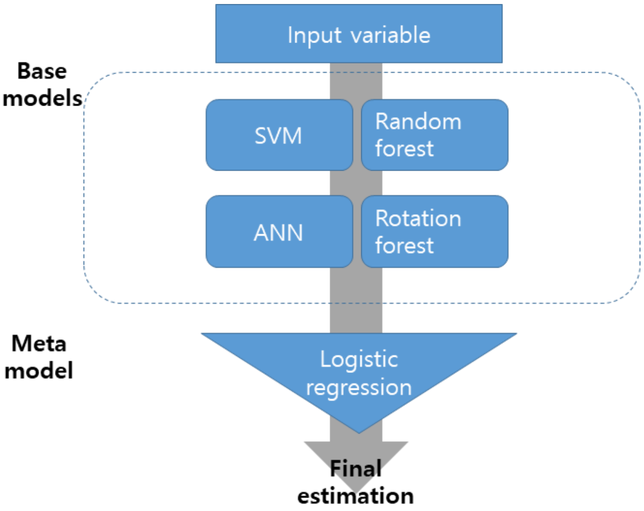


Fig. 5. The Structure of the Stacking Ensemble.

H. Validation of the Models

Each machine learning model was built through 5-fold cross-validation. This method validates the validity of learning by randomly dividing the entire sample into equal-sized five groups, validating it by using one of the groups as a validation dataset and the other groups as training datasets, and repeating this procedure five times. Root-mean-square-error (RMSE), index of agreement (IA), and variance of errors (Ev) were used to evaluate the prediction performance of the developed models. A lower RMSE indicates the higher accuracy of a prediction model. When IA is closer to 1 and Ev is lower, a model is more stable.

III. Results

Table I shows the general characteristics of adolescents from multicultural families in South Korea according to the presence of social discrimination experience. Among the all subjects (19,431 adolescents), 15.6% (3,035 adolescents) experienced social discrimination. The result of chi-square test revealed that residence, gender, highest level of education, the experience of using a support center for multi-cultural families, Korean speaking level, Korean listening level, Korean reading level, Korean writing level, career counseling, learning Korean, and Korean society adaptation training were significantly (p<0.05) different between adolescents from multicultural families with social discrimination experience and those without social discrimination experience.

TABLE I. General Characteristics of Adolescents from Multicultural Families in South Korea, n (%)

Variables	**Social discrimination experience**		**p**
	Yes (n=3,035)	**No (n=16,396)**	
Residence			<0.001
City	2,659 (16.2)	13,802 (83.8)	
Countryside	375 (12.6)	2,594 (87.4)	
Highest level of education			
Elementary school graduation and below	26 (12.9)	176 (87.1)	
Middle school graduation	394 (15.9)	2,084 (84.1)	
High school graduation	2,164 (15.3)	12,024 (84.7)	
College graduation or higher	451 (17.6)	2,112 (82.4)	
Gender			<0.001
Male	1,898 (19.3)	7,938 (80.7)	
Female	1,137 (11.8)	8,459 (88.2)	
Korean speaking level			<0.001
Good	2,176 (14.4)	12,939 (85.6)	
Average	635 (21.4)	2,330 (78.6)	
Poor	224 (16.6)	1,128 (83.4)	
Korean reading level			<0.001
Good	2,117 (14.0)	13,016 (86.0)	
Average	496 (19.6)	2,033 (80.4)	
Poor	422 (23.8)	1,348 (76.2)	
Korean writing level			<0.001
Good	2,033 (13.8)	12,731 (86.2)	
Average	506 (18.7)	2,196 (81.3)	
Poor	496 (25.2)	1,470 (74.8)	
Korean listening level			<0.001
Good	2,220 (14.4)	13,165 (85.6)	
Average	677 (24.1)	2,127 (75.9)	
Poor	137 (11.0)	1,105 (89.0)	
Economic activity			0.659
No	1,790 (15.7)	9,598 (84.3)	
Yes	1,245 (15.5)	6,798 (84.5)	
Korean society adaptation training			
No	2,888 (15.2)	16,104 (84.8)	
Yes	146 (33.3)	292 (66.7)	
Experience of career counseling			<0.001
No	2,611 (14.3)	15,602 (85.7)	
Yes	424 (34.8)	794 (65.2)	
Learning support experience			<0.001
No	2,526 (14.1)	15,336 (85.9)	
Yes	509 (32.4)	1,060 (67.6)	
Experience of using a support center for multi-cultural families			<0.001
Do not even know that such center exists	1,692 (18.8)	7,332 (81.3)	
Know the center but never used it before	1,125 (12.1)	8,198 (87.9)	
Not only know the center but also have used it before	218 (20.1)	867 (79.9)	

The prediction performance (i.e., RMSE, IA, and Ev) of the eight machine learning models for predicting social discrimination experience is presented in Fig. 6, 7, and 8, respectively. The analysis results of this study indicated that the prediction performance of the rotation Forst-logit regression model (RMSE = 0.15, IA = 0.72, and Ev =0.41) had the best performance.

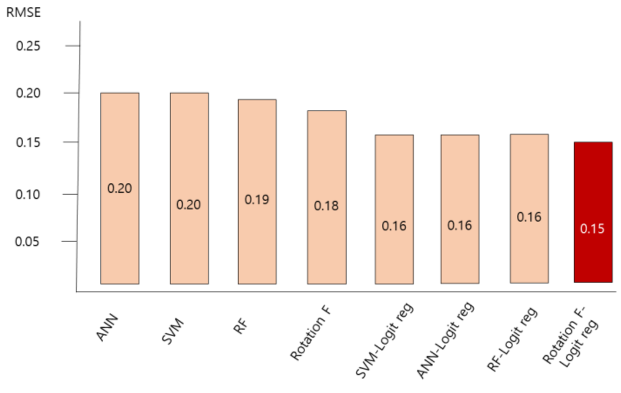


Fig. 6. RMSE Comparison of Machine Learning Models for Predicting Social Discrimination.

ANN=Artificial neural network; SVM=Support Vector Machine, RF=Random forest; Rotation F=Rotation forest; SVM-Logit reg=SVM-Logistic regression; ANN-Logit reg=Artificial neural network-Logistic regression; RF-Logit reg=Random forest-Logistic regression; Rotation F-Logit reg=Rotation forest-Logistic regression.

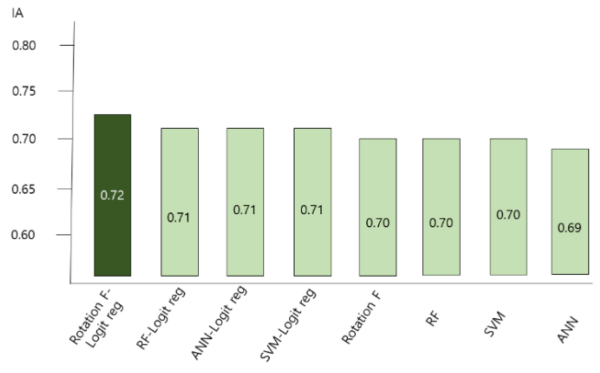


Fig. 7. IA Comparison of Machine Learning Models for Predicting Social Discrimination.

ANN=Artificial neural network; SVM=Support Vector Machine, RF=Random forest; Rotation F=Rotation forest; SVM-Logit reg=SVM-Logistic regression; ANN-Logit reg=Artificial neural network-Logistic regression; RF-Logit reg=Random forest-Logistic regression; Rotation F-Logit reg=Rotation forest-Logistic regression.

The normalized importance of each variable of the rotation forest-logit regression model is presented in Fig. 9. The model confirmed that Korean society adaptation training, learning Korean, gender, the experience of using a multicultural family support center, and career counseling were major variables with high weight in the social discrimination experience of children from multicultural families in South Korea. Especially, Korean society adaptation training was the most important factor in the final model.

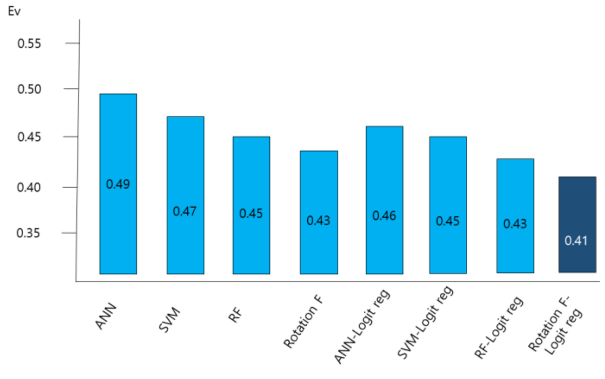


Fig. 8. Ev Comparison of Machine Learning Models for Predicting Social Discrimination.

ANN=Artificial neural network; SVM=Support Vector Machine, RF=Random forest; Rotation F=Rotation forest; SVM-Logit reg=SVM-Logistic regression; ANN-Logit reg=Artificial neural network-Logistic regression; RF-Logit reg=Random forest-Logistic regression; Rotation F-Logit reg=Rotation forest-Logistic regression.

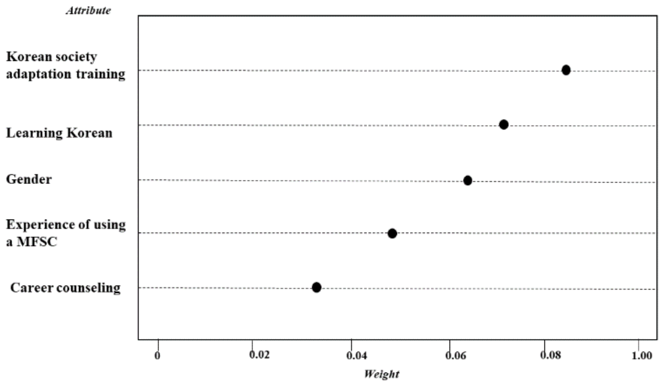


Fig. 9. The Importance of Variables in the Prediction Model for Discrimination Experience of Children from Multicultural Families in South Korea (only the top 5 Variables are Presented).

MFSC=multicultural family support center

IV. Conclusion

This study compared the accuracy of models for predicting the social discrimination experience of children from multicultural families in South Korea by using eight machine learning algorithms, and confirmed that the rotation forest-logit regression model based on the stacking ensemble algorithm had the best prediction performance. In particular, the prediction model based on the stacking ensemble had improved accuracy (RMSE = 0.04-0.05) than other models and more stable (IA= 0.02-0.03) than other models. The results of this study support the possibility that the meta-model's prediction performance can be superior to the single prediction model for not only unstructured data such as videos and images but also structured data such as social science data. However, Lee & Kim (2020) [33] also reported that stacking ensemble algorithms had a longer execution time (runtime) than single machine learning algorithms, a limitation. Therefore, future studies using stacking ensembles need to evaluate the prediction performance comprehensively by comparing execution time (runtime) as well as accuracy. It is also needed to explore stacking ensemble models with the best performance through combining a base model and a meta model by using various machine learning algorithms such as clustering and boosting.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091, NRF-2019S1A5A8034211, NRF-2021S1A5A8062526).

References

[1] B. G. Koo, Multiculturalism and transnational migrants: a case study of Wongok-dong, an immigrant-dominated area of Ansan city in South Korea. Cross - Cultural Studies, vol. 19, no. 2, pp. 5-51, 2013.

[2] B. J. Park, The role of local government for social Integration with multicultural family. Journal of North-east Asian Cultures, vol. 51, p. 285-307, 2017.

[3] E. H. Chae, Analysis of trends of ‘an investigation on multicultural families in Korea’ at the Korean Statistical Information Service(KOSIS), vol. 8, no. 11, pp. 11-20, 2018.

[4] J. H. Shin, A study on comparative home environmental factor effect of delinquency in multi-cultural youth. Journal of Korean Public Police and Security Studies, vol. 11, no.2, pp.1-20, 2014.

[5] J. Kim, Problem about policies for multi-cultural society and social integration. Low and Social Study, vol. 11, no. 2, pp. 349-368, 2011.

[6] K. S. Ahn, A plan of multi-cultural adolescent`s healthy upbringing. Korean Journal of Youth Studies, vol. 16, no. 7, pp. 99-126, 2009.

[7] M. Tienda, and R. Haskins, Immigrant children: introducing the issue. The Future of Children, vol. 21, no. 1, pp. 3-18, 2011.

[8] H. M. Kim, W. J. Seo, and S. H. Choi, Experiences of discrimination and psychological distress of children from multicultural families: examining the mediating effect of social support. Korean Journal of Social Welfare Studies, vol. 42, no. 1, pp. 117-149, 2011.

[9] S. Cha, and B. Hyeon, A systematic review on factors influencing multicultural acceptance in Korean adolescents. Journal of the Korea Academia-Industrial cooperation Society, vol. 19, no. 7, pp. 207-213, 2018.

[10] S. J. Kim, A study on the influence of experiences of discrimination to multicultural families' adolescents on their characteristics - focused on raw-data of National Survey of Multicultural Families 2012. The Journal of Asiatic Studies, vol. 58, no. 3, pp. 6-41, 2015.

[11] Y. S. Choi, Personal characteristics, ethnic identity, experience of discrimination, self-esteem, and problem behavior of Korean-Japanese multicultural adolescents. Korean Journal of Family Welfare, vol. 17, no. 2, pp. 49-71, 2012.

[12] A. R. Lee, J. Lee, and B. Y. Son, A qualitative study on the multi-cultural adolescents experience of career barrier. Korean Journal of Youth Studies, vol. 25, no. 11, p. 35-64, 2018.

[13] H. Byeon, Predicting the anxiety of patients with Alzheimer's dementia using boosting algorithm and data-level approach, International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, pp. 107-113, 2021.

[14] H. Byeon, Comparing ensemble-based machine learning classifiers developed for distinguishing hypokinetic dysarthria from presbyphonia. Applied Sciences, vol. 11, no. 5, pp. 2235, 2021.

[15] H. Byeon, S. Cha, and K. Lim, Exploring factors associated with voucher program for speech language therapy for the preschoolers of parents with communication disorder using weighted random forests. International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 12-17, 2019.

[16] H. Byeon, Developing a random forest classifier for predicting the depression and managing the health of caregivers supporting patients with Alzheimer's disease. Technology and Health Care, vol. 27, no. 5, pp. 531-544, 2019.

[17] Ministry of Gender Equality & Family, A study on the national survey of multicultural families. Ministry of Gender Equality & Family, Seoul, 2012.

[18] V. K. Chauhan, K. Dahiya, and A. Sharma, Problem formulations and solvers in linear SVM: a review. Artificial Intelligence Review, vol. 52, no. 2, pp. 803-855, 2019.

[19] D. Li, L. Xu, E. D. Goodman, Y. Xu, and Y. Wu, Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking. Integrated Computer-Aided Engineering, vol. 20, no. 3, pp. 201-216, 2013.

[20] A. Sarica, A. Cerasa, and A. Quattrone, Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in Aging Neuroscience, vol. 9, pp. 329, 2017.

[21] I. A. Ibrahim, T. Khatib, A. Mohamed, and W. Elmenreich, Modeling of the output current of a photovoltaic grid-connected system using random forests technique. Energy Exploration & Exploitation, vol. 36, no. 1, pp. 132-148, 2018.

[22] E. K. Sahin, I. Colkesen, and T. Kavzoglu, A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. Geocarto International, vol. 35, no. 4, pp. 341-363, 2020.

[23] J. Rodriguez, L. Kuncheva, and C. Alonso, Rotation forest: a new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pp. 1619-1630, 2006.

[24] E. K. Sahin, I. Colkesen, and T. Kavzoglu, A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. Geocarto International, vol. 35, no. 4, 341-363, 2020.

[25] H. Li, Z. Zhang, and Z. Liu, Application of artificial neural networks for catalysis: a review. Catalysts, vol. 7, no. 10, pp. 306. 2017.

[26] R. Adhikari, A neural network based linear ensemble framework for time series forecasting. Neurocomputing, vol. 157, pp. 231-242, 2015.

[27] Y. Ren, L. Zhang, and P. N. Suganthan, Ensemble classification and regression-recent developments, applications and future directions. IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 41-53, 2016.

[28] F. Divina, A. Gilson, F. Gomez-Vela, M. Garcia Torres, and J. F. Torres, Stacking ensemble learning for short-term electricity consumption forecasting. Energies, vol. 11, no. 4, pp. 949, 2018.

[29] R. Saini, and S. Ghosh, Ensemble classifiers in remote sensing, a review 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, pp. 1148-1152, 2017.

[30] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, A deep learning-based multi-model ensemble method for cancer prediction. Computer Methods and Programs in Biomedicine, vol. 153, pp. 1-9, 2018.

[31] L. Feng, Y. Li, Y. Wang, and Q. Du, Estimating hourly and continuous ground-level PM2.5 concentrations using an ensemble learning algorithm: the ST-stacking model. Atmospheric Environment, vol. 223, pp. 117242, 2020.

[32] J. Chen, J. Yin, L. Zang, T. Zhang, and M. Zhao, Stacking machine learning model for estimating hourly PM2. 5 in China based on Himawari 8 aerosol optical depth data. Science of The Total Environment, vol. 697, pp. 134021, 2019.

[33] S. Lee, and H. Kim, A new ensemble machine learning technique with multiple stacking. Society for e-Business Studies, vol. 25, no. 3, pp. 1-13, 2020.

Predicting DOS-DDOS Attacks: Review and Evaluation Study of Feature Selection Methods based on Wrapper Process

Kawtar BOUZOUBAA[1], Benayad NSIRI[3]
M2CS, Research Center STIS
National Graduate School of Arts and Crafts of Rabat
(ENSAM) Mohammed V University in Rabat
Rabat, Morocco

Youssef TAHER[2]
Center of Guidance and Planning (COPE)
Rabat, Morocco

***Abstract*—Now-a-days, Cybersecurity attacks are becoming increasingly sophisticated and presenting a growing threat to individuals, private and public sectors, especially the Denial Of Service attack (DOS) and its variant Distributed Denial Of Service (DDOS). Dealing with these dangerous threats by using traditional mitigation solutions suffers from several limits and performance issues. To overcome these limitations, Machine Learning (ML) has become one of the key techniques to enrich, complement and enhance the traditional security experiences. In this context, we focus on one of the key processes that improve and optimize Machine Learning DOS-DDOS predicting models: DOS-DDOS feature selection process, particularly the wrapper process. By studying different DOS-DDOS datasets, algorithms and results of several research projects, we have reviewed and evaluated the impact on used wrapper strategies, number of DOS-DDOS features, and many commonly used metrics to evaluate DOS-DDOS prediction models based on the optimized DOS-DDOS features. In this paper, we present three important dashboards that are essential to understand the performance of three wrapper strategies commonly used in DOS-DDOS ML systems: heuristic search algorithms, meta-heuristic search and random search methods. Based on this review and evaluation study, we can observe some of wrapper strategies, algorithms, DOS-DDOS features with a relevant impact can be selected to improve the DOS-DDOS ML existing solutions.**

Keywords*—DOS-DDOS attacks; feature selection; wrapper process; machine learning

I. INTRODUCTION

With the exponential proliferation of Internet users, the network traffic has known a massive generation of data. These data are coming from individuals, private and public organizations. Moreover, the hard complexity of the Internet architecture and its interdependent suffers from different vulnerabilities, threats and risks ([1], [2]). Consequently, the attackers find an impressive amount of vulnerable systems [3].

Nowadays, cybersecurity attacks are becoming increasingly sophisticated, particularly the infrastructure attacks that make security analysis systems more vulnerable to several failures [1]. One of these most famous threats is Denial Of Service attack (DOS) and its variant Distributed Denial Of Service (DDOS) ([4],[5]). These serious and dangerous attacks violate the availability of information systems, which is a pillar of information security ([6],[5]**)**. The attackers seek to target computer systems, network devices, services and web applications to consume their CPU power, bandwidth, memory and processing time ([7], [3]).

The DDOS attack has the same purpose but with the difference of using intermediate of multiple networks between the attacker and its target ([7],[8]). This technique allows the attacker to amplify its attack with orchestrating a simultaneous sending of an excessive number of unwanted computing requests to its victim to overload its computing capacity.

To deal with these DOS-DDOS attacks, some traditional mechanisms are deployed such as firewalls, software updates, antivirus, Intrusion Detection Systems (IDS), etc.

However, many challenges and limits hinder these traditional techniques [6]. To overcome these limitations and drawbacks, Machine Learning (ML) techniques can be used as artificial intelligence systems to enrich, complement and enhance the traditional security experiences.

One of the key and critical pre-processing phases to success these DOS-DDOS ML models is feature selection. This process selects the most representatives DOS-DDOS characteristics from the initially DOS-DDOS dataset by eradicating those that are redundant and insignificant. Consequently, the obtained features subset improves the execution time, the detection rate and the accuracy of the used DOS-DDOS models.

In this context, this investigation presents a review and evaluation study related to DOS-DDOS attacks prediction based on one of the effective methods to select relevant DOS-DDOS features: Wrapper process.

This paper is organized as follows: In Section 2 we study some traditional mitigation solutions and their limits. Section 3 describes the interest of using machine learning (ML) in DOS- DDOS attacks prevention. Section 4 exposes the impact of feature selection on DOS-DDOS machine learning projects. In Section 5 we review and we evaluate recent and relevant feature selection results obtained by using three commonly used wrapper strategies: heuristic search algorithms, meta-heuristic search and random search methods. Finally, Section 6 presents our conclusions.

II. Dealing with DOS and DDOS: Traditional Mitigation and Solutions

DOS-DDOS attacks can take many forms such as SYN flood, SYN-ACK-ACK flood, UDP flood, ICMP flood, and so on. To deal with these forms of threats, many traditional, external and internal DOS-DDOS mitigation solutions are developed such as bandwidth provisioning, software updates, firewalls, antivirus software and Intrusion Detection Systems (IDS), etc. In the paragraph below, we discuss briefly these traditional solutions and their limits.

Generally, the use of firewall solution provides many mitigation solutions such as filter-based forwarding at logical interfaces, blocking of certain types of packets to reach a routing engine and packet counter and protection of a routing engine from DOS-DDOS attacks ([9],[10]). However, firewall solutions suffer from many lacks of security. As an example, the attacker can modify his DOS-DDOS attacks and make it legitimate.

The software updates keep the software up to date to avoid DOS-DDOS attacks on the application layer (the highest abstraction layer of the TCP/IP model) [11]. However, the irregularity of these updates creates a gateway to the attackers to modify the contents of memories (buffer overflow).

The Intrusion Detection System IDS (Hardware/Software solutions) is a complemented security for the firewall solutions. This solution is a common way often used to analyze and detect DOS-DDOS attacks [12]. IDS techniques are used in the aims to detect, classify and respond to DOS-DDOS actions that affect the integrity, the confidentiality or the availability of any network resources [13]. These systems are mainly based on two detection methods [14]: Misuse Detection (MD) and Anomaly Detection (AD).

The Misuse Detection is also known as Signature Detection, Pattern Detection, Knowledge-Based or Rule-based detection. This technique is one of the most common methods of Antivirus. It filters malicious packet of the known attacks thanks to its signature database of known attacks. It detects efficiently known attacks with low false positive. Nevertheless, it shows limits on detecting new forms of threats and many variants of known attacks.

The Anomaly Detection supervises the behavior of network traffic. It alerts the system at the slightest changes compared to the normal behavior. This method can detect new forms of attacks but generates high false positives and doesn't give clear information about the malicious events in some forms of attacks. Moreover, it is not feasible to IDS to manipulate high dimensional variables. Consequently, this technique can affect the efficiency and the velocity in detecting intrusions ([15],[16], [17]).

In addition to the limitations and drawbacks mentioned above, traditional techniques are hindered by many others challenges [6]. As an example, many traditional strategies of security are not sufficient to protect information systems against the new forms of DOS-DDOS attacks, need extra-storage and computational resources due to the high level of network traffic, suffer from a lack of source attacks information and are unable to detect and prevent many DOS-DDOS attacks in real-time.

To overcome these drawbacks, Machine Learning has become one of the key techniques to enrich and complement these traditional security experiences. In the paragraph below we discuss briefly the benefits that can be attained by using ML- techniques in DOS-DDOS attacks prevention.

III. The Use of Machine Learning in DOS-DDOS Attacks Prevention

Machine Learning (ML) is an evolutionary field of Artificial Intelligence (AI) composed of a set of rules, methods and functions [18]. Applied to deal with many challenges in DOS-DDOS attacks, ML algorithms can learn from DOS-DDOS datasets and discover hidden knowledge from them [19].

By finding interesting DOS-DDOS patterns from training DOS-DDOS data, ML algorithms allow preventing and predicting many recent forms of DOS-DDOS behaviors.

Contrary to the traditional security solutions, ML models are powerful tools that can analyze in real time high dimensional DOS-DDOS traffic [20], classify the behavior of the DOS-DDOS traffic to determine the normal one from the abnormal and predict with high accuracy DOS-DDOS attacks before they happen.

Based on DOS-DDOS security modeling process (Fig. 1) and many common algorithms like K-Nearest Neighbors Algorithm (KNN), Support Vector Machines (SVM), Random Forest (RF) as well as Naïve Bayes (NB), etc. many recent research projects have shown other important preventing benefits of ML algorithms compared to the existing traditional solutions ([1], [12], [21]).

Feature selection is one of the critical pre-processing process to succeed and to improve the benefits mentioned above. In the paragraph below, we summarize the benefits of this process.

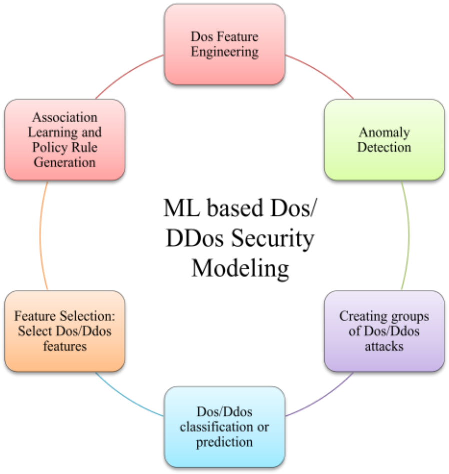

Dos Feature Engineering
Anomaly Detection
Creating groups of Dos/Ddos attacks
Dos/Ddos classification or prediction
Feature Selection: Select Dos/Ddos features
Association Learning and Policy Rule Generation
ML based Dos/ DDos Security Modeling

Fig. 1. Machine Learning DOS-DDOS Security Modeling Process

IV. Impact of Feature Selection Process DOS-DDOS Machine Learning Projects

Feature selection is one of the most critical pre-processing process in building DOS-DDOS Machine Learning (ML) models. This process is the first and crucial phase to improve the prediction accuracy, the detection rate and to reduce the execution time of DOS-DDOS models [22].

According to Bindra et al. [23], feature selection methods allow the DOS-DDOS security systems to distinguish DOS-DDOS attacks by using a minimum number of the most important features from network streams.

Applied to DOS-DDOS ML algorithms, feature selection is focused on selecting small and concise DOS-DDOS sets of characteristics describing the ML models [24]. It avoids the used features to contain redundant (correlation with other features) and noisier information of DOS-DDOS attacks without losing any piece of information. Consequently, it reduces the high memory requirements of security systems based on ML models ([25], [26], [27]).

Generally, the existing DOS-DDOS ML security systems use three commonly main categories of feature selection approaches: Filter, Wrapper and Hybrid methods [28].

The Filter methods are based on statistical methods which evaluate the relevance of DOS-DDOS features independently of any machine learning algorithms [27]. As a faster solution that computationally costs less, these methods are often used in high dimensional DOS-DDOS traffic ([29],[30]). However, the evaluation of individual information cannot take into consideration the correlation between the DOS-DDOS features. Consequently, the final DOS-DDOS subset can contain redundancy because some DOS-DDOS features can have the same ranking.

The wrapper strategies use a predetermined algorithm and its performance to assess the optimal DOS-DDOS subset features [31]. It executed in an iterative process, and at each iteration a new subset of DOS-DDOS features is generated to be evaluated by the classification algorithm [32]. The criterion of selection is principally based on the cross-validation accuracy during the DOS-DDOS training data [33].

The Hybrid method is a combination between filter method followed by wrapper approach, which offers the advantages of the two previous methods. It exploits their different criteria in different search stages [34].

V. Related Work

A. Objective of the Study

To detect and prevent DOS-DDOS attacks accurately, wrapper methods one of the *most effective* strategies to identify informative DOS-DDOS feature subsets from many high-dimensional DOS-DDOS network streams. This approach of feature selection is often addressed in many security solutions based on ML tasks. Indeed, increasing number of research projects have shown that many wrapper strategies can have an important impact on Accuracy, Detection Rate and time execution of existing DOS-DDOS ML systems.

In this context, we decided to focus our attention on the assessment of the performance of many DOS-DDOS experiments based on wrapper strategies and machine learning algorithms.

By studying different DOS-DDOS datasets, algorithms and recent results of several research projects, we review and we assess the impact of many recent wrapper strategies applied to predicting DOS-DDOS attacks. We have taken a more focused look at the impact of these strategies on number of DOS-DDOS features, detection rates, execution times and accuracies of DOS-DDOS attacks prediction.

We present four dashboards that are essential to understand the performances of three wrapper strategies commonly used in DOS-DDOS ML systems: heuristic search algorithms, meta-heuristic search and random search methods.

B. Review and Evaluation Study of Feature Selection Methods based on Wrapper Process

1) Used Datasets: To evaluate the performance of the wrapper strategies used in DOS-DDOS machine learning models, we start our review by studying relevant DOS-DDOS datasets commonly used by several DOS-DDOS research projects. These datasets are cited below:

The Knowledge Discovery and Data Mining (KDD'99) dataset was built based on the synthetic data captured in DARPA'98. This dataset is mainly composed of redundant records. Moreover, this configuration forces ML algorithms to learn less about infrequent records than the redundant ones. The inequality of attacks distribution between training and testing phase made the cross-validation more complicated.

This dataset is composed of four main families of attacks and forty one features.

The NSL_KDD was created to overcome the limits of the KDD'99 [35]. However, the main disadvantage of the NSL_KDD dataset, it does not include the modern low footprint attacks scenarios like the KDD'99.

The UNSW_NB15 is composed of nine family attacks and forty nine features. It includes a hybrid of the real modern normal behaviors and the synthetic attack activities [35].

Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) is a dataset mainly composed of hybrid modern normal activities and attacks behaviors. It is composed of forty-seven features[36].

2) Use model evaluation metrics: To evaluate the reviewed DOS-DDOS Wrapper strategies, we have selected different metrics [37]. These metrics namely are: Classification Accuracy (Acc), Detection Rate (DR), Recall (Re), Precision (Pr), Specificity (Sp), Sensitivity (Sen), F-Measure (FM), False Alert Rate (FAR), False Negative (FN) and Time model execution (T).

The formulas associated with these metrics are listed above:

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (1)$$

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN+FP})} \tag{5}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{(\mathrm{TP+FN})} \tag{6}$$

$$\mathrm{FM} = \frac{2\times\mathrm{Re}\times\mathrm{Pr}}{(\mathrm{Re+Pr})} \tag{7}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{8}$$

Where: TP is True Positive: correct positive prediction. TN is True Negative: correct negative prediction. FN is False Negative: incorrect negative prediction and FP is False Positive: incorrect positive prediction.

3) Impact of used DOS-DDOS datasets and algorithms on the wrapper process: Generally, the performance of DOS-DDOS prediction models based on the Wrapper process depends strongly on the used ML algorithms and datasets. As shown in Table I, many algorithms performed well in detecting DOS-DDOS attacks compared to others. The accuracy can range from Acc=62.5% by using KDD'99 dataset and SVM algorithm to Acc=99.92% with Decision Tree J.48 algorithm and KDD'99 dataset. Indeed, according to the experiment of Jalill et al. (2010) [38] based on the KDD'99 dataset, the Support Vector Machine (SVM) algorithm has a serious problem in accurately detecting DOS-DDOS attacks compared to the Decision Tree J.48 algorithm which shows high prediction accuracy that exceed 99%.

TABLE I. IMPACT OF USED DOS-DDOS DATASETS AND ALGORITHMS ON THE WRAPPER PROCESS

Reference	**Dataset**	**Algorithm**	**Accuracy(%)**
Jalill et al. [38]	KDD'99	SVM	62.5
		J.48	99.7
Katkar and Kulkarni. [40]	KDD'99	J.48	99.92
		REPTree	99.56
		NB	87.50
		BN	99.68
		Sequential Minimal Optimization (SMO)	99.72
		REPTree + J48 +BN	99.94
Bellouch et al. [39]	UNSW_NB15	SVM	92.28
		NB	74.19
		C4.5	95.82
		RF	97.49

The experiments based on the NB, C4.5, RF algorithms and UNSW_NB15 dataset realized by Bellouch et al. (2018) [39], has shown that the prediction accuracy obtained by RF (Acc__{RF} = 99.94%) is better than C4.5 ($\text{Acc\_}_{C4..5}$ = 95.82%) and SVM (Acc__{SVM} = 92.28%). The NB algorithm shows less accuracy (Acc__{NB} = 74. 19 %) compared to RF, C4.5 and SVM.

The Bayesian Network (BN) algorithm used in the experiment of Katkar and Kulkarni [40] achieved good accuracy (Acc__{BN} = 99.68%) in detecting DOS-DDOS attacks thanks to its capacity of detecting anomalies in a multi-class [41].

By comparing the experiments carried out by Jalill et al.[38] and Katkar and Kulkarni [40], we have observed that SVM algorithm predict DOS-DDOS more accurately on the dataset UNSW_NB15 compared to the KDD'99 dataset ($\text{Acc\_}_{SVM\_UNSW\_NB}$ = 92.28% > Acc__{SVM_KDD} = 62.5 %). This important difference according to W. Xingzhu [42] is caused by the redundant records on the KDD'99 dataset and SVM has slower training on high dimensional datasets.

4) DOS-DDOS feature selection based on wrapper process and heuristic search algorithms: Based on heuristic functions or cost measures, wrapper strategies using heuristic search algorithms optimize and iteratively improve the process of DOS-DDOS feature selection [43].

Many heuristic searches such as SFS (Sequential Forward search), SBS (Sequential Backward search), LRS (Plus L Minus R Selection), RELR (Random Effect Logistic Regression), and GFR (Gradually feature removal method) have been used by many recent important research projects to solve accurately the problem of DOS-DDOS feature selection.

We discuss these projects in the paragraph below. At the end of this subsection, we present our first dashboard (Tables IIA, IIB, IIC) to summarize and to compare the performances of these strategies.

As an example of wrapper strategies based on heuristic search algorithms, we can cite the important investigation of Kavitha and Chrita (2010) [44]. In this study, the authors used the Best First Search (BFS) method. They selected two subsets composed simultaneously of seven and fourteen DOS-DDOS features. They applied four classifying algorithms: ID3, J48, NB and One R. These experiments have shown that ID3 and J.48 using a subset composed of fourteen DOS-DDOS features has the highest accuracy (Acc = 99%). One R and NB performed well in execution time (T=0.5s) with only seven features. The NB classifier achieved the highest specificity with Sp__{NB} = 99% by using seven features and Sp__{NB} =100% by using fourteen features.

Mok et al. (2010) [45] used Random Effect Logistic Regression (RELR) with a fixed Logistic regression (LR). This method selected five DOS-DDOS features by using the Stepwise Variable Selection Search (SVSS) strategy based on the KDD'99 dataset. The method achieved an accuracy equal to 98.74%.

TABLE II. (A): WRAPPER METHOD BASED ON HEURISTIC SEARCH (HS)

DOS-DDOS feature selection projects based on wrapper methods	DOS- DDOS used dataset	Used wrapper strategy	Number of DOS - DDOS features		Used classifier	Used Metrics	Metrics Values with FS		Metrics Values without FS
Kavitha, and Chitra (2010) [44]	KDD’99	BFS	7	14	ID3	Accuracy Sensitivity Specifity Time (s)	97% 97% 97% 1.49	99% 100% 98% 4.01	99% 98% 100% NA
					J48	Accuracy Sensitivity Specifity Time (s)	97% 97% 97% 1.20	99% 99.5% 97.5% 1. 86	99.9% 97.8% 99.9% NA
					NB	Accuracy Sensitivity Specifity Time (s)	96% 92% 99% 0.05	97% 94% 100% 0.09	99% 98% 100% NA
					OneR	Accuracy Sensitivity Specifity Time (s)	86% 74% 99% 0.05	97% 72% 92% 0.16	99.5% 98% 99.7% NA
Mok et al. (2010) [45]	KDD’99	Stepwise	5		RLER	Accuracy	98.74%		NA
Ahmad et al. (2011) [46]	KDD’99	PCA-GA	12		MLP	Accuracy Time (h)	99% 72		NA
Yinhui et al. [47]	KDD’99	SBS-GFR	19		SVM	Accuracy Time(s)	98.62% 2.37		98.67% 3.97

TABLE II- (B): WRAPPER METHOD BASED ON HEURISTIC SEARCH (HS)

DOS-DDOS feature selection projects based on wrapper methods	DOS- DDOS used dataset	Used wrapper strategy	Number of DOS - DDOS features		Used classifier	Used Metrics	Metrics Values with FS		Metrics Values without FS
Zhang and Wang (2013) [48]	NSL_KDD	SBS-BN	11		BN	Accuracy Time(s)	98.98% 4.73		95.7% 18.94
Al-Jarrah et al.(2014) [49]	KDD’99	FSR-RF	15		RF	Accuracy	99.90%		99.89%
		BER-RF	14				99.88%		
Lee et al. (2017) [50]	NSL_KDD	SFFS-RF	10		C4.5	Accuracy Detection Rate FAR Time(s)	99.89% 99.9% 0.1 0.18		NA NA 1.07 NA
Harish and Manju (2018) [51]	KDD’99	FDR + PLR	20	40	KNN	Accuracy Time(s)	98.5% 17.98	99.0% 32.95	NA
		FDR +SFS	25			Accuracy Time(s)	98.27 17.74		NA
		FDR +SBS	40			Accuracy Time(s)	98.78% 32.18		NA NA
Houseini Soodeh and Mehrdad (2019) [52]	NSL_KDD	Forward Feature Selection	12		NB	Accuracy Precision Recall F-measure	93.1% 93.6% 87.3% 92.7%		NA
			14		RF	Accuracy Precision Recall F-measure	98.9% 99.6% 99.8% 99.7%		NA
			10		DT	Accuracy Precision Recall F-measure	98.2% 99.4% 99.8% 99.6%		NA

			20	MLP	Accuracy Precision Recall F-measure	96.1% 93.4% 91.8% 94.9%	NA
			11	KNN	Accuracy Precision Recall F-measure	97.7% 99.8% 99.8% 99.8%	NA
Malhotra and Sharma (2019) [53]	NSL_KDD	CfsSubsetEval + BestFirst	6	RF	Accuracy Time (s)	99,41% 66.82	99,91% 191.06
				Bagging	Accuracy Time (s)	99,35% 17,7	99,84% 109.9%
				PART	Accuracy Time (s)	99,37% 8.07	99,83% 99.1
				J48	Accuracy Time (s)	99,78% 7.95	99,78% 61.68
Wang et al.(2020) [54]	NSL_KDD	SBS-MLP	31	MLP	Accuracy Detection Rate FAR	97.66% 94.88% 0.62%	97.61% 94.78% 0.63%
Polat, and Cetin (2020) [55]	Their Dataset composed of 12 Features	SFFS	10	SVM	Accuracy Sensitivity Specificity Precision F_measure	92.15% 90.20% 97.26% 90.23% 90.21%	92.11% 88.71% 96.93% 91.42% 89.91%
			6	KNN	Accuracy Sensitivity Specificity Precision F_measure	98.30% 97.73% 99.45% 97.72% 97.70%	95.67% 93.87% 98.01% 97.05% 95.30%

TABLE II-(C): WRAPPER METHOD BASED ON HEURISTIC SEARCH (HS)

DOS-DDOS feature selection projects based on wrapper methods	**DOS- DDOS used dataset**	**Used wrapper strategy**	**Number of DOS -DDOS features**	**Used classifier**	**Used Metrics**	**Metrics Values with FS**	**Metrics Values without FS**
Polat, and Cetin (2020) [55]	Their Dataset composed of 12 Features	SFFS	6	ANN	Accuracy Sensitivity Specificity Precision F_measure	91.44% 87.82% 97.31% 88.11% 87.89%	91.07% 87.27% 96.58% 89.89% 88.45%
			8	NB	Accuracy Sensitivity Specificity Precision F_measure	94.87% 92.05% 98.43% 93.29% 92.01%	94.48% 91.77% 98.29% 92.94% 91.79%
Alabdulwahab and Moon (2020) [31]	NSL_KDD	CfsSubsetEval + BestFirst	6	RePTree	Accuracy Time(s)	99,44% 5,76	99,83% 3.59
				Logiboost	Accuracy Time(s)	94,15% 9,96	97,1% 18.3
				RBF	Accuracy Time(s)	90,6% 45.91	97,95% 81.01
				BayesNet	Accuracy Time(s)	96,26% 5.64	97,17% 4.69
				SMO	Accuracy Time(s)	89,09% 514.7	97,4% 1137.71
				NBTree	Accuracy Time(s)	99,46% 14.23	99,87% 213.18
Umar et al. (2020) [56]	UNSW_NB15	Best First Forward-DT	19	ANN	Accuracy Detection Rate FAR Time(s)	82.08% 97.94% 37.36% 240	86.00% 98.62% 29.45% 660

					SVM	Accuracy Detection Rate FAR Time(s)	79.11% 99.31% 45.64% 15540		81.6% 99.64% 40.51 10860	
					KNN	Accuracy Detection Rate FAR Time(s)	83.21% 96.44% 33.01% 600		84.78% 96.46% 29.53% 1020	
					RF	Accuracy Detection Rate FAR Time(s)	86.41% 97.95% 27.73% 37.8		86.82% 98.7% 27.74% 44.4	
					NB	Accuracy Detection Rate FAR Time(s)	55.61% 19.38% 0.01% 2.86		55.61% 19.39% 0.01% 4.64	
Umar and Chen (2020) [57]	UNSW_NB15	NSL_KDD	Best First - DT	20	ANN	Accuracy Detection Rate FAR Time(s)	94.32% 98.48% 14.56% 325	98.9% 99.0% 1.11% 123	94.62% 97.54% 11.64% 348	99.6% 99.6% 0.23% 94
					SVM	Accuracy Detection Rate FAR Time(s)	93.56% 99.54% 19.19% 10236.6	98.0% 97.1% 1.17% 921.6	93.67% 99.63% 19.14% 5213.4	98.5% 98.1% 1.08% 972.6
					KNN	Accuracy Detection Rate FAR Time(s)	95.8% 97.28% 7.36% 502.8	99.1% 99.2% 0.97% 331.2	93.81% 96.24% 11.42% 747.6	99.5% 99.4% 0.36% 563.4
					RF	Accuracy Detection Rate FAR Time(s)	98.51% 99.17% 2.89% 33.6	99.7% 99.7% 0.22% 13.2	95.74% 97.84% 8.77% 32.4	98.8% 99.7% 0.1% 15

Ahmad et al. (2011) [46] used Principal Components Analysis (PCA) to reduce the features and to choose the highest eighteen values. Genetic Algorithm (GA) was applied as wrapper method to the reduce space. This method selected twelve DOS-DDOS features. By using the Multi Layer Perceptron (MLP) as classifier on the output of GA and the KDD'99 dataset, this model has shown high accuracy (Acc_$_{MLP}$ = 99%) by using a minimum of features equal to 12 and the time of execution equal to 72 h.

L. Yinhui et al. (2012) [47] applied Gradually Feature Removal method (GFR) which selected nineteen best DOS-DDOS features. This strategy was based on SBS as search strategy and SVM as classifier. The accuracy of this model has been slightly reduced (Acc_$_{(19\ features)}$ = 98.62% < Acc_$_{(42\ features)}$ = 98.67%) by using a wrapper step. The execution time has been reduced from T $_{\_(42\ features)}$ = 18.94s to T_$_{(19\ features)}$ = 3.73 s.

Zhang and Wang (2013) [48] adopted SBS-BN and Bayesian network approach as a wrapper strategy. This experiment selected three best DOS-DDOS features and achieved good accuracy (Acc_$_{(3\ features)}$ = 98.98% > Acc_$_{(42\ features)}$ = 95.7%) with an interesting time of execution (T_$_{(3\ features)}$ = 2.37s < T_$_{(42\ features)}$ = 3.97s).

Al-Jarrah et al. (2014) [49] proposed a set of RF algorithm with forward and backward elimination ranking features selection techniques. This experiment demonstrated that FSR-RF outperforms with fifteen best features, BER-RF with fourteen features and RF with all used DOS-DDOS features: (Acc_$_{(15\ features)}$ = 99.98% > Acc_$_{(14\ features)}$ = 99.88 %) and (Acc_$_{(15\ features)}$ = 99.98% > Acc_$_{(42\ features)}$ = 99.89%).

J. Lee et al. (2017) [50] proposed SFFS-RFC to generate DOS-DDOS features subset and to measure the performance of each subset. This experiment has shown that SFFS-RFS improved the performance of the accuracy and the detection rate of attacks classification with only ten DOS-DDOS (Acc_$_{(10\ features)}$ = 99.89% and DR_$_{(10\ features\ )}$ = 99.9%). It realized a fewer FAR (FAR$_{\_(10\ features)}$ = 0.1% < FAR_$_{(41\ features)}$ = 1.7%) compared to the existing methods using the classifier C4.5 and reduced the execution time (T_$_{(10\ features)}$ = 0.18 s).

Harish and Manju (2018) [51] combined the Fisher Ratio Discrimination (FRD) with three different search strategies: SFS, SBS and LRS. They concluded that FDR using LRS, KNN and twenty DOS-DDOS features outperformed other methods. Thanks to its capacity to remove non-performing DOS-DDOS features from the initial subset, this strategy achieved a better accuracy with twenty features (Acc_$_{(20\ features)}$ = 98.87% > Acc_$_{SFS\_(25\ features)}$ = 98.27%) compared to FDR-SFS which selected 25 features. However, the execution time of FDR-SFS is less than FDR-LRS (T_$_{SFS\_(25\ features)}$ = 17.74 s < T_$_{SFS\_(20\ features)}$ = 17.98 s). On the other side the FDR-LRS with forty features showed a good accuracy compared to the accuracy of FDR-SBS with the same number of features (Acc_$_{LRS\_(40\ features)}$ = 99.09% > Acc_$_{SBS\_(40\ features)}$ = 98.78%). However the execution time of FDR-SBS is better compared to FDR-LRS (T_$_{SBS\_(40\ features)}$ = 32.18s < T_$_{LRS\_(40\ features)}$ = 32.95s).

Soodeh and Mehrdad (2019) [52] proposed a new framework composed of a hybridization of different algorithms. The objective of this framework is to handle new types of attacks better than other existing frameworks based on Forward Feature Selection (FFS). By using NSL_KDD dataset, this framework has shown that RF outperformed other algorithms with only thirteen features in attack detection accuracy ($Acc_{\_(13\ features)} = 98.9\%$). In the case of DOS-DDOS attacks, the KNN classifier has achieved the highest precision with eleven features ($Pr_{\_(11\ features)}= 99.8\%$). The classifiers RF, DT and KNN achieved the highest Recall value (Re = 99.8 %), and the highest F-measure ($FM_{\_RF\_(14\ features)} = 99.7\%$, $FM_{\_DT\_(10\ features)} = 99.6\%$, and $FM_{\_KNN\_(11\ features)} = 99.8\%$). The classifier NB showed the lowest measured values of all these metrics: $Acc_{\_NB} = 93.10\%$, $Pr_{\_NB} = 93.6\%$, Re = 87.3%, $FM_{\_NB} = 92.7\%$.

Malhotra and Sharma (2019) [53] used CfsSubsetEval and Best First as wrapper method. Based on NSL_KDD dataset and RF Bagging, PART and J.48 algorithms, this strategy selected eight best DOS-DDOS features. It increased slightly the accuracy and decreased significantly the execution time for all the classifiers. The accuracy of J.48 is 99.78% by using 6 and 42 features. However, this strategy decreased the execution time ($T_{\_J.48\_(42\ features)} = 61.68s > T_{\_J.48\_(6\ features)} = 7.95s$). The RF model decreased slightly the accuracy ($Acc_{\_RF\_(6\ features)} = 99.41\% < Acc_{\_RF\_(42\ features)} = 99.91\%$), and decreased drastically the execution time ($T_{\_RF\_(6\ features)} = 66.82s < T_{\_RF\_(42\ features)} = 191.06\ s$).

M. Wang et al. (2020) [54] combined SBS with Multi Layer Perceptron (MLP) to select the optimal DOS-DDOS features by using NSL_KDD dataset. This experiment showed that SBS-MLP can find an optimal DOS-DDOS feature subset and performed better accuracy than the full DOS-DDOS feature set among all the MLP-based detection methods ($Acc_{\_(31\ features)} = 97.66\% > Acc_{\_(42\ features)}= 97.61\%$). It enhanced the detection rate ($DR_{\_(31\ features)} = 94.88\ \%> DR_{\_(42\ features)} = 94.78\%$). It decreased the FAR value ($FAR_{\_(31\ features)} = 0.62\% < FAR_{\_(42\ features)} = 0.63\%$).

Polat et al. (2020) [55] evaluated the classifiers SVM, KNN, ANN and NB on their dataset initially composed of twelve features. This experiment used SFFS as a wrapper approach. They evaluated the performance of this approach by calculating many metrics: accuracy, sensitivity, specificity, precision and F-measure. By using a wrapper step and only selected DOS-DDOS features instead of all features, these different models increased the accuracy ($Acc_{\_ANN\_(6\ features)} = 91.44\% > Acc_{\_ANN\_(42\ features)} = 91.07\%) < (Acc_{\_SVM\_(10\ features)} = 92.15\% > Acc_{\_SVM\_(42\ features)} = 92.11\%) < (Acc_{\_NB\_(8\ features)} = 94.87\% > Acc_{\_NB\_(42\ features)} = 94.48\%) < (Acc_{\_KNN\_(8\ features)}= 98.30\% > Acc_{\_KNN\_(42\ features)}= 95.67$. However, the precision of SVM and KNN is slightly decreased by integrating the feature selection process compared to the initial set with all features ($Pr_{\_SVM\_(10\ features)} = 90.23\% < Pr_{\_SVM\_(42\ features)} = 91.42\%$), ($Pr_{\_ANN\_(6\ features)} = 88.11\% < Pr_{\_ANN\_(42\ features)} = 89.89\%$). The specificity is enhanced for all the used models, particularly by using a KNN model ($Sp_{\_SVM\_(10\ features)} = 97.26\%$, $Sp_{\_ANN\_(6\ features)} = 97.31\%$, $Sp_{\_NB\_(8\ features)} = 98.43\%$, $Sp_{\_KNN\_(6\ features)} = 99.45\%$).

Alabdulwahab and Moon (2020) [31] used the NSL_KDD dataset to evaluate different algorithms based on CfsSubsetEval and Best First as wrapper strategy. They tested CfsSubsetEval with six supervised classifiers: Logiboost, RBF, BayesNet, SMO and RepTree. By using six most relevant DOS-DDOS features, this experiment has shown an important improvement of the execution time ($T_{\_NBTree\_(6\ features)} = 14.23s < T_{\_NBTree\_(42\ features)} = 213.18s$, $T_{\_Logiboost\_(6\ features)} = 9.96s < T_{\_Logiboost\_(42\ features)} = 18.3s$. However, the accuracy was better without using the wrapper process ($Acc_{\_NBTre\_(6\ features)} = 99.46\ \% < Acc_{\_NBTree\_(42\ features)} = 99.87\%$). However, the RepTree algorithm decreased the accuracy and increased the execution time ($Acc_{\_RepTree\_(6\ features)} = 99.44\% < Acc_{\_RepTree\_(42\ features)} = 99.83\%$, $T_{\_RepTree\_(6\ features)} = 5.76s > T_{\_RepTree\_(42\ features)} = 3.59\ s$).

Umar et al. (2020) [56] applied Best First Forward as search strategy and DT to evaluate the performance of their detecting attacks model. This strategy selected nineteen best features by using UNSW_NB15 dataset. The assessment of this experiment was based on five metrics: Acc, DR, FAR and T. This method has shown that the execution time has overall decreased for different used classifiers ($T_{\_ANN\_(19\ features)} = 240s < T_{\_ANN\_(42\ features)}=660s$, RF ($T_{\_RF\_(19\ features)} =37.8s < T_{\_RF\_(19\ features)}= 44.4s$), NB ($T_{\_NB\_(19\ features)} = 2.86\ s < T_{\_NB\_(42\ features)} = 4.64\ s$).

By using nineteen DOS-DDOS features, the five metrics values of ANN, RF and SVM models are slightly the same as the baseline model.

The NB model achieved the worst detection rate ($DR_{\_NB\_(19\ features)} = 19.38\%$) and the same FAR value as the baseline model ($FAR_{\_NB\_(19\ features)} = FAR_{\_NB\_(42\ features)} = 0.01\%$).

The same performance was observed by the RF model ($FAR_{\_RF\ (19\ features)} = 27.73\% = FAR_{\_RF\_(42\ features)}=27.74\%$).

However, the classifiers KNN, SVM, ANN and RF increased the FAR value ($FAR_{\_ANN\_(19\ features)} = 37.36\% > FAR_{\_ANN\_(42\ features)} = 29.45\%$,

$FAR_{\_SVM\_(19\ features)} = 45.64\% > FAR_{\_SVM\_(42\ features)} = 40.51\%$).

Umar and Chen (2020) [57] used Best First as search strategy and DT as evaluator of their wrapper process. Based on UNSW_NB15, NSL_KDD datasets and four classifiers (ANN, SVM, KNN and RF), this process has selected twenty best DOS-DDOS features. The authors used five metrics to evaluate their models: Acc, DR, FAR and T. As results of this experiment, the RF algorithm outperformed the other used classifiers. By using the NSL_KDD dataset, the used wrapper process enhanced the accuracy and reduced the execution time ($Acc_{\_RF\_(20\ features)} = 99.7\ \% > Acc_{\_RF\_(42\ features)} = 98.8\ \%$, $T_{\_RF\_(20\ features)} = 13.2s < T_{\_RF\_(42\ features)} = 15s$). The use of UNSW_NB15 dataset and the wrapper step enhanced the RF accuracy and slightly increased the execution time due to the unnormalized data ($Acc_{\_RF\_(20\ features)} = 98.51\% > Acc_{\_RF\_(42\ features)} = 95.74\%$, $T_{\_RF\_(20\ features)} = 33.6s > T_{\_RF\_(42\ features)} = 32.4s$).

The performances of KNN, SVM and ANN were slightly lower by using twenty features, UNSW_NB15 and NSL_KDD datasets.

However, the SVM model increased drastically the execution time ($T\_{\text{SVM\_(20 features)}}$ = 10236.6s > $T\_{\text{SVM\_(42 features)}}$ = 5213.4s) by using the UNSW_NB15 and NSL_KDD datasets. The KNN and RF classifiers decreased the FAR value on the UNSW_NB15 dataset: ($FAR\_{\text{KNN (20 features)}}$ = 7.36% < $FAR\_{\text{RF\_(42 features)}}$= 11.42 %, $FAR\_{\text{RF\_(20 features)}}$ = 2.89 % < $FAR\_{\text{RF\_(42 features)}}$ = 8.77%).

However, the SVM model increased drastically the execution time ($T\_{\text{SVM\_(20 features)}}$ = 10236.6s > $T\_{\text{SVM\_(42 features)}}$ = 5213.4s) by using the UNSW_NB15 and NSL_KDD datasets. The KNN and RF classifiers decreased the FAR value on the UNSW_NB15 dataset: ($FAR\_{\text{KNN\_(20 features)}}$ = 7.36% < $FAR\_{\text{RF\_(42 features)}}$= 11.42 %, $FAR\_{\text{RF\_(20 features)}}$ = 2.89 % < $FAR\_{\text{RF\_(42 features)}}$ = 8.77%).

5) DOS-DDOS based on wrapper process and meta-heuristics search: Meta-heuristics are new optimization methods used in DOS-DDOS feature selection problems to provide near-optimal solution [34]. These methods are based on two main search strategies [58]. The first strategy is used to guarantee a global and efficient search to find a solution of DOS-DDOS feature selection. The second strategy is used to improve feature selection solutions.

Important research projects have applied meta-heuristic strategies to solve the problem of DOS-DDOS feature selection. In the paragraph below we discuss the important results of these investigations. At the end of this subsection, we present our second dashboard (Tables IIIA, IIIB, IIIC) to summarize and to compare the performances of these strategies.

TABLE III. (A): WRAPPER METHODS BASED ON META-HEURISTIC SEARCH (MHS)

DOS-DDOS feature selection projects based on wrapper methods	DOS- DDOS used dataset	Used wrapper strategies	Number of DOS -DDOS features	Used classifier	Used Metrics	Values metrics with FS	Values metrics without FS
Jun,et al. (2010)[59]	KDD'99	ABC	5	SVM	Accuracy Time (s)	99.92% 12.20	NA
Alomari and A. Othman (2012) [60]	KDD'99	BA	6	SVM	Accuracy Detection Rate FAR	93.36% 90.22% 4.56%	NA
De la Hoz et al. (2014) [61]	NSL_KDD	NGHA-II	25	GHSOM	Accuracy	99.5%	96.02%
Senthilnayaki, et al. (2015) [62]	KDD'99	GA	10	SVM	Accuracy	99.15%	82.45%
Gaikwad and Thool (2015) [63]	NSL_KDD	GA	15	Bagging (PART)	Accuracy Time (s)	99.71% 1589	NA
				PART	Accuracy Time (s)	77.79% 274	NA
				Bagging (C4.5)	Accuracy Time (s)	77.86% 1795	NA

TABLE III-(B): WRAPPER METHODS BASED ON META-HEURISTIC SEARCH (MHS)

DOS-DDOS feature selection projects based on wrapper methods	DOS- DDOS used dataset	Used wrapper strategies	Number of DOS -DDOS features	Used classifier	Used Metrics	Values metrics with FS	Values metrics without FS
Gaikwad and Thool (2015) [63]	NSL_KDD	GA	15	C4.5	Accuracy Time (s)	79.08% 176.05	NA
Wang Xingzhu (2015) [42]	KDD'99	ACO	10	SVM	Detection Rate Time(s)	97.09% 17.99	92.71% 23.51
Eesa et al. (2015) [64]	KDD'99	CFA	10	ID3	Accuracy Detection Rate FAR	92,83% 92.05% 3.9%	73,26% 71.08% 17.685%
Kang and Kim (2016) [65]	NSL_KDD	LSA- K-means	25	MLP	Accuracy Detection Rate FAR	99.37% 99.42% 0.66%	96.93% 93.38% 0.96%
Hosseinzadeh and Kabiri (2016) [66]	KDD'99	ACO	4	NN	Precision Recall F-measure	81.66% 99.78% 89.82%	87.86% 80.02% 83.76%

Khammassi and Krichen (2017) [26]	KDD’99	GA-LR	18	RF	Precision Recall	99.97% 99.98%	NA
	UNSW_NB15		20	C4.5	Precision Recall	36.09% 4.11%	NA
Enache et al. (2017) [67]	NSL_KDD	PSO	21	SVM	Detection Rate FAR	97.17% 1.6%	89.64% 6.88%
			20	NB	Detection Rate FAR	89.85% 5.34%	90.53% 6.66%
			20	C4.5	Detection Rate FAR	96.66% 2.62%	95.67% 3.02%
Yin Chunyong et al. (2017) [68]	KDD’99	ICSA	21	KNN	Accuracy FAR	99.5% 0.1%	-
Khorram and Baykan (2018) [69]	NSL_KDD	PSO	11	KNN	Accuracy Detection Rate Time (s)	96.04% 94.9% 52	93.9% 91.9% 291
				SVM	Accuracy Detection Rate Time (s)	96.02% 92.3% 309	91.4% 89.9% 722
		ACO	7	KNN	Accuracy Detection Rate Time (s)	98.13% 97.2% 67	93.9% 91.9% 291
				SVM	Accuracy Detection Rate Time (s)	95.6% 93% 142	91.4% 89.9% 722
		ABC	7	KNN	Accuracy Detection Rate Time (s)	98.9% 98.7% 53	93.9% 91.9% 291
				SVM	Accuracy Detection Rate Time (s)	97.1% 93.9% 341	91.4% 89.9% 722
	UNSW_NB15		15		Accuracy Specificity Sensitivity Time(s)	99.12% 91.76% 93.46% 1.32	85.56% NA NA NA

TABLE III-(C): WRAPPER METHODS BASED ON META-HEURISTIC SEARCH (MHS)

DOS-DDOS feature selection projects based on wrapper methods	**DOS- DDOS used dataset**	**Used wrapper strategies**	**Number of DOS -DDOS features**	**Used classifier**	**Used Metrics**	**Values metrics with FS**	**Values metrics without FS**
Mazini et al. (2019)[70]	NSL_KDD	ABC	25	AdaBoost	Accuracy Detection Rate FAR	98.90% 99.61% 0.01%	NA NA NA
Samadi Bonab et al. [58]	KDD	FFA-ALO	12	DT	Accuracy Specificity Sensitivity Time(s)	99.73% 99.67% 99.87% 2.90	97.99% NA NA NA
	NSL_KDD		16		Accuracy Specificity Sensitivity Time(s)	99.31% 97.10% 99.24% 1.50	99.31% NA NA NA
	UNSW_NB15		15		Accuracy Specificity Sensitivity Time(s)	99.12% 91.76% 93.46% 1.32	85.56% NA NA NA

As an example of relevant research projects based on wrapper process and meta-heuristic search, we can cite the important investigation of Jun Wang et al. [59]. In this study, the ABC-SVM approach was adopted as wrapper feature selection process. This wrapper strategy selected five DOS-DDOS best features from the KDD’99 dataset and found the best parameter to the SVM classifier. This method achieved good accuracy ($\text{Acc\_}_{\text{SVM\_(5 features)}}$ = 99.92%) and improved the time of execution ($\text{T\_}_{\text{SVM\_(5 features)}}$ = 12.20 s).

Alomari and Ali Othman (2012) [60] used an approach based on the Bees Algorithm (BA) as a wrapper feature method by using the classifier SVM. This experiment selected six DOS-DDOS features collected from the KDD'99 data set. They compared BA-SVM with other methods and concluded that their method achieved high detection rate and accuracy ($DR\_{}_{SVM\_(6\ features)}$ = 90.22%, $Acc\_{}_{SVM\_(6\ features)}$ = 93.36%) on detecting attacks with a low FAR ($FAR\_{}_{SVM\_(6\ features)}$ = 4.56%).

De La Hoz et al. (2014) [61] used a multi-objective procedure based on NSGA-II algorithm as wrapper feature selection to reduce the complexity of Growing Hierarchical Self-Organising Maps (GHSOM) algorithm. This wrapper method selected twenty-five representative features. As one of the multiple-objective based on the NSGA-II, the Jaccard index is evaluated after training the GHSOM. Their proposition improved the accuracy compared to the baseline model ($Acc\_(_{25\ features)}$ = 99.5% > $Acc\_(_{42\ features)}$ = 96.02%).

Senthilnayaki et al. (2015) [62] combined Genetic Algorithm (GA) with SVM. This study achieved high accuracy ($Acc_{\_(10\ features)}$ = 99.15%) with only ten best DOS-DDOS features compared to the baseline model ($Acc\_(_{42\ features)}$ = 82.45%).

Gaikwad and Thool (2015) [63] used Genetic Algorithm as wrapper feature selection which selected fifteen features. The authors used two classifiers Partial Decision Tree (PART) and C4.5, and they employed the Bagging on the two previous classifiers. This experiment has shown that using PART with the bagged classifier enhanced the accuracy and increased the execution time ($Acc\_{}_{Bagging\_PART}$ = 99.71% > $Acc\_{}_{PART}$ = 77.79%, $T\_{}_{Bagging\_PART}$ = 1589s > $T\_{}_{PART}$ = 274s). On the other side, using C4.5 with Bagging decreased the accuracy and increased drastically the execution time ($Acc\_{}_{Bagging\_C4.5}$ =77.86% < $Acc\_{}_{C4.5}$ = 79.08%, $T\_{}_{Bagging\_C4.5}$ = 1795s > $T\_{}_{C4.5}$ = 176.05s).

Wang Xingzhu (2015) [42] combined ACO feature weighting SVM. This wrapper strategy selected ten most important DOS.

DDOS features which achieved high detection rate and reduced the execution time ($DR_{\_(10\ features)}$ = 97.09% > $DR_{\_(42\ features)\ features)}$ = 92.71 %, $T_{\_(42\ features)}$ = 23.51s > $T_{\_(10\ features)}$ = 17.99s).

Eesa et al. (2015) [64] modified the Cuttle Fish Algorithm (CFA) and used it as wrapper feature selection method. They applied the classifier ID3 to detect attacks by using the KDD'99 dataset with ten best features. The process showed a real improvement of accuracy and detection rate compared to all used features ($Acc_{\_(10\ features)}$ = 92.83% > $Acc\_{}_{(42\ features)}$ = 73.26%, $DR_{\_(10\ features)}$ = 92.05% > $DR_{\_(42\ features)}$ = 71.08%). Moreover, the FAR value decreased from $FAR_{\_(42\ features)}$ = 17.68% to $FAR_{\_(10\ features)}$ = 3.9 %.

Kang and Kim (2016) [65] employed Local Search Algorithm (LSA) and K-means to find the optimal DOS-DDOS subset features, to reduce the training time and to avoid the over-fitting problem. This experiment evaluated the performance of twenty five selected DOS-DDOS features. The result has shown that using LSA-K-means as wrapper feature step with MLP enhanced the accuracy, increased the detection rate and reduced the FAR value ($Acc_{\_(25\ features)}$ = 99.37% > $Acc\_(_{42\ features)}$ = 96.93%, $DR_{\_(25\ features)}$ = 99.42% > $DR_{\_(42\ features)}$ = 93.38 %, $FAR_{\_(25\ features)}$ = 0.66% < $FAR_{\_(42\ features)}$ = 0.96%).

Hosseinzadeh Aghdam and Kabiri (2016) [66] build an intrusion detection system based on ACO (Ant Colony Optimization) feature selection method. This method converges faster to the optimal DOS-DDOS subset composed of four DOS-DDOS features. This strategy has increased the Recall and the F-measure values ($Re_{\_(4\ features)}$ = 99.78% > $Re_{\_(42\ features)}$ = 80.02 %, $FM_{\_(4\ features)}$ = 89.82% > $FM_{\_(42\ features)}$ = 83.76%). However, the precision is slightly decreased compared to the baseline model ($Pr_{\_(4\ features)}$ = 81.66 % < $Pr_{\_(42\ features)}$ = 87.86 %).

Khammassi and Krichen (2017) [26] combined Genetic Algorithm with Logistic Regression (LR) as Wrapper feature selection method. This experiment based on different decision tree classifiers (C4.5, RF, and NBTree) has maximized the accuracy by using the KDD'99 and UNSW_NB15 datasets with eighteen and twenty DOS-DDOS best features. The LR-RF strategy has achieved a high precision and Recall values ($Pr_{\_(18\ features)}$ = 99.97%, $Re_{\_(18\ features\ )}$ = 99.98%).

By using UNSW_NB15 dataset with twenty DOS-DDOS features, the LR-C4.5 process has achieved the worst Recall and precision values ($Re_{\_(20\ features)}$ = 4.11 %, $Pr\_(_{20\ features)}$ = 36.09%).

Enache et al. (2017) [67] conducted their experiment on the NSL_KDD dataset with many wrapper approaches (Algorithm (BA) ad Particle Swarm Optimization (PSO)). To evaluate these strategies they used the classifiers C4.5, SVM and BN.

The PSO-SVM process outperformed the other classifiers with only twenty-one features. It enhanced the detection rate and decreased the FAR value ($DR_{\_(21features)}$ = 97.17 > $DR_{\_(42\ features)}$ = 89.64 %, $FAR_{\_(21\ features)}$ = 1.6 % < $FAR_{\_(42\ features)}$ = 6.88 %).

By using eighteens selected features, the process BA-C4.5 achieved an interesting detection rate and increased slightly the FAR value ($DR_{\_(18\ features)}$ = 96.01 % > $DR_{\_(42\ features)}$ = 95.67 %, $FAR_{\_(18\ features)}$ = 3.20 % > $FAR_{\_(42\ features)}$ = 3.02 %).

Yin Chunyong et al. (2017) [68] used an artificial immune system as wrapper method which improved the Clonal Selection Algorithm (ICSA). This method based on the theory of biological immune system learning process selected twenty-one features from the KDD'99 dataset.

This subset realized a good accuracy and low FAR value ($Acc_{\_(21\ features)}$ = 99.5%, $FAR_{\_(21\ features)}$ = 0.1%).

Khorram and Baykan (2018) [69] tested and compared the performances of three wrapper feature selection methods by using two classifiers: SVM and KNN. The used wrapper methods are Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and Artificial Bee Colony (ABC).

This experiment showed that ABC-KNN strategy with seven features outperformed the use of all features ($T\_{}_{ABC\text{-}KNN\_(7\ features)}$ = 53s < $T\_{}_{KNN\_(42\ features)}$ = 291s, $Acc\_{}_{ABC\text{-}KNN\_(7\ features)}$ = 98.9 % > $Acc\_{}_{KNN\_(42\ features)}$ = 93.9%,

$\text{DR\_}_{\text{ABC-KNN\_(7 features)}}$ =98.7 % > $\text{DR\_}_{\text{KNN\_(42 features)}}$ = 91.9 %).

Mazini et al. (2019) [70] employed ABC as wrapper process to optimize their IDS by using NSL_KDD dataset, the classifier AdaBoost and the parameters regulation method.

This strategy selected twenty-five DOS-DDOS features and achieved a high accuracy, detection rate and low FAR values ($\text{Acc}_{\text{\_(25 features)}}$ = 98.90%, $\text{DR}_{\text{\_(25 features)}}$ = 99.61%, $\text{FAR}_{\text{\_(25 features)}}$ = 0.01%).

Samadi Bonab et al. [58] proposed an improved version of IDS based on the hybrid method Fruit-Flu algorithm (FFA) and the Lion Optimizer algorithm (ALO) as wrapper approach. This strategy based on the datasets KDD'99, NSL_KDD and UNSW_NB15 reduced the used features from 41 to 12 on KDD'99, from 41 to 16 on NSL_KDD and from 48 to 15 on UNSW_NB15. It applied the DT algorithm as a classifier on these different datasets. The performances are evaluated by using five metrics: Acc, Sp, Sen and T. This experiment has shown an enhanced accuracy and reduced the execution time on KDD'99 and UNSW_NB15 datasets ($\text{Acc\_}_{\text{KDD'99\_(12 features)}}$ = 99.73% > $\text{Acc\_}_{\text{KDD'99\_(42 features)}}$ = 97.99%, $\text{Acc}_{\text{\_UNSW\_NB15\_(15 features)}}$ = 99.12% > $\text{Acc}_{\text{\_UNSW\_NB15\_(42 features)}}$ = 85.56%). On the NSL_KDD dataset the use of this wrapper process didn't change the accuracy ($\text{Acc}_{\text{\_NSL\_KDD\_(16 features)}}$ = $\text{Acc}_{\text{\_NSL\_KDD\_(42 features)}}$ =93%). However, the specificity was lower on UNSW_NB15 and NSL_KDD compared to KDD'99 ($\text{Sp\_}_{\text{UNSW\_NB}}$ = 91.76 % < $\text{Sp\_}_{\text{NSL\_KDD}}$ = 97.10% < Sp__{KDD}= 99.67%).

The Tables IIIA, IIIB, IIIC summarize and compare the performances of all wrapper process and meta-heuristic strategies discuss above.

6) DOS-DDOS feature selection based on wrapper process and Random search methods: Random search methods applied DOS-DDOS feature selection projects to evaluate the DOS-DDOS features on random sampling around the problem region. These stochastic methods are mainly used to solve the global problem optimizations [71].

To optimize the DOS-DDOS feature subsets, many important research projects have used wrapper process and random search methods to solve this problem. We discuss these projects in the paragraph below. At the end of this subsection, we present our third dashboard (Table IV) to summarize and to compare the performances of these strategies.

As an example of these important investigations, we can cite the important study of Lin et al. (2012) [72] which combined Simulated Annealing (SA) with SVM algorithm to get the best feature subset. This experiment selected twenty three best DOS-DDOS features which evaluated by SA as random search and C4.5 decision tree as classifier. Compared to the initial set of features, the selected subset achieved a high accuracy equal to 99.96%.

Chowdhury et al. (2016) [36] used a wrapper feature selection method based on SA as random search and the ACCS dataset. This strategy selected three best features to detect attacks.

By applying the SVM algorithm with SA, this experiment has showed better accuracy, low FAR and FN values compared to all used features ($\text{Acc\_}_{\text{SVM\_(3 features)}}$ = 98.76% > $\text{Acc\_}_{\text{SVM\_(42 features)}}$ = 88.03%, $\text{FAR\_}_{\text{SVM\_(3 features)}}$ = 0.09% < $\text{FAR\_}_{\text{SVM\_(42 features)}}$ = 4.2%, $\text{FN\_}_{\text{SVM\_(3 features)}}$ = 1.15% < $\text{FN\_}_{\text{SVM\_(42 features)}}$ = 7.77 %).

Hasan Md El Mehedi et al. (2016) [73] adapted the Random Forest algorithm (RF) to select twenty-five best features by using the KDD'99 dataset. The performances evaluation is based on 3 metrics: accuracy, precision and FAR. Compared to the initial used dataset with all features, this wrapper strategy increased the accuracy, the precision and decreased the FAR value ($\text{Acc}_{\text{\_(25 features)}}$ = 91.90% > $\text{Acc}_{\text{\_(42 features)}}$ = 91.41%, $\text{Pr}_{\text{\_(25 features)}}$ = 98.94% > $\text{Pr}_{\text{\_(42 features)}}$ = 98.91%, $\text{FAR}_{\text{\_(25 features)}}$ = 5.82% < $\text{FAR}_{\text{\_(42 features)}}$ = 7.52%).

TABLE IV. WRAPPER METHOD BASED ON RANDOM METHODS

DOS-DDOS feature selection projects based on wrapper methods	DOS- DDOS used dataset	Used wrapper strategies	Number of DOS -DDOS features	Used classifier	Used Metrics	Values metrics with FS	Values metrics without FS
Lin et al. [72]	KDD'99	SA-SVM	23	SA-DT	Accuracy	99.96%	NA
Chowdhury et al. [36]	ACCS	SA	3	SVM	Accuracy FAR FN	98.76% 0.09% 1.15%	88.03% 4.2% 7.77%
Hasan Md El Mehedi et al. [73]	KDD'99	RF	25	RF	Accuracy Precision FAR	91.90% 98.94% 5.82%	91.41% 98.91% 7.52%
Najeeb and Dhannoon (2018) [74]	NSL_KDD	BFA	15	NB	Accuracy	94.83%	89.9%
Almasoudy et al. (2019) [75]	NSL_KDD	DE	9	ELM	Detection Rate Precision F_measure	91.5% 81.18% 86.03%	79.55% 94.90% 80.44%

Najeeb and Dhannoon (2018) [74] proposed an IDS model that combined the Binary Firefly (BFA) method with the Naïve Bayes (NB) classifier by using the NSL_KDD dataset. The BFA is initialized by a binary sequence contrary to the Firefly (FA) algorithm. This model was iterated two hundred times with fifteen selected features and achieved better accuracy compared to all used features ($Acc\_{(25\ features)}$ = 94.83% > $Acc\_{(42\ features)}$ = 89.9%).

Almasoudy et al. (2019) [75] has realized an IDS experiment based on Differential Evolution (DE) as wrapper based approach by using the NSL_KDD dataset. Nine candidate features are randomly selected. The Extreme Learning Machine (ELM) is used as classifier to compute the accuracy of DOS-DDOS features until it achieved high accuracy. Applied to DOS-DDOS attacks predicting, this method achieved high detection rate, high F-measure and decreased slightly the precision ($DR\_{(9\ features)}$ = 91.5% > $DR\_{(42\ features)}$ = 79.55%, $FM\_{(9\ features)}$ = 86.03 % > $FM\_{(42\ features)}$ = 80.84%, $Pr\_{(42\ features)}$ = 94.90 % > $Pr\_{(9\ features)}$ = 81.18%).

VI. Conclusion

Nowadays, cybersecurity attacks grow over time, especially the Denial of Service attack (DOS) and its variant Distributed Denial of Service (DDOS). These famous attacks continue to threaten private and public activities everywhere.

Dealing with these threats by using Machine Learning (ML) models can hold a great promise in DOS-DDOS security systems. By learning from and identifying a large amount of network traffic, these predictive models can efficiently handle the DOS-DDOS threats and overcome several limits and performance issues of the traditional security solutions.

One of the key preprocessing phases to success and optimize these DOS-DDOS cybersecurity intelligence models is feature selection step, particularly the feature selection method based on the Wrapper strategies.

Using Wrapper techniques improved significantly the selection of the relevant DOS-DDOS features and enhanced the performance of many existing ML solutions.

In this paper, we have advanced the development of this previous work by studying different DOS-DDOS datasets, algorithms and the results of several research projects. We have reviewed and evaluated the impact of many important wrapper strategies used by many existing DOS-DDOS security systems.

We have summarized the findings in three dashboards that are essential to understand the performance of three wrapper strategies commonly used in DOS-DDOS ML models: heuristic search algorithms, meta-heuristic search and random search methods.

This study shows that many wrapper strategies, algorithms, DOS-DDOS features with a relevant impact can be selected to improve the DOS-DDOS ML existing solutions.

Acknowledgment

I would like to express my sincere gratitude to my Professors and my family for the continuous support of my study and related research, for their patience, motivation, and immense knowledge.

References

[1] F. Ullah, M. Ali Babar, “Architectural Tactics for Big Data Cybersecurity Analytics Systems: A Review,” Journal of Systems and Software, 151, 81–118, 2019, doi:10.1016/j.jss.2019.01.051.

[2] R. Vishwakarma, K. Ankit Jain, “A survey of DDoS attacking techniques and defence mechanisms in the IoT network,” Telecommunication Systems: Modelling, Analysis, Design and Management, 73(1), 3–25, 2020.

[3] K.M. Prasad, D.A.R.M. Reddy, D.K.V. Rao, “DoS and DDoS Attacks: Defense, Detection and Traceback Mechanisms - A Survey,” Global Journal of Computer Science and Technology, 2014.

[4] J.-H. Cho, J.-Y. Shin, H. Lee, J.-M. Kim, G. Lee, “DDoS Prevention System Using Multi-Filtering Method,” Atlantis Press: 774–778, 2015, doi:10.2991/cmfe-15.2015.182.

[5] S. Qadir Mir, S. Quadri, “Information Availability: An Insight into the Most Important Attribute of Information Security,” Journal of Information Security, 07, 185–194, 2016, doi:10.4236/jis.2016.73014.

[6] M. Sachdeva, G. Singh, K. Saluja, K. Singh, “DDoS Incidents and their Impact: A Review,” Int. Arab J. Inf. Technol., 7, 14–20, 2010.

[7] X. Liang, T. Znati, “On the performance of intelligent techniques for intensive and stealthy DDos detection,” Computer Networks, 164, 106906, 2019, doi:10.1016/j.comnet.2019.106906.

[8] Ibrahim Salim M., T.A. Razak, “A study on IDS for preventing Denial of Service attack using outliers techniques,” in 2016 IEEE International Conference on Engineering and Technology (ICETECH), 768–775, 2016, doi:10.1109/ICETECH.2016.7569352.

[9] Y.V. Srinivasa Murthy, K. Harish, V. Varma, K. Sriram, B. Revanth, “Hybrid Intelligent Intrusion Detection System using Bayesian and Genetic Algorithm (BAGA): Comparitive Study,” International Journal of Computer Applications, 99, 1–8, 2014, doi:10.5120/17342-7808.

[10] O. Salem, M. HOTTE, Q.-E. LUTTIN, T. ASCOET, Protection contre les attaques de déni de service dans les réseaux IP, Paris Descarte IUT, ECTEI: 31, 2015.

[11] J. Jang-Jaccard, S. Nepal, “A survey of emerging threats in cybersecurity,” Journal of Computer and System Sciences, 80(5), 973–993, 2014, doi:10.1016/j.jcss.2014.02.005.

[12] K.R. Bandara, T. Abeysinghe, A. Hijaz, D. Darshana, H. Aneez, S.J. Kaluarachchi, K.D. Sulochana, M. DhishanDhammearatchi, “Preventing DDoS attack using Data mining Algorithms,” International Journal of Scientific and Research Publications, 6(10), 390–400, 2016.

[13] L. Gnanaprasanambikai, N. Munusamy, “Data Pre-Processing and Classification for Traffic Anomaly Intrusion Detection Using NSLKDD Dataset,” Cybernetics and Information Technologies, 18, 2018, doi:10.2478/cait-2018-0042.

[14] S.X. Wu, W. Banzhaf, “The use of computational intelligence in intrusion detection systems: A review,” Applied Soft Computing, 10(1), 1–35, 2010, doi:10.1016/j.asoc.2009.06.019.

[15] A. Alazab, M. Hobbs, J. Abawajy, M. Alazab, “Using feature selection for intrusion detection system,” in 2012 International Symposium on Communications and Information Technologies (ISCIT), IEEE, Gold Coast, Australia: 296–301, 2012, doi:10.1109/ISCIT.2012.6380910.

[16] V.O. Ferreira, V.V. Galhardi, L.B.L. Gonçalves, R.C. Silva, A.M. Cansian, “A model for anomaly classification in intrusion detection systems,” Journal of Physics: Conference Series, 633, 4, 2015, doi:10.1088/1742-6596/633/1/012124.

[17] M. Bataghva, “Efficiency and Accuracy Enhancement of Intrusion Detection System Using Feature Selection and Cross-layer Mechanism,” Electronic Thesis and Dissertation Repository, 2017.

[18] I.H. Sarker, A.S.M. Kayes, S. Badsha, H. Alqahtani, P. Watters, A. Ng, “Cybersecurity data science: an overview from machine learning perspective,” Journal of Big Data, 7(1), 41, 2020, doi:10.1186/s40537-020-00318-5.

[19] J.B. Fraley, J. Cannady, “The promise of machine learning in cybersecurity,” in SoutheastCon 2017, 1–6, 2017, doi:10.1109/SECON.2017.7925283.

[20] F. Ullah, M. Ali Babar, "Architectural Tactics for Big Data Cybersecurity Analytic Systems: A Review," Journal of Systems and Software, 151, 2018, doi:10.1016/j.jss.2019.01.051.

[21] S. Sambangi, L. Gondi, "A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression," Proceedings, 63(1), 51, 2020, doi:10.3390/proceedings2020063051.

[22] R. Panthong, A. Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm," Procedia Computer Science, 72, 162–169, 2015, doi:10.1016/j.procs.2015.12.117.

[23] N. Bindra, M. Sood, "Evaluating the Impact of Feature Selection Methods on the Performance of the Machine Learning Models in Detecting DDoS Attacks," Romanian Journal of Information Science and Technology, 3, 250–261, 2020.

[24] M. Joshi, T.H. Hadi, "A Review of Network Traffic Analysis and Prediction Techniques," Network Traffic Analysis and Prediction, 23, 2015.

[25] Z. Foroushani, Y. Li, "Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique," in ICSCA 2018: Proceedings of the 2018 7th International Conference on Software and Computer Applications, Kuantan, Malaysia: 119–123, 2018, doi:10.1145/3185089.3185114.

[26] C. Khammassi, S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," Computers & Security, 70, 255–277, 2017, doi:10.1016/j.cose.2017.06.005.

[27] V. Bolón-Canedo, N.S. Maroño, A. Alonso-Betanzos, Feature Selection for High-Dimensional Data, Springer International Publishing, 2015, doi:10.1007/978-3-319-21858-8.

[28] F. Amiri, "Mutual information-based feature selection for intrusion detection systems," Journal of Network and Computer Applications, 34(4), 1184–1199, 2011, doi:10.1016/j.jnca.2011.01.002.

[29] V. Bachu, J. Anuradha, "A Review of Feature Selection and Its Methods," Cybernetics and Information Technologies, 19, 3, 2019, doi:10.2478/cait-2019-0001.

[30] V. Kumar, S. Minz, "Feature selection: A literature review," Smart Computing Review, 4, 211–229, 2014, doi:10.1145/2740070.2626320.

[31] S. Alabdulwahab, B. Moon, "Feature Selection Methods Simultaneously Improve the Detection Accuracy and Model Building Time of Machine Learning Classifiers," Symmetry, 12(9), 1424, 2020, doi:10.3390/sym12091424.

[32] S. Dwivedi, M. Vardhan, S. Tripathi, "Defense against distributed DoS attack detection by using intelligent evolutionary algorithm," International Journal of Computers and Applications, 1–11, 2020, doi:10.1080/1206212X.2020.1720951.

[33] K. Yan, D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," Sensors and Actuators B: Chemical, 212, 353–363, 2015, doi:10.1016/j.snb.2015.02.025.

[34] N. Mlambo, W. Cheruiyot, M.W. Kimwele, "A Survey and Comparative Study of Filter and Wrapper Feature Selection Techniques," The International Journal Of Engineering And Science, 5(10), 57–67, 2016.

[35] N. Moustafa, J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 Military Communications and Information Systems Conference, 1–6, 2015, doi:10.1109/MilCIS.2015.7348942.

[36] M.N. Chowdhury, K. Ferens, M. Ferens, "Network Intrusion Detection Using Machine Learning," in Computer Science, CSREA Press: 30–35, 2016.

[37] M.E. Elhamahmy, H.N. Elmahdy, I.A. Saroit, "A New Approach for Evaluating Intrusion Detection System," in CiiT International Journal of Artificial Intelligent Systems and Machine Learning, 290–298, 2010.

[38] Kamarularifin Abd Jalil, Muhammad Hilmi Kamarudin, Mohamad Noorman Masrek, "Comparison of Machine Learning algorithms performance in detecting network intrusion," in 2010 International Conference on Networking and Information Technology, 221–226, 2010, doi:10.1109/ICNIT.2010.5508526.

[39] M. Belouch, S. El Hadaj, M. Idhammad, "Performance evaluation of intrusion detection based on machine learning using Apache Spark," Procedia Computer Science, 127, 1–6, 2018, doi:10.1016/j.procs.2018.01.091.

[40] V.D. Katkar, S.V. Kulkarni, "Experiments on detection of Denial of Service attacks using ensemble of classifiers," in 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 837–842, 2013, doi:10.1109/ICGCE.2013.6823550.

[41] M.H. Bhuyan, D.K. Bhattacharyya, J.K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," IEEE Communications Surveys Tutorials, 16(1), 303–336, 2014, doi:10.1109/SURV.2013.052213.00046.

[42] W. Xingzhu, "ACO and SVM Selection Feature Weighting of Network Intrusion Detection Method," International Journal of Security and Its Applications, 9(4), 259–270, 2015, doi:10.14257/ijsia.2015.9.4.24.

[43] J.J. Lu, M. Zhang, Heuristic Search, Springer, New York, NY: 885–886, 2013, doi:10.1007/978-1-4419-9863-7_875.

[44] B. Kavitha, S. Karthikeyan, B. Chitra, Efficient Intrusion Detection with Reduced Dimension Using Data Mining Classification Methods and Their Performance Comparison, Springer Berlin Heidelberg, Berlin, Heidelberg: 96–101, 2010, doi:10.1007/978-3-642-12214-9_17.

[45] M.S. Mok, S.Y. Sohn, Y.H. Ju, "Random Effects Logistic Regression Model for Anomaly Detection," Expert Syst. Appl., 37(10), 7162–7166, 2010, doi:10.1016/j.eswa.2010.04.017.

[46] I. Ahmad, A. Abdullah, A. Alghamdi, M. Hussain, K. Nafjan, "Intrusion Detection Using Feature Subset Selection based on MLP," Scientific Research and Essays, 6(34), 6804–6810, 2011, doi:10.5897/SRE11.142.

[47] L. Yinhui, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," Expert Systems with Applications, 39, 424–430, 2012, doi:10.1016/j.eswa.2011.07.032.

[48] F. Zhang, D. Wang, "An Effective Feature Selection Approach for Network Intrusion Detection," in 2013 IEEE Eighth International Conference on Networking, Architecture and Storage, 307–311, 2013, doi:10.1109/NAS.2013.49.

[49] O.Y. Al-Jarrah, A. Siddiqui, M. Elsalamouny, P.D. Yoo, S. Muhaidat, K. Kim, "Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection," in 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW), 177–181, 2014, doi:10.1109/ICDCSW.2014.14.

[50] J. Lee, D. Park, C. Lee, "Feature Selection Algorithm for Intrusions Detection System using Sequential Forward Search and Random Forest Classifier," KSII Transactions on Internet and Information Systems, 11(10), 5132–5148, 2017.

[51] B.S. Harish, N. Manju, "Hybrid Feature Selection Method Using Fisher's Discriminate Ratio to Classify Internet Traffic Data," in Proceedings of the 4th International Conference on Frontiers of Educational Technologies, ACM, New York, NY, USA: 75–79, 2018, doi:10.1145/3233347.3233369.

[52] H. Soodeh, A. Mehrdad, "The hybrid technique for DDoS detection with supervised learning algorithms," Computer Networks, 158, 35–45, 2019, doi:10.1016/j.comnet.2019.04.027.

[53] H. Malhotra, P. Sharma, "Intrusion Detection using Machine Learning and Feature Selection," International Journal of Computer Network and Information Security, 11(4), 43–52, 2019, doi:10.5815/ijcnis.2019.04.06.

[54] M. Wang, Y. Lu, J. Qin, "A dynamic MLP-based DDoS attack detection method using feature selection and feedback," Computers & Security, 88, 101645, 2020, doi:10.1016/j.cose.2019.101645.

[55] H. Polat, O. Polat, A. Cetin, "Detecting DDoS Attacks in Software-Defined Networks Through Feature Selection Methods and Machine Learning Models," Sustainability, 12(3), 1–16, 2020.

[56] M.A. Umar, C. Zhanfang, Y. Liu, "Network Intrusion Detection Using Wrapper-based Decision Tree for Feature Selection," in Proceedings of the 2020 International Conference on Internet Computing for Science and Engineering, ACM, Male Maldives: 5–13, 2020, doi:10.1145/3424311.3424330.

[57] M.A. Umar, Z. Chen, Effects of Feature Selection and Normalization on Network Intrusion Detection, 2020, doi:10.36227/techrxiv.12480425.

[58] M. Samadi Bonab, A. Ghaffari, F. Soleimanian Gharehchopogh, P. Alemi, “A wrapper‐based feature selection for improving performance of intrusion detection systems,” International Journal of Communication Systems, 33, 2020, doi:10.1002/dac.4434.

[59] W. Jun, L. Taihang, R. Rongrong, “A real time IDSs based on artificial Bee Colony-support vector machine algorithm,” Suzhou, Jiangsu, China: 91–96, 2010, doi:10.1109/IWACI.2010.5585107.

[60] O. Alomari, Z. Ali Othman, “Bees Algorithm for feature selection in Network Anomaly detection,” 8(3), 1748–1756, 2012.

[61] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, A. Martínez-Álvarez, “Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps,” Knowledge-Based Systems, 71, 322–338, 2014, doi:10.1016/j.knosys.2014.08.013.

[62] B. Senthilnayaki, K. Venkatalakshmi, A. Kannan, “Intrusion detection using optimal genetic feature selection and SVM based classifier,” in 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 1–4, 2015, doi:10.1109/ICSCN.2015.7219890.

[63] D.P. Gaikwad, R.C. Thool, “Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier,” Procedia Computer Science, 49, 92–98, 2015, doi:10.1016/j.procs.2015.04.231.

[64] A.S. Eesa, Z. Orman, A.M.A. Brifcani, “A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,” Expert Systems with Applications, 42(5), 2670–2679, 2015, doi:10.1016/j.eswa.2014.11.009.

[65] S.-H. Kang, K.J. Kim, “A feature selection approach to find optimal feature subsets for the network intrusion detection system,” Cluster Computing, 19(1), 325–333, 2016, doi:10.1007/s10586-015-0527-8.

[66] M. Hosseinzadeh Aghdam, P. Kabiri, “Feature Selection for Intrusion Detection System Using Ant Colony Optimization,” International Journal of Network Security, 18, 420–432, 2016.

[67] A. Enache, V. Sgârciu, M. Togan, “Comparative Study on Feature Selection Methods Rooted in Swarm Intelligence for Intrusion Detection,” in 2017 21st International Conference on Control Systems and Computer Science (CSCS), 239–244, 2017, doi:10.1109/CSCS.2017.40.

[68] C. Yin, L. Ma, L. Feng, “Towards accurate intrusion detection based on improved clonal selection algorithm,” Multimedia Tools and Applications, 76(19), 19397–19410, 2017, doi:10.1007/s11042-015-3117-0.

[69] T. Khorram, N. Baykan, “Feature selection in network intrusion detection using metaheuristic algorithms,” International Journal Of Advance Research, Ideas and Innovations in Technology, 4(4), 704–710, 2018.

[70] M. Mazini, B. Shirazi, I. Mahdavi, “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms,” Journal of King Saud University - Computer and Information Sciences, 31(4), 541–553, 2019, doi:10.1016/j.jksuci.2018.03.011.

[71] H.E. Romeijn, Random search methods, Springer US, Boston, MA: 3245–3251, 2009, doi:10.1007/978-0-387-74759-0_556.

[72] S.-W. Lin, K. Ying, C. Lee, Z.-J. Lee, “An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection,” Appl. Soft Comput., 12(10), 3285–3290, 2012, doi:10.1016/j.asoc.2012.05.004.

[73] M.A.M. Hasan, M. Nasser, S. Ahmad, K.I. Molla, “Feature Selection for Intrusion Detection Using Random Forest,” Journal of Information Security, 7(3), 129–140, 2016, doi:10.4236/jis.2016.73009.

[74] R.F. Najeeb, B.N. Dhannoon, “Improving Detection Rate of the Network Intrusion Detection System Based on Wrapper Feature Selection Approach,” Iraqi Journal of Science, 59(1.B), 426–433, 2018, doi:10.24996/ijs.2018.59.1B.23.

[75] F. Almasoudy, W. Al-Yaseen, A. Idrees, “Differential Evolution Wrapper Feature Selection for Intrusion Detection System,” Procedia Computer Science, 167, 1230–1239, 2019, doi:10.1016/j.procs.2020.03.438.

Intelligent Data Aggregation Framework for Resource Constrained Remote Internet of Things Applications

Abhijith H V[1]
Department of Information Science and Engineering
Sai Vidya Institute of Technology, Affiliated to Visvesvarya
Technological University, Belagavi, Bangalore, India

Dr. H S Ramesh Babu[2]
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, Affiliated to Visvesvarya
Technological University, Belagavi, Bangalore, India

***Abstract*—Internet of Things (IoT) is a technology that can connect everything to the Internet. IoT can be used in a wide range of applications which includes remote applications like Underwater networks. Remote applications involve the deployment of several low-power, low-cost interconnected sensor nodes in the specific region. With a massive amount of devices connected to the IoT and the considerable amount of data associated with it, there remain concerns about data management. Also, the amount of data generated in an extensive IoT-based remote sensing network is usually enormous for the servers to process, and many times data generated are redundant. Hence there is a need for designing a framework that addresses both aggregations of data and security-related issues at various aggregation points. In this paper, we are proposing an intelligent data aggregation mechanism for IoT-based remote sensing networks. This method avoids redundant data transmission by adapting spatial aggregation techniques. The proposed method was tested through simulations, and the results prove the efficiency of the proposed work.**

Keywords—Wireless sensor networks; Internet of Things; intelligent boundary determination; sensor nodes; data aggregation

I. Introduction

Internet of Things refers to connecting various things to the internet. IoT can be used in a wide range of applications. IoT applications include remote applications also. Remote IoT applications include military applications such as surveillance, battlefield monitoring, underwater applications such as oil spill detection, analysis of aquatic animals lifecycle, smart farming, forest applications such as fire monitoring, etc. Remote IoT applications use resource-constrained networks. It includes deploying several low-cost and low-power sensor nodes that can sense, process, and communicate the data. Each sensor node has a limited transmission range, limited energy, limited processing capabilities, and limited memory.

As sensor nodes are resource-constrained devices, several sensor nodes will be deployed in a smaller region. This leads to the generation of redundant data. Transmitting redundant data will leads to the wastage of energy and other resources. Redundant data transmission can be controlled by data aggregation techniques. There are different data aggregation techniques such as In-network data aggregation, Tree-based data aggregation, cluster-based data aggregation, Grid-based data aggregation, and hybrid data aggregation. There can be single-level or multiple levels of data aggregation. Data aggregation can be performed in homogeneous networks or in heterogeneous networks. Homogeneous networks include nodes with similar configurations and capabilities in the entire network. Heterogeneous networks include nodes with different configurations and capabilities organized in the multiple levels of hierarchy [15][17].

Data aggregation involves just combining the data from multiple sources into a single packet and forwarding it to the next node or destination. Traditional approaches always aggregate the data based on the count, average, sum, etc. The reduction of redundant data is not that significant through normal data aggregation techniques. Hence there is a need for adding some intelligence to nodes to decide whether to transmit the data or not [16]. In the constrained remote IoT applications where we have limited resources, communication consumes more energy than compared to processing. In this paper, we are proposing an intelligent data aggregation technique, which aggregates the data as well as provides intelligence to nodes to detect whether to send the data or not. Also, our present work adds features to detect the boundary of certain event occurrences in event-driven networks.

The paper is organized as follows: Related work is given in Section 2, Section 3 describes proposed work, Section 4 describes the simulation and results, the conclusion is provided in Section5.

II. Related Work

There are different traditional data aggregation mechanisms; they are flat-based [1][2][3] and hierarchical-based techniques.

In a flat network, the base station transmits a query message requesting data from the sensor nodes within the network. The nodes which have data relevant to the query sent will respond with the requested data. In this method, the base station performs excessive communications and computations. Because of this, if the base station fails, then the network connection will be lost with the outer world.

Under hierarchical data aggregation, many approaches have been proposed for energy efficiency and scalability. There exist four different types of hierarchical data aggregation they are the cluster-based data aggregation [4][5], tree-based data aggregation [9], grid-based data aggregation [8]., and chain-based data aggregation [6][7].

These traditional data aggregation schemes will just combine the data from multiple sources and forward it to the next node. In the majority of applications, normal data aggregation is not sufficient. We need to add some intelligence to the data aggregation and data transmission process within the limits of constrained resources.

J. Chen, S. Kher, and A. Somani, et al. proposed a scheme in [9], which involves Majority Voting. This approach is a data outlier detection method based on spatial correlation. Here if a reading of the local sensor node is different from the majority of its neighboring nodes, then it will be classified as abnormal.

Y. Sun, H. Luo, and S. K. Das et al. Proposed a scheme in [10]. Here according to the trustworthiness ranked by comparison with historical data and neighbor data, Weight will be assigned to every sensor data. Then weighted mean value will be calculated at the aggregator. This will be considered as the aggregated data.

S. Din, A. Ahmad, et al. Proposed a scheme in [11], where the nodes closer to the sink node performs direct communication and remains unclustered; the nodes which are one-hop away from the sink node perform multi-hop communication and remains clustered.

Bo Yin et al. in [12] specified a Tree-based scheme that involves the construction of an aggregation tree for complex queries. This ensures minimum communication cost.

Xiong Li et al. proposed a scheme in [13] where there will be three participants, they are edge server, terminal device, and public cloud center. The data from the terminal devices is encrypted and communicated to the edge server, then the edge server performs data aggregation of the data from terminal devices and forwards the aggregated data to the public cloud center. The aggregated plaintext data can be recovered by public cloud center through its private key.

S. Kumar and V. K. Chaurasiya et al. in [14] proposes a scheme, where data mining techniques are used to generate more accurate, consistent, and useful information than that generated by any individual sensor node.

III. Proposed Work

A. Assumptions and Architectural Setup

Following are the assumptions made in deployment of nodes in the WSN test bench.

1) Grid based deployment of nodes: nodes are deployed on a crisscross grid, to specifically determine location of each individual node. This assumption is both safe and valid as, there is a commercially available underwater robot capable of deploying nodes at specific coordinates of latitude and longitude. Such a deployment is also helpful in making intelligent inferences based on which nodes are communicating sensed data.

2) Number of nodes per grid is fixed: this assumption helps us compute statistically significant results after data collection is done and analysis is to be performed.

3) Node density is uniform: non-uniform node deployment does not allow implementation of intelligence.

4) Each node in a grid represents a single unit of area: data collected is representative of a fixed area under surveillance.

5) Unit squares (Grids) form a Level -1 cell: this logical rule imposed on grid helps in aggregation of data across levels of nodal deployment.

6) Every level-1 cell has a level-1 cell aggregator node: this is an architectural requirement of WSNs. aggregator nodes act as data aggregator for sensed data before being forwarded to upper layers.

7) Four level-1 cells combine to form a level 2-cell: multiple layers in the architecture will always help in implementation of an additional second layer of intelligence and additional aggregation.

8) One of the unit cell's cell-sink also works as a sink for four adjacent level-2 cells: this is strictly based on energy levels of nodes, suitable cell-sink is elected as the sink of adjacent cells based solely on this criteria.

9) All level-2 cells combine to form the region under observation: this is the highest level of aggregation implemented under our framework.

B. Node Deployment Phase and Network Establishment

1) Four different levels of nodes are deployed.

a) Ground Level Sensor Nodes: They are low power sensor nodes with capability to sense and forward the data to next level.

b) Grid Head: They are high power nodes. They can aggregate the data and also sense the information.

c) Level -1 Aggregator: They act like second level grid head. They gather the data from multiple grids and transfer it to Level-2 Aggregator after processing.

d) Level-2 Aggregator: They are at the top level of hierarchy, they can provide minor conclusions about the sensed information.

2) These nodes are deployed at the required locations.

3) The area of interest is divided into equal sized grids. Ground level nodes and Grid heads are deployed at each grid. High power node will be the grid head of that particular grid.

4) Each High power node broadcast a beacon message to the ground level sensor node to indicate its presence in the grid.

5) Each ground level nodes respond back to their grid heads by sending the beacon response. Same Procedure is followed between grid heads and anchor nodes, Anchor nodes and surface buoyant nodes.

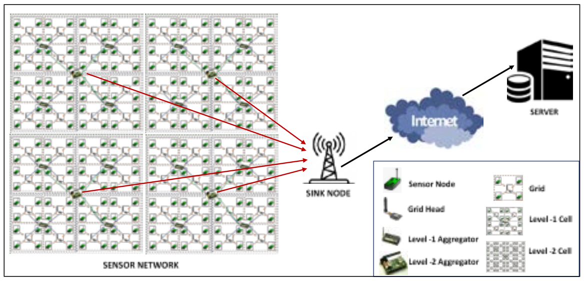


Fig. 1. Network Architecture.

Fig. 1 shows the proposed network architecture. The network is divided into smaller grid. Grids are combined to form level-1 Cell. Level-1 Cells are combined to form level -2 Cells. There will be sink node at the border of area of interest. Sink node forwards the data to server through internet connection.

C. *Data Aggregation*

Tier- 1Local aggregation (level 1):

Spatial aggregation: this primary level of aggregation is based on the percentage of area generating similar readings. Each grid has a designated grid head. The grid head communicates reading only if number of nodes reporting similar readings are greater than threshold percentage.

Spatial Aggregation is adapted in Grid as well as level-1 Cell by Grid head and Level-1 Aggregator.

Tier-2 Aggregation (level 2)

TABLE I. SYMBOLS USED IN THE ALGORITHM

Symbol	Description
S	Ground Level Sensor Node
G	Grid Head
L1	Level 1 Aggregator
L2	Level 2 Aggregator
SI	Sink Node
Ts	Threshold Percentage used by G for Level 1 Aggregation
Tg	Threshold Percentage used by L1 for Level 1 Aggregation
R1	Threshold Radius 1 for level 2 Aggregation
R2	Threshold Radius 2 for level 2 Aggregation
t_i	Time interval, i=1,2,3….
Ns	No. of Sensor Nodes which has sent similar readings
Ng	No. of Grid Heads which has sent similar readings
Ps	Packet sent by sensor node
Pg	Packet sent by Grid Head
PL1	Packet sent by Level-1 Aggregator
PL2	Packet sent by Level-2 Aggregator
D	Distance between L2 and L1
Dist	List of D computed in 1 time interval

Boundary value aggregation: region under observation has a Level -2 Aggregators that acts as a data aggregator. Radial survey is made periodically at three lengths of radii- minimum, nominal and maximum. Minimum is close to the Level -2 Aggregators, maximum is distance from Level -2 Aggregators to boundary of region under observation, and nominal is an intermediate distance.

Algorithm 1 shows how sensor senses and sends the data. Algorithm 2 and 3 depicts the level-1 aggregation techniques (Spatial aggregation). Level -2 aggregation is given in Algorithm 4. Table I describes the symbol used in the algorithm.

Algorithm 1: Sensing at Ground Level

```
1. For each ti i=1,2,3….
2.   S senses the data
3.   Prepare the packet Ps
4.   Send Ps to its G
5. End
```

Algorithm 2: Level-1 Aggregation at Grid Level

```
1. For each ti, i=1,2,3…..
2.   G: Compare the Ps received by all S under G and compute Ns
3.   If Ns >= Ts
4.       Aggregate the data
5.       Prepare the packet Pg
6.       Send Pg to L1
7.   Else
8.       Discard the data
9. End
```

Algorithm 3: Level-1 Aggregation at Level-1 Cell Level

```
1. For each ti, i=1,2,3…..
2.   L1: Compare the Pg received by all G under L1 and compute Ng
3.   If Ng >= Tg
4.       Aggregate the data
5.       Prepare the packet PL1
6.       Send PL1 to L2
7.   Else
8.       Discard the data
9. End
```

Algorithm 4: Level 2 Aggregation at Level-2 Cell level

```
Consider coordinates of L2: (XL2 , YL2) as centre point.
Let (XL1, YL2) be the coordinates of L1
For each L1 which has sent data to L2
    Find distance between the L1 and L2
    $D = Sqrt( (XL2 – XL1 )2 – (YL2 – YL1 )2 )$
    Add D to list Dist
End
Find largest D in Dist
If largest Di is within radius R1
    Radial_survey = minimum: Event spread in shorter region
    Prepare PL2 with a field Radial_survey
    Send PL2 to Sink
Else if Largest Di is between radius R1 and R2
    Radial_survey = nominal: Event spread in Moderate region
    Prepare PL2 with a field Radial_survey
    Send PL2 to Sink
Else if Largest Di is greater than R2
    Radial_survey = maximum: Event spread in longer region
    Prepare PL2 with a field Radial_survey
    Send PL2 to Sink
```

IV. Simulation and Results

In the proposed research work, we have used MATLAB as the simulation tool to check the efficiency of our proposed work. Initially, we considered the area of the application region with dimensions 800m (length) x 800m (breadth). Then the area under consideration is divided into equal-sized grids of size 100m each at ground level. Table II shows the various parameters used in the simulation. Fig. 2 depicts the topology created through simulation. Here blue color nodes represent sensor nodes deployed at ground level. Red color nodes represent the grid head; the magenta color nodes indicate nodes that are Level-1 Aggregators, Black color represents Level-2 Aggregators, and Green color indicates sink node.

Fig. 3 shows the amount of time required to detect the boundary of the event spread through Level-2 aggregation. In certain remote IoT applications detecting the spreading rate of a certain event is very important to take immediate measures. For Ex: the spread of fire in the forest, oil leakage in the ocean, etc.

Fig. 4 helps us to understand that the number of redundant transmissions in various techniques compared with our work. It is evident that redundant data transmission is significantly dropped compared to other traditional approaches. Our method requires a lesser number of packet transmissions compared to without aggregation deployment, traditional in-network architecture, and cluster-based approach. It is clear that the proposed work reduces redundant data transmission significantly.

TABLE II. Simulation Parameters

Parameter Type	**Parameter Value**
Area of application region	800 X 800 m^2
Grid range	100 m
Number of Ground level node in each grid	4
Number of High power nodes in each grids	2 (1Grid Head)
Number of Level-1 Aggregators in each group of Level-1 Cell	1
Number of Level-2 Aggregators in each group of Level-2 Cell	1
Simulation Time	150sec
Payload length	512 Bytes
Parameters of Ground level sensor node	
Initial Energy	4J
Transmission range	20m
Data rate	4kbps
Parameters of High power node/Grid Head	
Initial Energy	7J
Transmission range	25m
Data rate	6kbps
Parameters of Level-1 Aggregator	
Initial Energy	10J
Transmission range	30m
Data rate	8kbps
Parameters of Level-2 Aggregator	
Initial Energy	14J
Transmission range	35m
Data rate	12kbps

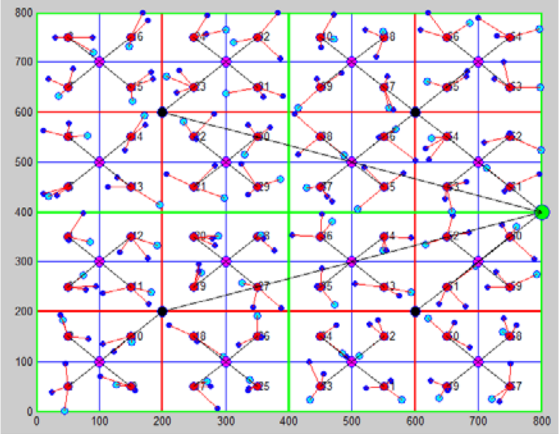


Fig. 2. Topology Setup.

Fig. 5 shows the number of the packet transmitted to sink. In this result also it is clear that our proposed method is better compared to others. The proposed method involves very little data transmission.

Fig. 6 shows the comparison of the proposed method with other traditional methods in terms of energy consumption. Result clearly shows that the proposed method consumes lesser energy compared to other techniques.

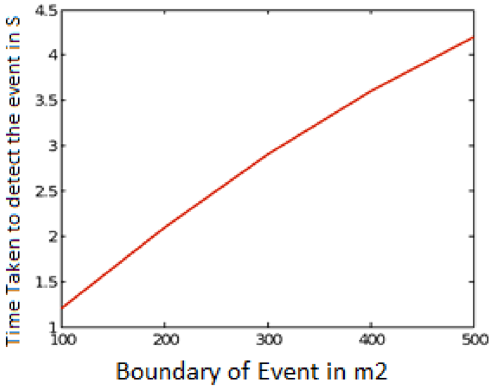


Fig. 3. Time Required for Detecting the Boundary of Event.

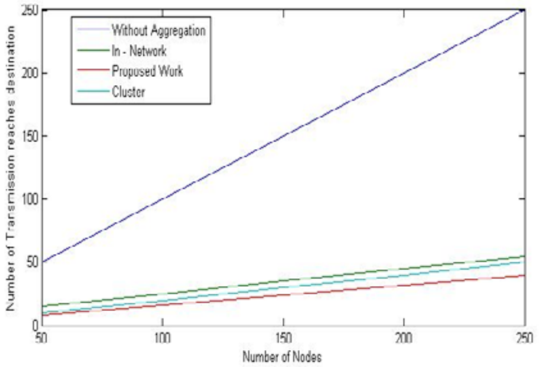


Fig. 4. Number of Redundant Transmission in Various Approaches.

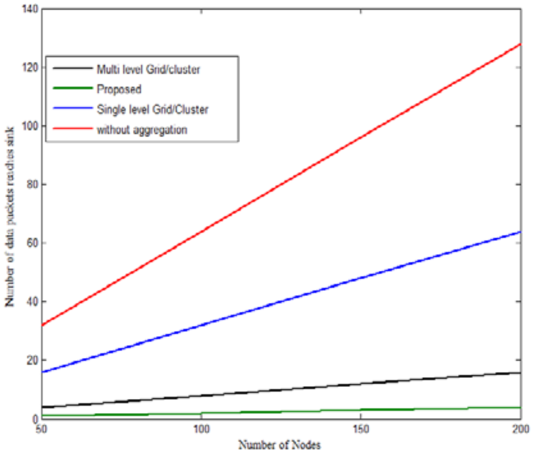


Fig. 5. Number of the Data Packets Transmitted to the Sink.

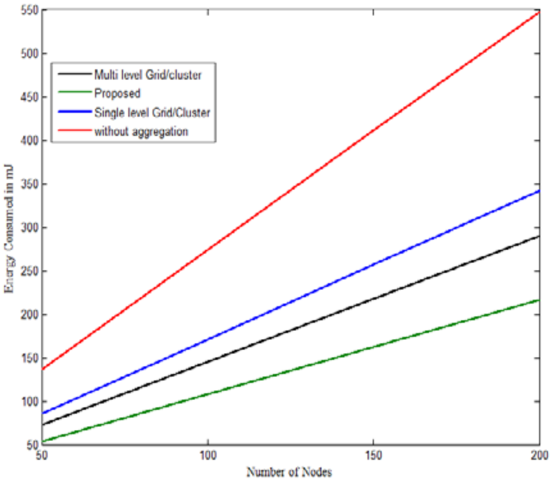


Fig. 6. Energy Consumed Per Round in Various Approaches.

V. Conclusions

Controlling redundant data transmission is a significant challenge in Constrained IoT Remote applications. In this proposed research work, Data Aggregation is addressed through spatial and boundary value aggregation, which is the novel scheme. The aggregation technique adapted reduces the number of redundant transmissions significantly. The efficiency of the proposed approach is compared with existing methods through simulation.

References

[1] J.Kulik.W.R.Heinzlman & H.Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks", WirelessNetwork, Vol 8, Mar-2002, pp.169-185.

[2] C.Intanagonwiwat,R.Govindan and D.Estrin, "Directed Diffusion: A Scalable and robust communication paradigm for sensor networks", 6th annual international conference on mobile-computing and networking, Aug-2000.

[3] B.Krishnamachari and J.Heidemann, "Application specific modeling of information routing in wireless sensor networks", IEEE international performance, computing, communications conference, Vol-23, pp.717-722, 2004.

[4] W.R Heinzelman, "Application-specific protocol architectures for wireless networks", PhD-Thesis, MIT, June-2000.

[5] O.Younis and S.Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks" IEEE transactions on mobile computing, Vol-3, Dec-2004,pp.366-79.

[6] Lindsey.S, Raghavendra.C, "PEGASIS Power-efficient Gathering in sensor Information system", IEEE Aerospace Conference 2002, pp-1125- 1130.

[7] Kuong-Ho-Chen, Jyh-Ming-Huang, Chieh-Chuan-Hsiao "CHIRON: An Energy-Efficient Chain-Based Hierarchical Routing Protocol in Wireless Sensor Networks", IEEE. 2009.

[8] Liyang-Yu, Neng-Wang, Wei-Zhang, Chunlei-Zheng, "GROUP: a Gridclustering Routing Protocol for Wireless Sensor Networks", 2006 IEEE International conference on wireless-communication networking & mobile-computing, china 2006.

[9] J. Chen, S. Kher, and A. Somani, "Distributed Fault Detection of Wireless Sensor Networks" ACM Workshop on Dependability Issues in Wireless Ad-hoc Sensor Netw. 2017.

[10] Y. Sun, H. Luo, and S. K. Das, "A Trust-based Framework for FaultTolerant Data Aggregation in Wireless Multimedia Sensor Networks," IEEE Trans. Depend. Sec. Comput., 2017.

[11] S. Din, A. Ahmad, A. Paul, M. M. U. Rathore, and J. Gwanggil, ``A cluster based data fusion technique to analyze big data in wireless multi-sensor system," IEEE Access, vol. 5, pp. 50690 5083, 2017.

[12] Bo Yin et. all. “Communication-Efficient Data Aggregation Tree Construction for Complex Queries in IoT Applications”, IEEE INTERNET OF THINGS JOURNAL, 2018 .

[13] Xiong Li et. all. “Privacy Preserving Data Aggregation Scheme for Mobile Edge Computing Assisted IoT Applications”, IEEE INTERNET OF THINGS JOURNAL, 2018.

[14] S. Kumar and V. K. Chaurasiya, "A Strategy for Elimination of Data Redundancy in Internet of Things (IoT) Based Wireless Sensor Network (WSN)," in IEEE Systems Journal 2018.

[15] Abhijith H V, & Sindhu M P. (2015). Energy efficient multilevel hierarchical data aggregation mechanism for wireless sensor networks. 2015 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2015.7154688.

[16] H.V., Abhijith and Raj, S. Deepak and Babu, H. S. Ramesh, Intelligent Boundary Determination of Oil Spill Detection Using IOT (April 21, 2018). Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018 held at Malaviya National Institute of Technology, Jaipur (India) on March 26-27, 2018, Available at SSRN: https://ssrn.com/abstract=3167315 or http://dx.doi.org/10.2139/ssrn.3167315.

[17] F. Al-Doghman, Z. Chaczko and J. Jiang, "A Review of Aggregation Algorithms for the Internet of Things," 2017 25th International Conference on Systems Engineering (ICSEng), Las Vegas, NV, 2017, pp. 480-487. doi: 10.1109/ICSEng.2017.43.

Investigative Study of the Effect of Various Activation Functions with Stacked Autoencoder for Dimension Reduction of NIDS using SVM

Nirmalajyothi Narisetty[1] ,Gangadhara Rao Kancherla[2]
Basaveswararao Bobba[3]
Dept. of CSE, Acharya Nagarjuna University
Guntur, 522510, India

K.Swathi[4]
Dept. of CSE
NRI Institute of Technology
Agiripalli, India

Abstract*—Deep learning is one of the most remarkable artificial intelligence trends. It remains behind numerous recent achievements in various domains, such as speech processing, and computer vision, to mention a few. Likewise, these achievements have sparked great attention in utilizing deep learning for dimension reduction. It is known that the deep learning algorithms built on neural networks contain number of hidden layers, activation function and optimizer, which make the computation of deep neural network challenging and, sometimes, complex. The reason for this complexity is that obtaining an outstanding and consistent result from such deep architecture requires identifying number of hidden layers and suitable activation function for dimension reduction. To investigate the aforementioned issues linear and non-linear activation functions are chosen for dimension reduction using Stacked Autoencoder (SAE) when applied to Network Intrusion Detection Systems (NIDS). To conduct experiments for this study various activation functions like linear, Leaky ReLU, ELU, Tanh, sigmoid and softplus have been identified for the hidden and output layers. Adam optimizer and Mean Square Error loss functions are adopted for optimizing the learning process. The SVM-RBF classifier is applied to assess the classification accuracies of these activation functions by using CICIDS2017 dataset because it contains contemporary attacks on cloud environment. The performance metrics such as accuracy, precision, recall and F-measure are evaluated along with theses classification time is being considered as an important metric. Finally it is concluded that ELU is performed with low computational overhead with negligible difference of accuracy that is 97.33% when compared to other activation functions.

Keywords—Auto-encoder; cloud computing; dimension reduction; intrusion detection system; machine leaning

I. Introduction

The Cloud services availability to the individuals, organizations, and Governments connected through web-enabled devices across the world on pay-as-you-go premise [1] have become very common. The security and privacy problems get magnified as new type of attacks when internet environment migrate to Cloud [2]. Among these types of malicious activities Distributed Denial of Service (DDoS) attacks are easily invoked by the attackers basically with malicious intent of denying the Cloud services.

This type of attacks causes the interruption of cloud services to legitimate users by inordinate resource consumption which would automatically results in Service Level Agreement (SLA) violation. Most of the Cloud services are inherently elastic so the DDoS attacks are damaging the cloud service provider (CSP) economically but not its physical system or server assets [3]. This phenomenon is known as Economic Denial of Sustainability (EDoS) EDoS attack.

Due to increase in migration to the cloud the security and privacy problems are also increased in cloud environment with new types of malicious activity penetrated by professionals with cutting edge technologies with malicious intent. Over the last three decades the problems related to security and privacy are major research problems and addressed by several researchers with the evaluation of Network Intrusion Detection Systems (NIDS). It is a necessity to improve the NIDS to mitigate the new type of attacks on cloud because several users are getting migrated to cloud environment. It is a necessary to improve the NIDS to mitigate the attacks on cloud environment and this problem is addressed by several researchers. They have identified significant measures to detect and mitigate such types of attacks using statistical, Machine Learning (ML) techniques and knowledge based approaches. The NIDS needs to be more robust to increase the users trust in adoption of cloud computing in future. Therefore Network traffic analysis of cloud is necessary to identify the patterns inorder to discriminate malicious users and legitimate users. ML approaches offer great strength and diversity in research for anomaly detection in NIDS using classification task [4].

Lot of significant research work is going on over the past two decades on dimension reduction using Wrapper methods, Filter based Approaches, etc. to address the curse of dimensionality [5]. One of the advantages of NIDS is the availability of huge collection of network data related to cloud environment on which machine learning algorithms can be applied to detect attacks. Such a complex and huge data may disgrace the performance metrics of classifiers [6]. Dimensionality reduction or feature learning is one of the stages in classification which helps to extract the relevant information from the original data to reduce the computational time without compromising the other performance metrics.

Recently Deep learning methods from the family of Machine Learning approaches are successfully applied to extract good feature representation automatically [7]. Due to its capability for extracting valuable and useful information from large data yields better classification process and lower complexity.

The chosen activation function plays an important role in deep learning models to improve the accuracy rate and reduce computational complexity [8]. Activation function is one of the principal factors which will affect the performance of the neural networks [9]. Activation functions are basically divided into linear or non-linear relying upon the function it represents. Activation functions are leveraged to transfer and control the outputs of neural networks, across various domains from object recognition and classification.

This research work proposes a comprehensive investigative study intended to identify suitable activation function of SAE. By implementing this activation function an attempt is made to extract an optimized feature subset. The SAE extracts key features from the CICDS2017 which is exposed to lot of vulnerabilities as it is on Internet. Then a SVM classifier stage is used to classify the attacks and distinguish between normal and anomalous traffic.

After examining the contemporary studies the SAE model is identified as one of the best model for dimension reduction of various types of bigdata analytical content like image, text and network intrusion detection. There is a need to identify the proper activation function that will be more effective in the process of dimension reduction using SAE by satisfying the criteria like high classification accuracy and minimum classification time.

In view of the above the problem is selected to identify suitable activation function of SAE for dimension reduction and evaluate the classification accuracies. To achieve this objective the study is carried out with following contributions.

- Sharp decisions are needed within a stipulated time which plays an important role for NIDS. So to suggest a novel framework for better NIDS with suitable activation function and classifier.
- To compare the different activation functions of SAE using SVM classifier with RBF kernel.
- To evaluate the performance metrics and computational time with adoption of CICIDS2017 dataset.
- Finally identified the effective activation function based on experimental results.

The rest of this paper is organized as follows: the related previous work is outlined in Section 2 and Section 3 is explains the description of the CICIDS2017 dataset. The methodology and experimental setup of the proposed model is depicted in Section 4. The results and discussions are presented in Section 5. Finally conclusions and future scope of this work are given in Section 6.

II. Literature Review

An IDS has been studied for the last two decades using various machine learning approaches. This section will discuss some of the approaches proposed by researchers to analyze the effect of activation functions for IDS as well as for feature selection.

In [10], Investigated the performance of different types of rectified activation functions using Convolutional neural network. The standard rectified activation functions of rectified linear unit (ReLU), leaky rectified linear unit (Leaky ReLU), parametric rectified linear unit (PReLU) and a new randomized leaky rectified linear unit (RReLU) are evaluated with three different dataset NDSB, CIFAR-10 and CIFAR-100. Observations are discussed based on the exploratory results. However, this study is limited to different types of rectified activation functions and not pertaining to large datasets.

In [11], studied the effect of Activation functions on classification accuracy using Deep Artificial Neural Networks (ANN). The ANN is used to classify real multi-spectral Landsat 7 satellite images and therefore the accuracy of the classification was evaluated with twelve different activation functions. The accuracy of classifier can be improved by selecting good activation function. However, the activation functions sigmoid and bipolar activation functions are recommended to specific field of remote sensing.

It is Observed and stated in [12] that activation functions play pivotal role in understanding neural network. “DBN suffers from vanishing gradient problem due to the saturation characteristic of activation function. Therefore, the selection of activation function in DBN is critical to reduce the network complexity and to improve the performance of pattern recognition”. DBN based classification with different types of activation functions like Sigmoid, Hyperbolic Tangent, MSAF, ReLU and LReLU can be used to examine their performance with MNIST dataset. Besides that, the randomization of training samples would significantly improve the performance of DBN. The experimental results showed that hyperbolic tangent activation function achieved the lowest error rate which is 1.99% on MNIST handwritten digit dataset.

Kunang, Y. N et al. in [13] presented a deep learning based dimension reduction approach in the first step using stacked autoencoder. The input neurons 120 were reduced to 8 neurons with seven activation functions i.e., linear, sigmoid, ReLU, SoftMax, Softplus, Softsign and tanh and couple of loss functions like Mean Squared error and cross entropy to seek out the suitable activation function. In the second step reduced data is fed as input to supervised classifier SVM with RBF kernel to evaluate the output of first step. Based on the results it’s shown that the ReLU and linear activation functions with cross entropy yields better results over the other combinations.

The authors in [14] compared non-linear activation functions alternative to sigmoid in deep neural networks on Image Dataset, and also different weight initialization methods Gaussian distribution, uniform distribution, learning rate from 0.01 to 0.2, batch size 50 and 100 epochs and effect of hidden layers were tested. Based on the experimental results it was noticed that the accuracies of ReLU and its variants are higher than the sigmoid activation function and the learning rate is faster in ELU and SELU compared to the ReLU and Leaky ReLU. However, this study did not pay attention to computational complexity.

In [15] fast activation function adaptive Linear Function (ALF) is proposed for anomaly detection to increase the speed and accuracy of the deep leaning structure for real-time applications. Deep Belief Network with new activation function ALF and Sigmoid, Tanh, and ReLU are used for classification on four datasets namely NSL KDD, Kyoto, KDDCUP'99 and CSIC 2010. The experiments were conducted with a combination of each dataset using all four activation functions. Exploratory results have shown that learning structure using the fast ALF activation function outperforms the state-of-the-art network Stacked Sparse AutoEncoder Based Extreme Learning Machine (SSAELM) in accuracy and convergence time. However, the proposed approach needs to be evaluated on contemporary datasets.

Another research work [16] built IDS model with a blend of two models, autoencoder and deep artificial neural network. In the first stage Stacked autoencoder based dimension reduction with sigmoid activation function in input/hidden layer/output layer using 4 hidden layers and reconstruction error is calculated with mean squared loss function performed. In the next stage deep artificial neural network model was built utilizing different activation functions like hard sigmoid, relu, sigmoid, softplus and tanh to further classify the network traffic. Conclusions are drawn based on F1-score which is best with relu activation function in comparison to other activation functions. However, this study is limited to binary classification and need to, be tested on multi-class classification.

Feng, J., & Lu, S in [9] studied the characteristics of linear and non-linear activation functions Sigmoid, Tanh, ReLU, LeakyReLU, PReLU, RReLU, and ELU are compared. Activation functions advantages and disadvantages in Artificial Neural Network are also discussed and concluded that choosing of suitable activation function depends on our aim and network structure.

A Novel method was proposed by [17] to test SVM classifier with different kernels and Deep Convolutional Neural Network (DCNN) method for classification with different activation functions ReLU, Sigmoid, SoftMax, and Tanh. Before performing the intrusion detection chi-square based feature selection was performed to lessen the size of NSL-KDD dataset. The experiments were conducted with different activation functions by fixing one function from the above said four activation functions in input and hidden layers .Then the remaining are used at output layer respectively for each experiment. DCNN yields better results with sigmoid as output layer activation function with any of the activation function in input/hidden layers. The performance of SVM and DCNN are analyzed and concluded that the DCNN performed well compared to SVM classifier in terms of accuracy. This study focused on binary classification.

Le, T. T. H. et al. explored six non-linear Activation functions i.e., Softplus, ReLU, Tanh, Sigmoid, ELU and Leaky ReLU in [18] and additionally the impact of those activation functions was analyzed with Recurrent Neural Network (RNN) model to find the best activation function for intrusion detection. The KDD cup dataset was employed to conduct experiments. Among the employed activation functions Leaky ReLU the best results in terms of performance metrics accuracy, precision, recall and False Alarm Rate (FAR). The study is associated with performance metrics but does not concentrate on time complexity.

Most of the above mentioned studies focused mainly on comparing the various activation functions in evaluation of classifying the performance of neural network models and dimension reduction carried out with autoencoders. They considered the image datasets. In my knowledge and opinion very few studies are carried out by the researchers on network intrusion detection as well as autoencoders being used for dimension reduction on outdated datasets like KDD Cup ' 99 and NSL-KDD etc. To fulfill these research gaps the present research is carried out with adoption of CICIDS2017 dataset as it contains modern type of attacks and generated on cloud environment.

III. Description of CICIDS2017 Dataset

The CICIDS2017 dataset generated by [19] is chosen to train the model, which contains benign and the up-to-date common attacks, which resembles the real world data. It includes the network traffic from 09:00 on Monday, July 3rd and continuously ran for an exact duration of 5 days, ending at 17:00 on Friday July 7th. Network data is extracted using CIC Flow Meter, 2017 with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files). The dataset is publicly available in PCAP files[1]. In the current study only Wednesday data is considered for intrusion detection. It consists of 6,92,703 instances and 85 feature columns including a label with 6 classes such as Benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed. The distribution of records label wise is shown in below Table I.

The feature names and the Attacks distribution of the CICIDS2017 dataset are available in [20].

TABLE I. LABEL WISE DISTRIBUTION OF CICIDS2017 DATASET

Category	Class	Number of Records
Anomaly	DoS GoldenEye	10,293
	DoS Hulk	231,073
	DoS Slowhttptest	5,499
	DoS slowloris	5,796
	Heartbleed	11
Total Anomaly data	---	252,672
Normal	Benign	440,031
Total		692,703

IV. METHODOLOGY

A novel framework is developed for analyzing the performance of various activation functions of SAE for dimension reduction. For evaluation of the performance of these dimension reduction models SVM-RBF classifier is chosen. For this purpose CICIDS2017 dataset is used for conducting experiments. This methodology consists of three phases: A) Data pre-processing, B) Dimension Reduction, and C) SVM-RBF classification. The flow of the proposed framework is depicted in Fig. 1.

A. Data Preprocessings

In the first phase data cleaning and normalization operations are carried out.

Data cleaning: It involves finding the null records as Machine Learning algorithms cannot build a model or test them. Since the percentage of null records associated with each class is small so these records are deleted from the dataset.

Based on the statistical measures of the dataset given in Table II, such as maximum, minimum, standard deviation of each feature it is observed that all the values of attributes Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, CWE Flag Count, Fwd Avg Bytes, Fwd Avg Packets, Fwd Avg Bulk Rate, Bwd Avg Bytes, Bwd Avg Packets, Bwd Avg Bulk Rate are null in the dataset. These columns are also deleted from dataset because they potentially don't have any contribution towards classification.

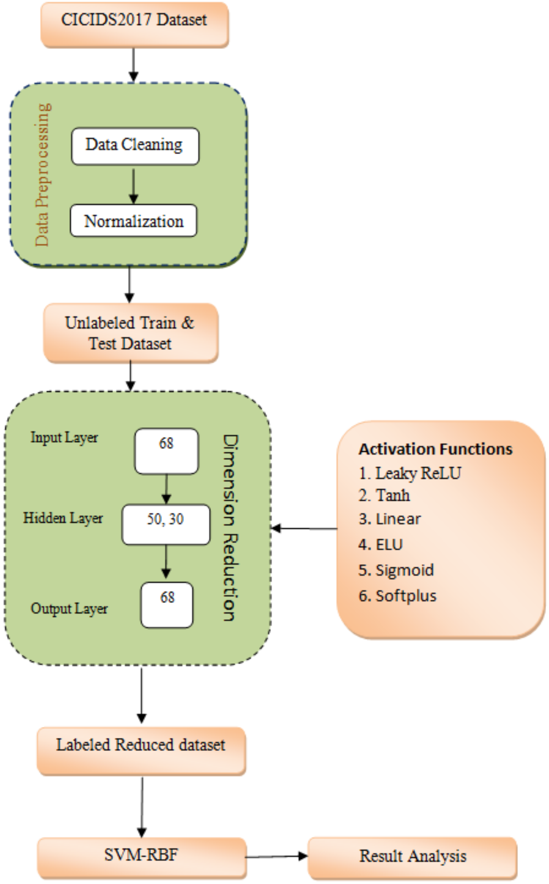


Fig. 1. Novel Framework for Comparison of Activation Functions for NIDS.

TABLE II. STATISTICAL MEASURES OF FEATURES WITH ZERO VALUES

	Bwd PSH Flags	Fwd URG Flags	Bwd URG Flags	CWE Flag Count	Fwd Avg Bytes/ Bulk	Fwd Avg Packets/Bulk	Fwd Avg Bulk Rate	Bwd Avg Bytes/Bulk	Bwd Avg Packets/Bulk	Bwd Avg Bulk Rate
count	692703	692703	692703	692703	692703	692703	692703	692703	692703	692703
mean	0	0	0	0	0	0	0	0	0	0
Std	0	0	0	0	0	0	0	0	0	0
min	0	0	0	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0	0	0
75%	0	0	0	0	0	0	0	0	0	0
max	0	0	0	0	0	0	0	0	0	0

TABLE III. LABEL WISE DISTRIBUTION OF CICIDS2017 DATASET AFTER DATA CLEANING

Category	Class	Number of Records
Anomaly	DoS GoldenEye	10293
	DoS Hulk	2,30,124
	DoS Slowhttptest	5,499
	DoS slowloris	5,796
	Heartbleed	11
Total Anomaly data	---	2,51,723
Normal	Benign	4,39,972
Total	---	6,91,695

Additionally two features Byte/s and Flow Packet/s have infinity/NaN values in very few records which are replaced with zeros to avoid difficulties in applying the ML algorithms. After performing data cleaning operation the distribution of the remaining dataset is given Table III.

Normalization: [1]By nature the features of the dataset are quantitative. Several features of the dataset lie in a wide range mostly between the highest possible value and lowest possible value.

The features with higher values may dominate the features with low values. To avoid this, features to be normalized to eliminate such dominance before applying ML algorithms.

Normalization is a scaling technique in which values are transformed and rescaled so that they fall in specific range [0, 1]. In this paper features are scaled using min-max normalization. The feature f values with a range between fmin and fmax , then the normalization is defined by the equation of fnor= (fi - fmin)/(fmax-fmin), Where fnor is a normalized value of the ith value of feature f. An illustrative example of normalization for one record is given in Fig. 2.

B. Dimension Reduction using SAE

This section explores the impact of the different activation functions of SAE with specified hyper parameters as given below for feature dimensionality reduction.

The performance of NIDS using neural networks for dimension reduction as well as classification is dependent on three criteria. They are i) to identify suitable activation function, ii) the number of hidden layers to be used and iii) to adjust the proper weights for minimizing the loss between input and output. Using SAE for dimension reduction the chosen of activation function is a crucial problem.

Among the three the first one plays a crucial role for optimizing the feature subset. So this study considers to find the impact of the first criterion and the remaining are fixed at number of hidden layers are 3 and Mean Square Error as a loss function. The main focus of this study is to compare various prominent activation functions that are used in SAE for NIDS and evaluate them through the SVM classifier. Proper activation function also affects the classification time along with classification accuracy [11]. For this purpose six activation functions are chosen. The following Fig. 3 depicts the typical structure of the SAE which consists of input and output layers with three hidden layers, the same is being considered for conducting experiments.

0.352941, 0.000319,0, 3.67E-06, 4.90E-06, 9.57E-09, 0.000242, 0.002906, 0.001293, 0, 0.000307, 0.003026, 0.001373, 0, 0.005764, 0.40001, 0.000319,0, 0.000319, 0.000319,0,0,0,0, 6.67E-08,0,0,0,0,0,0, 7.23E-06, 3.67E-06, 8.70E-06, 1.31E-05, 0.004144, 0.000242, 0.002632,0,0,0,0,0,0,1,1,0, 0.023256, 0.003446, 0.001293, 0.001373, 7.23E-06,0, 4.90E-06, 3.67E-06, 9.57E-09, 0.003906, 0.01445,0, 0.344262,0,0,0,0,0,0,0,0,0

Fig. 2. Resulting One Sample Record after Normalization.

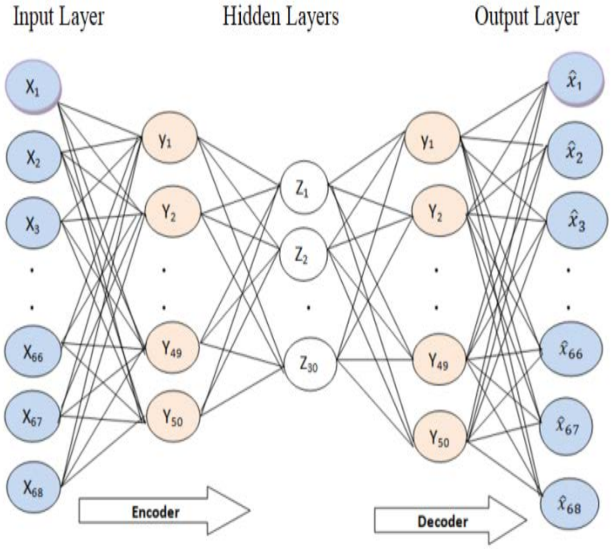


Fig. 3. Typical Structure of Stacked Autoencoder.

[1] http:[1]//www.unb.ca/cic/datasets/IDS2017.html.

The below SAE has 68 neurons in both input and output layers which are equal to the features of normalized train and test dataset. The numbers of neurons in three hidden layers are 50, 30 and 50 respectively. The role of activation function in neural network is to facilitate the transfer of data from input layer neurons to output layer neurons [18]. The chosen six activation functions are being applied.

Several different kinds of optimizers are used by different researchers even though their common purpose is minimizing the loss in learning process. Among those Adaptive Moment Estimation (Adam) is chosen for this study [21]. The following function is adopted for reconstruction error computation [22]. The mean square error function is calculated for the samples of $\mathrm{x_i}$ and reconstructed samples $\hat{\mathrm{x}}_\mathrm{i}$ as shown in equation 1. To find the optimal structure of SAE for dimension reduction with afore mentioned optimizer and loss function used. The unlabeled normalized train and test datasets of CICIDS2017 are taken as input datasets, because SAE is an unsupervised dimension reduction approach.

$$\mathrm{d}\,(x\ ,\hat{x}\ )=\frac{1}{N}\sum_{i=1}^{N}||x_i-\hat{x}_i||^2\ , \qquad (1)$$

where N is the number of input samples.

The following algorithms takes various hyper parameters weight vector W, bias vector, batch_size, epochs, number of layers (L), learning rate α and neurons_structure in each layer are assigned `to initial values. The training procedure of unsupervised SAE model for dimension reduction is given in Algorithm1 and Algorithm2. The input to Algorithm1 is number of SAE structures, number of hidden layers in each structure. Step3 inputs the number of neurons in each hidden layer. Followed by in step4 compute_MSE ( ) is invoked for each structure to setup the corresponding SAE architecture and iterated for all epochs with given activation function then returns the MSE error which is stored in loss. Input to Algorithm2 are train and test datasets {x1, x2… x68} where sample X∈R where, structures of SAE and the number of neurons in each structures. For each layer l the parameters are initialized to zero. In each iteration the hidden representation vector of layer l is computed based on the previous layer. After that the loss is calculated equation (4) in step 2.2.2, Next step is to update θ′ based on loss and completes all iterations similarly. Finally mean of all the epochs MSE loss is calculated and returned. Once the MSE losses of different structures are calculated then the minimum loss is identified and the corresponding structure is selected as optimized SAE structure for dimension reduction. All the above steps are repeated for each activation function.

Algorithm1: The training procedure to identify optimal SAE structure for dimension reduction is given in Algorithm1

```
Algorithm1: Optimized SAE model for dimension reduction

Input: S                        // number of SAE structures
       L                        // number of layers in structure
       neurons_structure [ ] [ ]  // array of SAE structures

Output: Optimized SAE structure

Step 1:  // Initialization
         S=0, neurons_structure [ ] [ ] <- 0, loss [ ]

Step 2: read S //Total Number of Structures

Step 3: for I in 1 to S do // for each structure I
    Step 3.1:  read L [I] // Number of Layers in the structure
    Step 3.2: for J in 1 to L [I] do
        Step 3.2.1:  read N // number of neurons in Layer J of structure I
        Step 3.2.2: store neurons_structure [I] [J] =N
    end for
end for
Step 4: for s in 1 to S do
    loss[s] = compute_MSE(s, neurons_structure [I], L [I], bias[I])
end for

Step 5: betsvalidation_loss= loss [1]
    Step 5.1 for k in 2 to S
        Step 5.1.1 if (bestvalidation_loss > loss[k])
            then
                bestvalidation_loss=loss[k]
                location = k
            end if
    end for

Step 6:  optimal structure= neurons_structure [location] [  ]
```

As a part of initiation towards optimization various structures of neural networks are used to reduce the MSE as per equation7. In the lines of [13] several experiments were conducted for the number of features and the corresponding MSEs are listed in Table IV. It is observed from the above table.

Algorithm2: The training procedure of unsupervised SAE model for dimension reduction is given in Algorithm2

```
Algorithm2: Compute_MSE (s, neurons_structure [ ], L) // Training SAE and calculate MSE

Input: Unlabeled normalized train dataset (UNTR) and test dataset (UNTS), UNTR & UNTS ⊆ R &
          UNTR ∩ UNTS=∅, data sample X ∈R where {$x_1$, $x_2$… $x_{68}$}
Output: Z      // Mean Square Error

Step 1:  /*Initialization

    h [ ][]<- 0 // output of each neurons of each hidden layer
    $\hat{x}_i$[ ][ ] <- 0 //reconstructed data
    batch_size =256 // number of samples per each iteration
    bias [ ] // bias vector of each hidden layer
    epoch=10 // number of passes to execute learning algorithm for entire dataset
    learning_rate =0.001 // weight adjusting control variable w.r.t gradient
    mse_sum=0, mean_loss=0
Step 2: create and configure the stacked autoencoder network with Adam optimizer
    Step 2.2: for e ∈ 1 to epoch do
        Step 2.2.1: for b ∈ 1 to b do
            n=neurons_structure [1]; // number of neurons for the first layer
            W = 0 // weight vector W= {$w_1$, $w_2$...,$w_n$}
            define h[1] [ ] = f (WX +bias) // First hidden layer
            for l in 2 to L do
                n=neurons_structure[l];
                W= 0 // weight vector W= {$w_1$, $w_2$...$w_n$}
                bias =1  // bias value
                h[l][ ]= f (W * $h_{l-1}$+bias)  // compute $h_l$ from $h_{l-1}$ of both
                                                     encoder and decoder
            end for
        end for
        Step 2.2.2 $\hat{x}_i$[ ][ ]=h[l][ ]
            for I in 1 to $\hat{x}_i$   do
            Z <-  $\frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i[\,][\,]||^2$
            sum = mse_sum + Z
            end for
        update layer parameters θ′l = (W, B) s. t Z is minimum
    end for

Step 4: mean_loss = $\frac{\text{mse\_sum}}{e}$

Step 5: return mean_loss
```

It is observed from the above table that the optimized MSE for majority of the activation functions is w.r.t to neurons structure 68-50-30-50-68. Hence this neurons structure is considered to reduce the dimensionality to 30.Once better optimized structure is identified then series of experiments are conducted for dimension reduction on training and testing data. The inner most hidden layer i.e. last layer of encoder provides the richer representation of CICIDS2017 i.e. 30 features with reduction of 44% of data. Evaluation of the classification efficiency of various activation functions used for dimension reduction are evaluated with SVM-RBF classification model. The next section explores the classification evaluation model.

TABLE IV. MEAN SQUARE ERRORS FOR SOME OF THE STRUCTURES OUT OF THE CONDUCTED EXPERIMENTS

Neurons Structure\ Activation function	Leaky ReLU	Tanh	Linear	ELU	Sigmoid	Softplus
SAE(68-34-17)	0.00007	0.000139	0.000173	0.000072	0.000217	0.000123
SAE(68-50-40-30-10-5)	0.000324	0.0003	0.003887	0.000208	0.002995	0.000851
SAE(68-50-35-20)	0.000373	0.000082	9.52548E-05	0.000039	0.00083	0.000091
SAE(68-50-40-30-20)	0.00028	0.000083	0.000096	0.000039	0.002366	0.000498
SAE(68-34-17-8)	0.000276	0.000284	0.001656	0.000166	0.001296	0.000249
SAE(68-50-40-30)	0.001802	0.000053	2.07121E-05	0.000031	0.00051	0.000208
SAE(68-40-20)	0.000052	0.000091	0.071884	0.000042	0.000224	0.000114
SAE(68-50-30)	0.000042	0.000048	0.00002014	0.000121	0.000136	0.000091

C. SVM-RBF Classification Model

In this module SVM classifier with RBF kernel is used for multi–class classification. Initially SVM-RBF base model is trained and tested with default values of c and gamma. From the existing literature it has been found that the usage of SVM-RBF model in the context of multiclass classification is more suitable than any other existing classifier. Complete dataset which contains 68 features after due preprocessing has been given as an input to the python program with scikit library intended for dimension reduction. Later SVM-RBF classifier is trained with derived reduced datasets of six different SAE models using different activation functions which would serve as a thorough evaluation of the activation functions along with the corresponding confusion matrices. Further the standard performance metrics vis-a-vis accuracy, precision, recall and F-measure are calculated. These experiments were conducted on an experimental setup, Intel core i5 with 1.80 GHz processor with 8GB RAM, windows 10.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section discusses about the effect of six activation functions pertaining to SAE with the adoption of MSE as a loss function for dimension reduction. Further comparison of different classification metrics which are derived through SVM-RBF classifier is done. The experimental results w.r.t various performance metrics and computational time are given in below Tables V to VII as well as from Fig. 4 to 8.

TABLE V. VARIATION OF ACCURACY IN TRAINING AND TESTING FOR DIFFERENT ACTIVATION FUNCTIONS

Activation Function	**Training Accuracy**	**Testing Accuracy**	**Average of Training and Testing Accuracy**
Without Feature Selection	96.15	96.2	96.175
Leaky ReLU	96.89	96.96	96.925
tanh	96.98	97.05	97.015
Linear	97.19	97.26	97.225
ELU	97.26	97.33	97.295
Sigmoid	97.36	97.38	97.37
Softplus	97.33	97.4	97.365

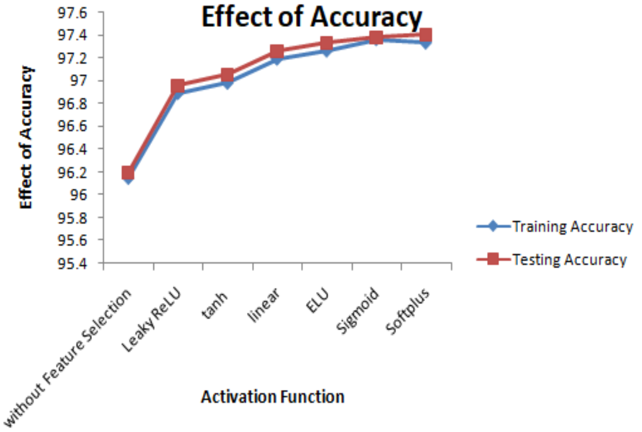


Fig. 4. Effect of the Accuracy of Training and Testing for different Activation Functions.

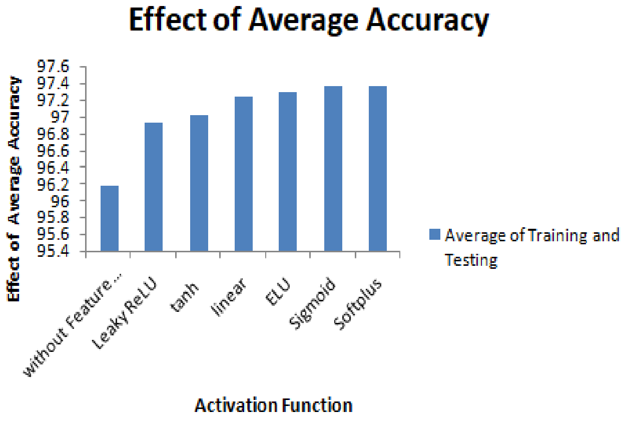


Fig. 5. Effect of Average Computational Time of Training and Testing Time for different Activation Functions.

It is observed that from Fig. 4 and 5:

- The six activation functions exhibits similar behavior for both training and testing.
- Amongst the six softplus gives better accuracy.
- Sigmoid and ELU follows softplus with a minor difference of 0.02 and 0.07, respectively.

TABLE VI. VARIATION OF COMPUTATIONAL TIME IN TRAINING AND TESTING FOR DIFFERENT ACTIVATION FUNCTIONS

Activation Function	**Training Time(Sec.)**	**Testing Time (Sec.)**	**Average of Training and Testing(Sec.)**
Without Feature Selection	24690.515	1123.724	12907.12
Leaky ReLU	4948.109	393.355	2670.732
tanh	7254.742	401.517	3828.1295
Linear	7587.699	302.599	3945.149
ELU	3115.355	306.596	1710.9755
Sigmoid	9128.278	364.093	4746.1855
Softplus	3837.29	338.254	2087.772

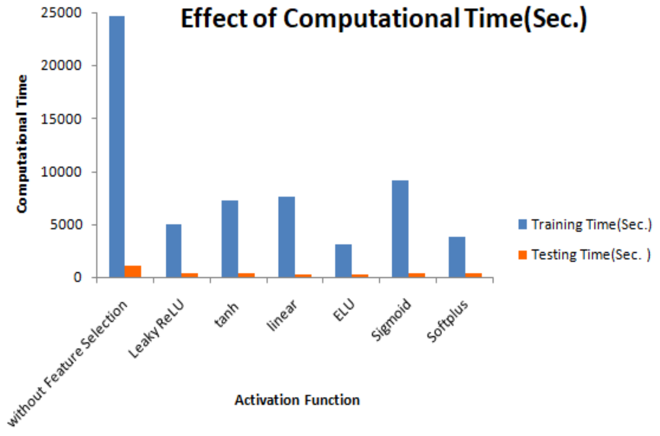


Fig. 6. Effect of the Computational Time of Training and Testing Time for different Activation Functions.

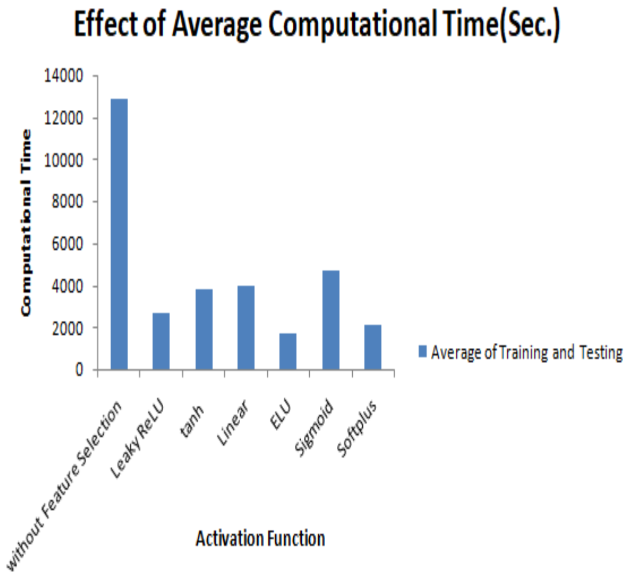


Fig. 7. Effect of Average Computational Time of Training and Testing Time for different Activation Functions.

The computational time of different methods of activation functions are presented in Table VI and Fig. 6 and 7. It is observed that the training and testing time of ELU activation function is minimum when compared to remaining activation functions. The activation function sigmoid takes highest total execution time. The other two activation functions tanh and linear exhibit more or less same computation time. Therefore ELU activation function must be the choice of researchers to reduce total training time and testing of SVM-RBF classifier using CICIDS2017 dataset.

The experimental results for metrics Precision, Recall, F-measure are presented in Table VII and Fig. 8. In case of precision linear and ELU activation functions exhibits highest performance with a value 0.97 and remaining all four activation function Leaky ReLU, tanh, Sigmoid and Softplus are on the lower side with a value of 0.96. The recall value of linear and ELU activation functions are higher and equal compared to other methods with a value of 0.95. The activation function linear shows next highest value with 0.95 followed by Sigmoid and Softplus which show with an equal value of 0.94. The activation functions linear and ELU obtained best measure results compared to other activation functions. It can be observed that the remaining activation functions are performing more or less equally with a minor difference ranging between 0.1 and 0.5.

TABLE VII. VARIATION OF PRECISION, RECALL AND F-MEASURE ACCURACY IN TRAINING AND TESTING FOR DIFFERENT ACTIVATION FUNCTIONS

Activation Function	**Precision**	**Recall**	**F-measure**
Without Feature Selection	0.79	0.68	0.72
Leaky ReLU	0.96	0.88	0.91
tanh	0.96	0.93	0.95
Linear	0.97	0.95	0.96
ELU	0.97	0.95	0.96
Sigmoid	0.96	0.94	0.95
Softplus	0.96	0.94	0.95

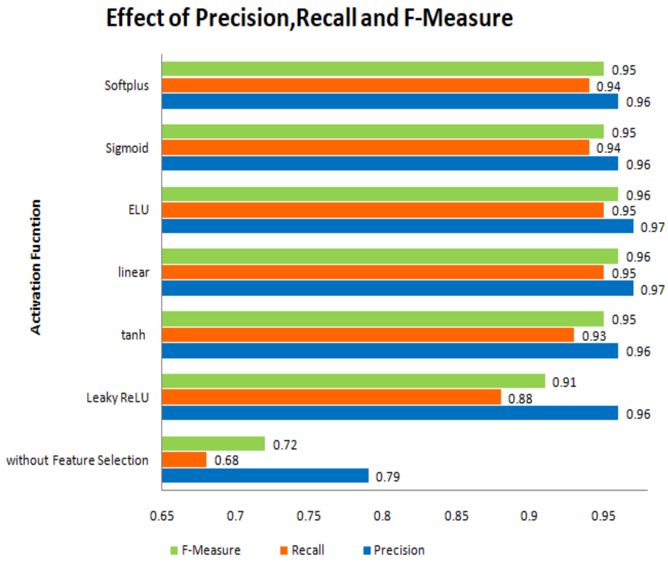


Fig. 8. Effect of Precision, Recall and F-Measure for different Activation Functions.

VI. Conclusion

This paper is intended to evaluate the SAE model for dimension reduction through six activation functions. The Experimental results exhibit that the activation function ELU gives better performance in terms of computational time. In the context of precision, recall and F-measure ELU and linear activation functions leads other functions. When classification accuracies are compared the Softplus yields marginal performance. From these experimental results it is observed that ELU is a better activation function with respect to computational time. Whereas ELU and linear provides better performance with respect to other performance metrics. To consider both computational time and classification performance the ELU is better with compromising of negligible difference of accuracy. Finally it is concluded that this comparative study will be of great help to the defenders to design suitable framework for NIDs on cloud environment to defend the intruders within a stipulated time. The extension for this study could be to compare the performance evaluation of different kernel functions of SVM with conducting of more experiments. As an enhancement of this study one can think about evaluating the performance metrics on a real time environment.

References

[1] Ahmed, H. A. S., Ali, M. H., Kadhum, L. M., Zolkipli, M. F., & Alsariera, Y. A. (2017). “A review of challenges and security risks of cloud computing”. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *9*(1-2), 87-91.

[2] Khalil, I. M., Khreishah, A., & Azeem, M. (2014). “Cloud computing security: a survey”. *Computers*, *3*(1), 1-35.

[3] Bulla, S., Rao, B. B., Rao, K. G., & Chandan, K. (2018). “An experimental evaluation of the impact of the EDoS attacks against cloud computing services using AWS”. *International Journal of Engineering & Technology*, *7*(1.5), 202-208.

[4] Liu, H., & Lang, B. (2019). “Machine learning and deep learning methods for intrusion detection systems: A survey”. applied sciences, 9(20), 4396.

[5] Goswami, S., & Chakrabarti, A. (2014). “ Feature selection: A practitioner view”. *International Journal of Information Technology and Computer Science (IJITCS)*, *6*(11), 66.

[6] Attallah, O., Sharkas, M. A., & Gadelkarim, H. (2020). “ Deep Learning Techniques for Automatic Detection of Embryonic Neurodevelopmental Disorders “. *Diagnostics*, *10*(1), 27.

[7] Bhati, B. S., & Rai, C. S. (2020). “Analysis of support vector machine-based intrusion detection techniques. *Arabian Journal for Science and Engineering”*, *45*(4), 2371-2383.

[8] Chigozie Enyinna Nwankpa et al, 2018, “ Activation Functions: Comparison of Trends in Practice and Research for Deep Learning”.

[9] Feng, J., & Lu, S. (2019, June). “Performance analysis of various activation functions in artificial neural networks”. In *Journal of Physics: Conference Series* (Vol. 1237, No. 2, p. 022030). IOP Publishing.

[10] Xu, B., Wang, N., Chen, T., & Li, M. (2015). “Empirical evaluation of rectified activations in convolutional network”. *arXiv preprint arXiv:1505.00853*.

[11] Serwa, A. (2017). “Studying the Effect of Activation Function on Classification Accuracy Using Deep Artificial Neural Networks”. *Journal of Remote Sensing & GIS*, *6*(03), 6-11.

[12] Lau, M. M., & Lim, K. H. (2017, April). “Investigation of activation functions in deep belief network”. In *2017 2nd international conference on control and robotics engineering (ICCRE)* (pp. 201-206). IEEE.

[13] Kunang, Y. N., Nurmaini, S., Stiawan, D., Zarkasi, A., & Jasmir, F. (2018, October). “Automatic Features Extraction Using Autoencoder in Intrusion Detection System”. In *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)* (pp. 219-224). IEEE.

[14] Pedamonti, D. (2018). “Comparison of non-linear activation functions for deep neural networks on MNIST classification task”. *arXiv preprint arXiv:1804.02763*.

[15] Alrawashdeh, K., & Purdy, C. (2018, May).” Fast activation function approach for deep learning based online anomaly intrusion detection”. In *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 5-13). IEEE.

[16] Catak, F. O., & Mustacoglu, A. F. (2019).” Distributed denial of service attack detection using autoencoder and deep neural network”. *Journal of Intelligent & Fuzzy Systems*, *37*(3), 3969-3979.

[17] Sstla, V., Kolli, V. K., Voggu, L. K., Bhavanam, R., & Vallabhasoyula, S. “Predictive Model for Network Intrusion Detection System Using Deep Learning Predictive Model for Network Intrusion Detection System Using Deep Learning”.

[18] Le, T. T. H., Kim, J., & Kim, H. (2016). “Analyzing Effective of Activation Functions on Recurrent Neural Networks for Intrusion Detection”. *Journal of Multimedia Information System*, *3*(3), 91-96.

[19] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January). “Toward generating a new intrusion detection dataset and intrusion traffic characterization” . In *ICISSP* (pp. 108-116).

[20] Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., & Abuzneid, A. (2019).”Features dimensionality reduction approaches for machine learning based network intrusion detection”. Electronics, 8(3), 322.

[21] El Mrabet, Z., Ezzari, M., Elghazi, H., & El Majd, B. A. (2019, March). “Deep Learning-Based Intrusion Detection System for Advanced Metering Infrastructure”. In *Proceedings of the 2nd International Conference on Networking, Information Systems & Security* (pp. 1-7).

[22] Farahnakian, F., & Heikkonen, J. (2018, February).” A deep auto-encoder based approach for intrusion detection system”. In *2018 20th International Conference on Advanced Communication Technology (ICACT)* (pp. 178-183). IEEE.

GAAR: Gross Anatomy using Augmented Reality Mobile Application

Wan Aezwani Wan Abu Bakar[1], Mohd Airil Solehan[3]
Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin (UniSZA)
Besut Campus, 22200 Besut, Terengganu, Malaysia

Mustafa Man[2], Ily Amalina Ahmad Sabri[4]
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu (UMT)
21030 Kuala Nerus, Terengganu, Malaysia

***Abstract*—Covid-19 pandemic has forced the teaching and learning activity into real time and real-world education meetings. The traditional physical and face-to-face meetings are avoided in accordance to reducing the close physical contacts among individuals. Thus, a new paradigm shift towards teaching and learning needs to be highly enforced. Teaching and learning on medical field especially require for real world anatomy against human or living things body. In response to providing the facility for medical teachers and learners, Gross Anatomy Augmented Reality (GAAR) is introduced. GAAR is an android mobile Augmented Reality (AR) learning tool to assist the educators and learners in internalizing 3D human anatomy with more fun and interactivity. The AR methodology is implied to attract the personal impacts and feelings towards operating of close to 'real' organ during anatomy practices. Traditional learning methods are changed with AR technology through small digital device. This application may be able to show the students the actual form of human gross anatomy and assist the teachers or educators' in explaining the sciences behind human body in more interactive and interesting. Furthermore, this application uses a 3-dimensional object, video and interactive info so that students are interested in using this application. The AR for education and learning is vital in bridging the digital divide among all generations through the conversion of static pictures into real-like 3D animation. The implementation results show that, through the real visualization, small to adult learners can imitate the real truth on human organ and how this can motivate them to take care of their bodies that would lead to a healthier living styles as well as easy memorizing of the subject contents.**

Keywords*—Augmented reality; gross anatomy; learning tool; android mobile application; 3D human anatomy

I. Introduction

Augmented reality (AR) is a technology of putting the static object into real and dynamic object. In AR, each object is enhanced to become a real physical world that is achieved through the use of digital visual elements, sound, or other sensory stimuli. The augmented technology aims to be an alternative method for educators and learners that offer for interactivity as compared to traditional methods. In response to Covid19 pandemic where the education and learning is conducted through virtual, the education sectors especially is facing greatest challenge in delivering the contents of any subject matters to all learners and key persons. Thus, some domain fields such as engineering, medical, pure sciences that requires total theory and practice are in demand for the technology that could build for a full interest among educators and learners.

Hence, Gross Anatomy Augmented Reality (GAAR) is developed as a touch-up and a complimentary tool as a combination of gross anatomy study into augmented reality (or close to reality) during early education level. This mobile AR application aims to enhance teacher's ability to educate students, drives their interest and help them to internalize through visual implicitly in the field of science and human anatomy. The environment of first, learning the theory, second make it visualize in real into their small and handy gadgets (i.e. smartphone device) could assist the implementation of GAAR to be run at all school.

The ultimate objective of this project is to design an application that include 3D models of anatomy such as liver, heart, kidney, stomach and skeleton system that have been developed during creation of mobile application environment combined with AR for the study of basic human anatomy. To achieve this objective, the following purpose hypothesis are tested: The development of GAAR mobile application specific purpose is for the field of human anatomy in three dimension (3D) in combination of AR is feasible and achievable, in order to be used as an educational tool for the student in primary and secondary schools. This acts as an alternative or replacement of a traditional teaching method and a description course on human anatomy.

Prior to the development of a good working prototype of the application, the next hypothesis is it usability. The developed GAAR serves as an alternative mechanism in teaching and learning human gross anatomy. The real implementation on the feasibility for the infrastructure and to evaluate errors in the model (human organ) and overall "look-and-feel" of the application is depending upon the feedback of future users of GAAR. This initiative aims to fulfill the successful run of operating scenarios of use as well as user satisfaction. The rest section of the study is organized as follows: Section 2 reviews on the related works and tools used for modeling and AR, Section 3 discusses on the methodology undertaken, Section 4 illustrates the experimentation of GAAR. Discussion is outlined in Section 5 and Section 6 summarizes the conclusion and future work of GAAR.

II. Related Works

A. Reviews of AR Apps

Teaching and learning demands for a best-fit approach for a sustainable long-life learning environment [1]. Especially in Medical Schools [2], medication practitioners practice the use of recorded videos [3] and podcasts [4] for virtual learning and during the online learning process [5]. The traditional method for an anatomy curriculum during 1980s demands for an improvements [6] and many technological advances in educational sectors has been proposed with the introduction of innovations such as virtual anatomy courses and dissections [7], three-dimensional (3D) atlas as initiated in [8] and "mARble" application that embeds augmented reality (AR) feature for iPhone users [9]. The use of AR applications in medical empower each user experiences and describes the appearance of a digital element generated with a computer [10]. The AR generates a fascinating output since the user can see in 360 angle degree from the human body until its anatomy [11]. Besides, an application that uses AR which called ROAR Augmented Reality created by Roar Interactive on 13 March 2016. *BARETA* [12] is an AR tool that combines AR and 3D modeling objects to provide simulation for touching as well as sighting.

As reported in School of Medicine of the Aristotle University of Thessaloniki, Greece, indicates the educational innovations in for teaching anatomy is being stigmatized with Andrea Vesalius paradigm [13]. This could be the facts towards dissection in anatomy. The actual intention of AR is not to replace anatomy dissection activity, but it tries to offer the alternative efforts to supplement the traditional curriculum [14-15]. Wikitude for instance, is coupled with its SDK7 that includes simultaneous localizing and mapping for recognition and tracking in 3D objects, image recognizing and tracking, cloud recognition, location-based services, smart glasses as well as external plugin integration in Unity [16].

B. Reviews of AR Platforms

1) Android OS: The Android Operating System is meant for the mobile device platform that underlies the Linux kernel. It was initiated by Open Handset Alliance on November 2007 using Java programming language [17]. Android is dedicated for open expansion platform where it offers upgrades across various versions. When Android starts releasing its source code, anyone can change it according to individual interests. The project uses the Android Software Developer Kit (SDK), that comes together with the tools and APIs. We integrate the SDK into a graphical user IDE (Integrated Development Environment). The novice can also use App Inventor application to develop an Android-based applications that can be browsed through online.

2) iOS: The iOS is Apple-mobile operating system (OS) for any Apple-manufactured devices includes Apple iPhone, Apple iPad, Apple iPod Touch and Apple TV. These products are popular with the niche style of identity recognition by gestures for instance the gestures to swipe, tap and also pinch. These gestures movement are performed on multi-touch capacities on screen touching displays that provides instant response actions and identify the fingers and pen that act as input.

C. Reviews of AR SDK

1) Vuforia: Vuforia is an augmented reality software development kit (SDK) that has been developed for mobile devices. It enables the creation of augmented reality applications. It uses computer vision technology to recognize and track planar images and 3D objects in real time. Vuforia acts as the first batch of AR environment for an AR application development that comes with a broad set of features. The Vuforia SDK for AR features comprises of a multi-objects recognition for objects like cylinders and boxes as well as the 3D objects like images, recognition for text of approximately 100,000 words in its vocabulary, the ability to customize the *VuMarks*, some form of high quality objects than any typical QR-code, feature for 3D geometric map of any environment using Smart terrain, feature to turn static image into full motion video, feature for Unity Plugin and lastly support for cloud and local storage. The Vuforia is supported in iOS, Android, Universal Window Pattern and Unity.

2) ARToolkit: An open-ended source tool for developing an AR application. Equipped with rich set of features for tracking, which include *Unity3D* and *OpenSceneGraph* that includes the facility for both single and dual camera, an apps with featuring object based on location such as GPS and compass support. Also added with the feature to create real-time AR Application, can be integrated with smart glasses, support for multiple language and automatic camera calibration. The ARToolkit supports for Android, iOS, Linux, Window, Mac OS and Smart Glasses and since it is an open source, it is free to use.

3) Wikitude: Wikitude and its SDK7 are included for parallel localizing and mapping. Offers for 3D image recognizing and tracking, recognized within cloud environment with location-based services and integrated with smart glasses and external plugins include Unity. The Wikitude is supported in Android, iOS and smart glasses platform.

The main scope of this article is to describe and demonstrate that development of a free educational mobile application that, with the use of AR, could also help teacher and students interact with each other and provide the user with basic anatomical knowledge for the human internal organs and skeleton system. The breakthrough aim of this application is to introduce the new technological techniques and tools that may reinforces and embarks Malaysia education processes when involving anatomy subject in primary or secondary schools.

III. Methodology

The research methodology is depicted in Fig. 1. There are four phases i.e. Phase 1: Creation of 3D Models, Phase 2: Development of Augmented Reality (AR), Phase 3: Combination of 3D Models and AR and final phase, Phase 4: Development of GAAR Mobile Apps.

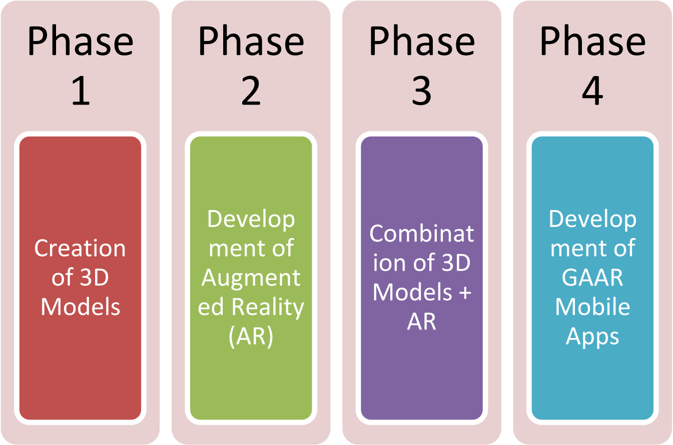

Phase 1
Creation of 3D Models
Phase 2
Development of Augmented Reality (AR)
Phase 3
Combination of 3D Models + AR
Phase 4
Development of GAAR Mobile Apps

Fig. 1. GAAR Methodology.

To embark the teaching and learning approach with interactivity, the conceptual model of GAAR is outlined as in Fig. 2. The AR technology is to bridge the gap between the real-world surroundings (physical area) and the digital area. The teaching and learning activities are done with interactivities via a tool where there is only need to scan and doing anatomy exercises and see references from the resources provided in GAAR library. The specific content of GAAR is shown in Fig. 3. The traditional textbooks such as Nota Sains Sukan Tahap 1, Skim Persijilan Kejurulatihan Kebangsaan (SPKK) [18] and Augmented Reality for the Study of Human Heart Anatomy [19] are used to evaluate the 3D model of human anatomy such as heart, lung, kidney, human skeleton and liver to design for GAAR content structure.

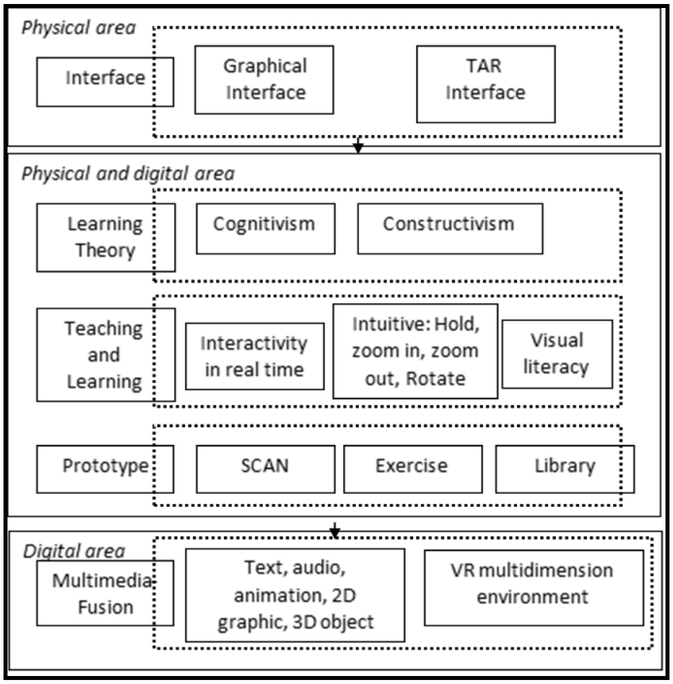


Fig. 2. GAAR Conceptual Model.

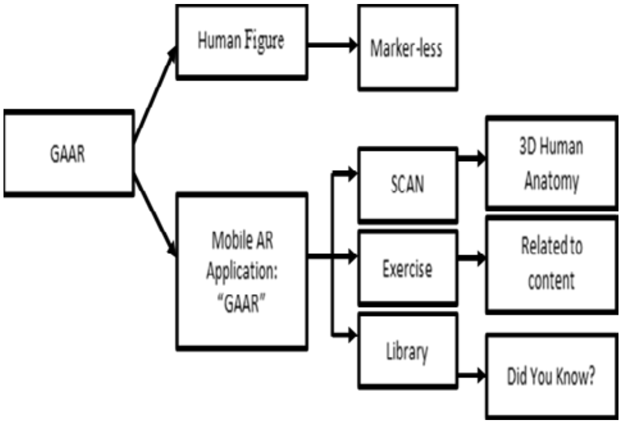


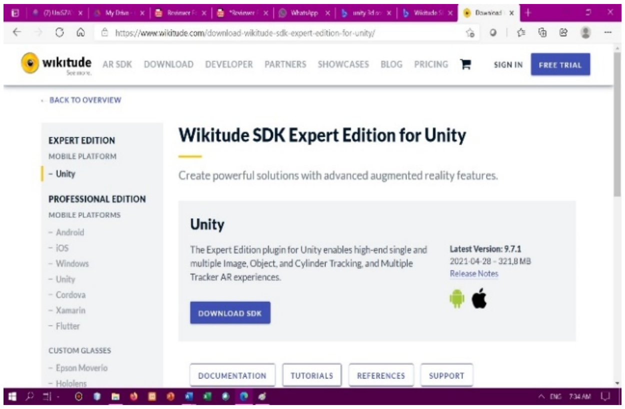
Fig. 3. GAAR Content Structure.

IV. Experimentation

A. Development

The interactive AR mobile application is only for Android mobile platform use only. Many external code libraries were used: Wikitude SDK, Android SDK, for the 3D models and the Autodesk Maya program for a 3D modelling and animation software. C# programming is used for creating mobile application using Unity3D. The GAAR mobile application used the concept of marker-less which mean that this mobile application does not required any medium or special paper to interact as marker. The only requirement for GAAR mobile application to work functionally need mobile phone with camera, android version 7.0 above and a flat surrounding to display the models. Using English language, the main menu of the mobile application is divided in seven categories i.e. Book Section, Quiz Game, Anatomy AR, About App, Option, Guide, Exit App. The main features which crucial for GAAR implementation which are anatomy AR, book anatomy and quiz game about anatomy to allow for better understanding about the structure of human skeleton and several main organs including their main function which suitable for young students to understands more about human science anatomy by providing 3D model which resemble to the actual organ and skeleton. Follows are the steps taken during development in Phase 1 of GAAR methodology.

Step 1: choose the correct marker-less Wikitude SDK for Unity as in Fig. 4.


https://www.wikitude.com/download-wikitude-sdk-expert-edition-for-unity/

Wikitude SDK Expert Edition for Unity

Create powerful solutions with advanced augmented reality features.

Unity

The Expert Edition plugin for Unity enables high-end single and multiple Image, Object, and Cylinder Tracking, and Multiple Tracker AR experiences.

Latest Version: 9.7.1
2021-04-28 - 321,8 MB
Release Notes

Fig. 4. Choosing Marker-less Wikitude SDK.

Step 2: 3D modeling of human body using Autodesk Maya as in Fig. 5.

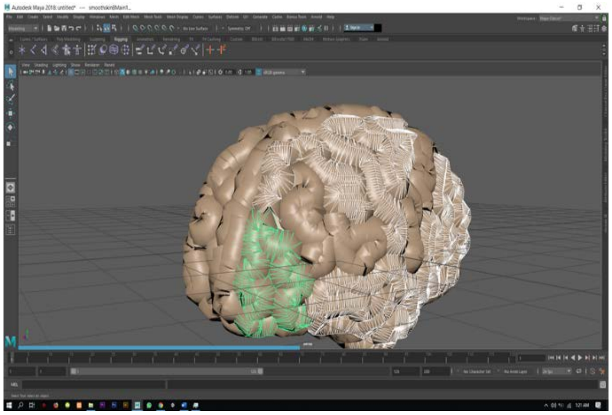


Fig. 5. Modelling in 3D.

Step 3: 3D model texturing in Autodesk Maya. Texturing is a process of making the surface model to look like real-world counterpart as in Fig. 6.

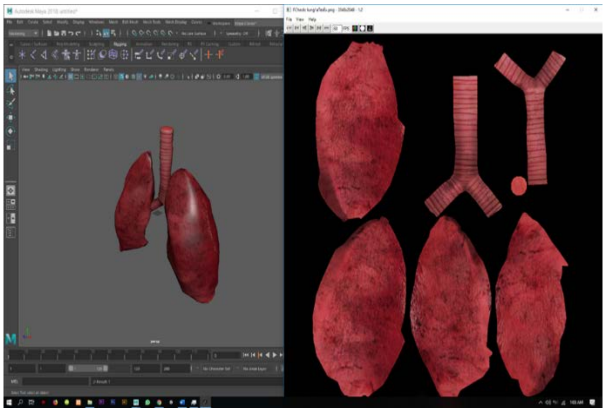


Fig. 6. Texturing.

Step 4: 3D model can be animated by rigging on a character as in Fig. 7. Rigging or skeletal animation means taking a static mesh to create an internal digital skeleton, and creating relationships between the mesh and the skeleton. This process also called as skinning, enveloping or binding process and adding a set of controls that the animator can use to push and pull the character around.

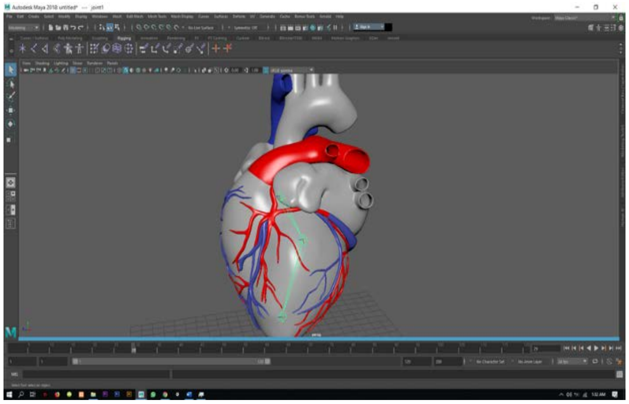


Fig. 7. Rigging.

Step 5: Building 3D model of human counterpart by importing into Unity3D software as in Fig. 8.

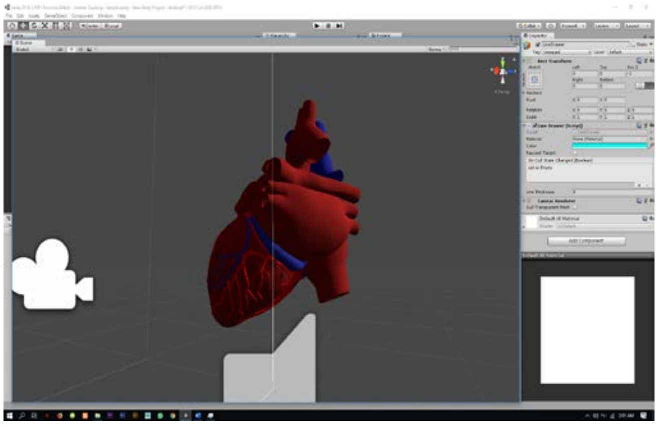


Fig. 8. Building 3D Model into Unity 3D.

Phase 2 is the programming or GAAR while Phase 3 is the combination of 3D models with AR where the *apk* files are created. Then in the last phase, the coding and programming are performed for GAAR mobile application.

B. *Interfaces*

Fig. 9 shows the home menu of the GAAR mobile application. When user first using the application, they will go to this home menu which consists seven buttons for user to explore. Fig. 10 shows the guide button feature. When user click on this button, they show the guide on what do each button state. As such, user can see which button that they need to know and be notify in early stage. These buttons guide with the symbol of boy reading manual, book section with the book symbol, quiz game with the symbol star. The middle button which is called Anatomy AR button which contain eight model of anatomy. The next three buttons is about application, option button and exit button.

Fig. 11 indicates the book section which contains seven buttons that indicated the several information about heart, lung, kidney, stomach, brain, liver and skeleton. Some model such as heart and lung have been provided with sound and animations. All the button give user the 3D model of anatomy and produce four main information about each button that suitable for young students and children. On the top, we have two button which are home button and Anatomy AR button. For the main part of GAAR mobile application is Anatomy AR button. Fig. 12 illustrates the interface of Anatomy AR. In this figure, the interface provides an information for start-up button which is used to explain the step that user needs to know before using Gross Anatomy AR. Then, once clicking on the initialize button, user will be directed with eight different icons which represent the 3D model organs. The GAAR contains eight different 3d models which are heart with animation, lung with animation, skeleton system, liver, kidney, brain, stomach and intestines (refer to Fig. 13). User also be notice with guides when using Anatomy AR as shown in Fig. 14. This small guide informs them on what user should do if user want to zoom, move, and rotate the 3d object. If the screen is full of 3d models, they can easily reset it all by pressing Reset button above go to Home page.

In Fig. 15, a 3D skeleton system model is displayed on the screen without any marker. The model also displays the name of bones and description about it. For each model, user have privileges to rotate using arrow provided or zoom by using two finger or move the model to another location. Next is Quiz game as stated in Fig. 16. The GAAR mobile apps provides simple questions which can test user memory based on the information that book section. Based on the figure above, there are nine multiple answer questions. Each question has its own countdown with test user capabilities to answer questions with at short timing. Next button act as submit answer after user pick an answer or not that allows user and audience know either the answer is true or false.

After finish answering those nine questions, the game shows the score and notify them the highest score of the previous challenge. If user want to try again, user can do this by pressing play again button or go to homepage at the home button below high score (see Fig. 17). Fig. 18 shows the graphic and game volume for GAAR mobile application. As for the graphic, it provided user low, medium and high resolution of pictures in the app. Game volume can be adjust either high or low depends on user. Hit apply button to set the options or hit back button to go to home menu. Lastly, the exit button used for user to check out from GAAR mobile application. In the button, it gives user the privilege for user to exit from using the app or stay using it (see Fig. 19).

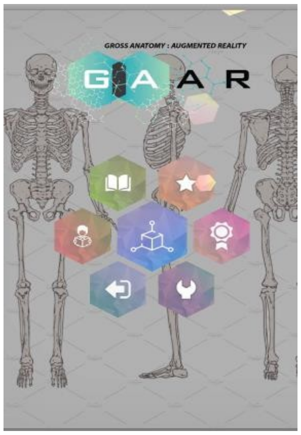
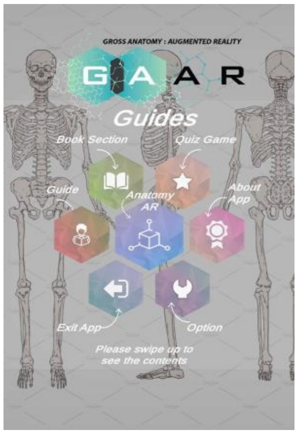
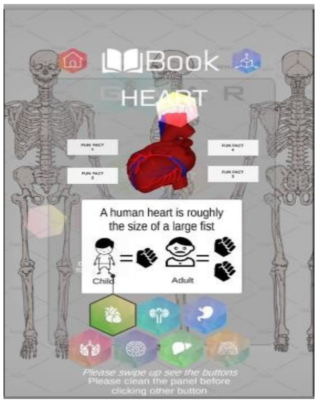


Fig. 9. Main Menu.



Please swipe up to see the contents

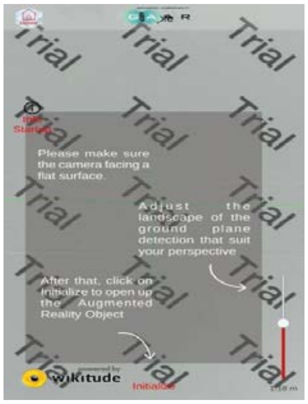
Fig. 10. Guide Buttons.



A human heart is roughly the size of a large fist

Please swipe up see the buttons
Please clean the panel before clicking other button

Fig. 11. Book Section.



Please make sure the camera facing a flat surface.

Adjust the landscape of the ground plane detection that suit your perspective

After that, click on Initialize to open up the Augmented Reality Object

Fig. 12. Guide to use GAAR.

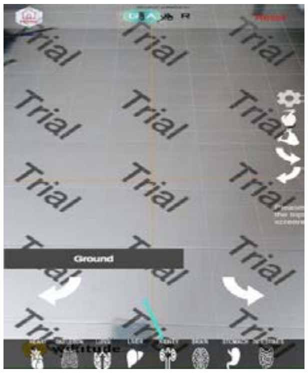
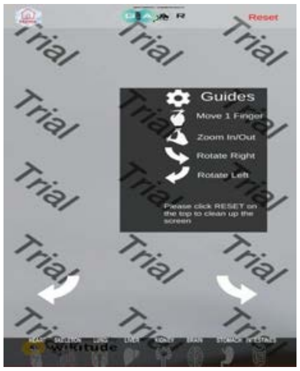
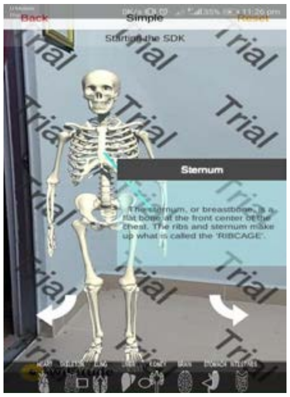


Fig. 13. GAAR Anatomy.



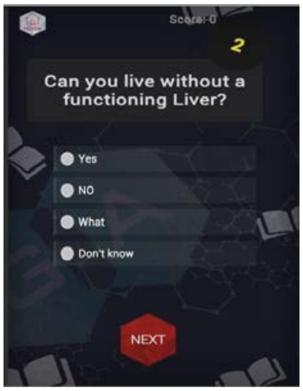
Please click RESET on the top to clean up the screen

Fig. 14. Guide in Anatomy AR.



The sternum, or breastbone, is a flat bone at the front center of the chest. The ribs and sternum make up what is called the 'RIBCAGE'.

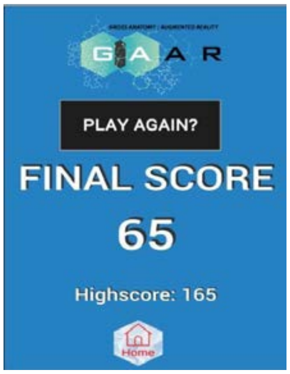
Fig. 15. AR Skeleton.



Can you live without a functioning Liver?

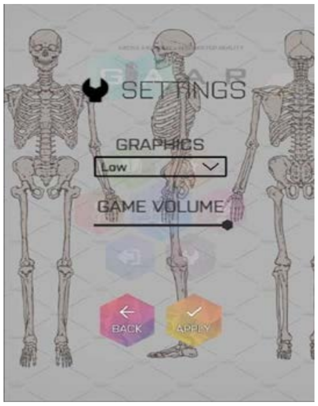
- ○ Yes
- ○ NO
- ○ What
- ○ Don't know

Fig. 16. Quiz and Game Interface.



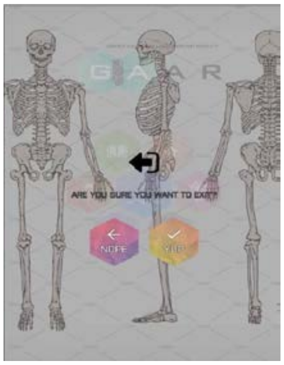
FINAL SCORE
65
Highscore: 165

Fig. 17. Score.



GRAPHICS: Low
GAME VOLUME

Fig. 18. Setting/Options.



ARE YOU SURE YOU WANT TO EXIT?

Fig. 19. Exit.

V. Discussion

The advantage of a virtual human anatomy is that the gross appearance of the human anatomy drives user to explore for more since user can easily see the model in 360 degree from any angle and observe the shape and where it is located inside a human body. The color, shape, animation and sound give user a real-life experience that they explore human body in reality. Based on the results of our research an interactive mobile application can be created to supplement the traditional descriptive course. This form of application has the advantage that the student can access it from home. Factors such as the resolution of the camera, or the mobile phone have an impact on the in full color appearance but they do not cause deformation of the virtual object.

VI. Conclusion and Future Recommendations

With this application, the medical lectures, laboratory exercises and the anatomical models, the targets of medical curriculum are accomplished. Because anatomical models and academic lessons is tough to study by oneself, perhaps with GAAR, they are capable of using their own mobile device to interact with 3D model. This GAAR is a mechanism to build interest in human anatomy with immersion. There are a few suggestions that can be used to upgrade GAAR mobile application to be more efficient in future. Our recommendation is to use a paid license version of Wikitude SDK to get rid of the trial watermark in AR camera. Others would be to add more sound to give impact to user when using the application when studying, to improve with a consistent layout of each scene in GAAR mobile app to make it more appeal to user experience, to add more games or interesting activities that can increase user understanding and allow them to memorize the info through game and fun activity. Also, this GAAR could provide more unusual and unseen info but remains a crucial fact about human internal and external body.

The use of AR in education is regards to bridge the digital divide and to provide the static pictures into real-like 3D animation. Through the real visualization, small to adult learners can imitate the real truth on human organ and how this can motivate them to take care of their bodies that would lead to a healthier living styles as well as easy memorizing of the subject contents [20-21]. The GAAR is a mobile application that is useful for all generations of students as well as teachers which give more interactions to explore human body by using smartphone in real environment. Student does not need to understand something about human anatomy since they can now use the current textbook and explore it using GAAR. This not only help students to increase their knowledge but also help teacher explain more about human anatomy in their science subject. It is anticipated that by using GAAR, young student can achieve more knowledge and see much more than just using textbook only.

ACKNOWLEDGMENT

A special thanks to the financial support under UMT Grant with Vot: 53290 and UniSZA grant with code UniSZA/2018/PKP/01 (An intelligent mosquito home system). A sincere gratitude goes to the AMHS co-founder, One Team Network Sdn. Bhd. and Dr. Mos Healthcare Sdn. Bhd. for the idea creation and all faculty members in supporting our work.

REFERENCES

[1] F. S. Motlagh & J. A. Pour, “Designing a model for developing students' needs skills of high schools for using virtual learning,” International Journal of Electronics Communication and Computer Engineering, 2014, 3(6), pp. 1444-1448.

[2] S. Standring, Gray's anatomy: the anatomical basis of clinical practice, expert consult (4st Ed). Spain: Churchill Livingstone, 2008.

[3] K. Romanov & A. Nevgi, “Do medical students watch video clips in e-Learning and do these facilitate learning?,” Med Teach, 2007, vol. 29, pp. 484-488.

[4] S. Shantikumar, “From lecture theatre to portable media: students’ perceptions of an enhanced podcast for revision,” Med Teach, 2009, 31, pp. 535-538.

[5] D. A. Back, N. Haberstroh, A. Antolic, K. Sostmann, G. Schmidmaier & E. Hoff, “Blended learning approach improves teaching in a problem-based learning environment in orthopedics - a pilot study,” BMC Medical Education, 2014, 14, pp.17.

[6] M. A. Aziz, J. C. McKenzie, J. S. Wilson, R. J. Cowie, S. A. Ayeni & B. K. Dunn, “The human cadaver in the age of biomedical informatics,” Anat Rec (New Anat), 2002, 269, pp. 20-32.

[7] H. Petersson, D. Sinkvist, C. Wang & O. Smedby, “Web-based interactive 3D visualization as a tool for improved anatomy Learning,” Anat Sci Ed, 2009, 2, pp. 61-8.

[8] J. Li, L. Nie, Z. Li, L. Lin, L. Tang & J. Ouyang, “Maximizing modern distribution of complex anatomical spatial information: 3D reconstruction and rapid prototype production of anatomical corrosion casts of human specimens,” Anat Sci Ed, 2012, 5, pp. 330-9.

[9] U. Von Jan, C. Noll, M. Behrends & U. Albrecht, “mARble - augmented reality in medical education,” Biomed Tech 2012, 18, pp. 67-7.

[10] H. Hazidar & R. Sulaiman, “Visualization cardiac human anatomy using augmented reality mobile application,” IJECCE, 2014, 5(3), pp. 497-501

[11] T. Blum, V. Kleeberger, C. Bichlmeier & N. Navab, “Mirracle: an augmented reality magic mirror system for anatomy education,” Virtual Reality Short Papers and Posters, 2012, pp. 169-170.

[12] R. Thomas, N. John & M. Delieu, “Augmented reality for anatomical education,” J Vis Commun Med, 2010, 33, pp. 6-15.

[13] M. Schoonheim, R. Heyden, J.M. Wiecha, “Use of a virtual world computer environment for international distance education: lessons from a pilot project using Second Life,” BMC Medical Education, 2014, 14, pp. 36.

[14] S. B. Issenberg, W. C. McGaghie, E. R. Petrusa, G. D. Lee & R. J. Scalese, Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005; 27: pp.10-28.

[15] Z. A. Galchenko,. "Six top tools to build augmented reality mobile apps." *Infoq. com Retrieved From: https://www. infoq. com/articles/augmented-reality-best-skds*, 2018.

[16] F. Mauro, et al. "Augmented reality as a new media for supporting mobile-learning." *Virtual and Augmented Reality: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2018, pp.1625-1643.

[17] K. Iggy, and D. Cummings. "History and Evolution of the Android OS." *Android on x86*. Apress, Berkeley, CA, 2013, pp. 1-8.

[18] YM Dr. Tengku Fadilah Bt. Tengku Kamalden, 2013, Akademik Kejurulatihan Kebangsaan, Nota Sains Sukan Tahap 1, Skim Persijilan Kejurulatihan Kebangsaan (SPKK).

[19] Kiourexidou, Matina, et al. "Augmented reality for the study of human heart anatomy." *International Journal of Electronics Communication and Computer Engineering* 6.6 (2015): 658.

[20] D. Schwarz, P. Štourač, M. Komenda, H. Harazim, M. Kosinová, J. Gregor, R. Hůlek, O. Smékalová, I. Křikava, R. Štoudek & L. Dušek, “Interactive algorithms for teaching and learning acute medicine in the network of medical faculties MEFANET,” J Med Internet Res, 2013, 15(7), pp. e135.

[21] P.E. Antoniou, C.A. Athanasopoulou, E. Dafli, P.D. Bamidis, “Exploring design requirements for repurposing dental virtual patients from the web to second life: a focus group study,” J Med Internet Res, 2014, 16(6), pp. e151.

GRASP Combined with ILS for the Vehicle Routing Problem with Time Windows, Precedence, Synchronization and Lunch Break Constraints

Ettazi Haitam[1], Rafalia Najat[2], Jaafar Abouchabaka[3]
LaRI Laboratory, Faculty of Sciences, University Ibn Tofail[1,2,3]
Kenitra, Morocco

***Abstract*—In this era of pandemic especially with COVID-19, many hospitals and care structures are at full capacity regarding availability of beds. This problem leads to ensure giving specific cares to people in need either in illness or disability in their own homes. Home Health Care (HHC) proposes this kind of services for patients demanding it. These services have to be done at the request of the patient which appears to be the client in a way that gives satisfaction to the requester of the service. Often, these demands are bound by a specific time that the workforce (caregivers) are obligated to respect in addition to the precedence (priority) constraint. The main purpose of the HHC structures is to provide a service that is good in term of quality, minimize the overall costs and shorten the losses. To reduce the costs of these HHC structures, it is mandatory to find comprehensible and logical ways to do it, for it is not permissible to touch the caregiver salary, HHC structures find themselves in the obligation to optimize by other means such as reducing the travel cost. Note that these structures give cares in one's home, which means that the travel aspect is important and is considered the core spending charges of the institution. Another fact is the satisfaction of the patients toward caregivers; this is an essential element to optimize in order to obtain a good quality service, to give a realistic aspect for the problem the lunch break of caregivers is introduced as a parameter. For those arguments, a conception of an efficient planning of caregivers involves using decision tools and optimization methods. A caregiver (vehicle) is attributed to a patient (customer) to do a number of cares with several options in accordance to the customer wishes like time windows requirement often specified by the client, the priority or precedence constraints are usually performed if a care have to be performed before another and could need the intervention of more than one caregiver and must have at least one lunch break a day and it is not always taken at a set time of the day and must be versatile to optimize customer demand satisfaction. To resolve this issue which is called VRPTW-SPLB, a mathematical model of the problem is proposed and explained as a Mixed Integer Linear Programming (MILP) and a greedy heuristic based on a Greedy Randomized Adaptive Procedure (GRASP) is proposed, two strategies based on local search and two metaheuristics, and a metaheuristic resultant of an hybridization of the two metaheuristics. At the end of the paper, results are shown on a benchmark extracted from the literature.**

Keywords—Optimization; VRP; home health care; ILS; tabu search; metaheuristics

I. Introduction

The world nowadays know a considerable increase in growing rate of population along with the several diseases that emerges day by day leading to a decrease in the number of beds in hospital per person. This situation arises many problems such offering adequate cares to needy people without delay. The Home Health Care services (HHC) becomes more complex and must adapt and evolve to cover all this variables by applying decisions in operational research context. To identify the source of the problem and its solution, defining the major actors is essential, these services are made essentially for disabled/elderly people that request many needs like medical care, hygienic care, assistance... In order to fulfill these demands, HHC structures offer the opportunity to perform such care in patient's homes along with the adequate and necessary equipment and resources may be caregivers or devices and gives the patient's the feeling that they are treated like if they receive treatment in a traditional hospital. This overall situation obligate to create a fine and good collaboration and coordination between human and material resources in order to establish a good and optimized planning that increases the quality of services offered by home health care with a controlled and reduced generated costs.

In this study, a generalization of the classical problem known as vehicle routing problem is proposed to deal with special and particular aspects linked to HHC. The problem studied in this paper consists of generating a set of routes executed by a group of vehicles (qualified caregivers) in order to visit a number of dispatched clients (patients) while satisfying a set of preferences aspired by the patients towards caregivers so that the operation is done in the most effective way. To enrich this study, the cares (2 or more) are of two categories according to the synchronization aspect, either simultaneously (at the same time) or in a given order(precedence), this case need the coordination of a high level of caregivers. To give the problem a realistic aspect that most of the studies neglect in their work is by considering the lunch break of caregivers which are operated only after the termination of a care by a caregiver.

To summarize what is being said, the problem studied is called the Vehicle Routing Problem with Time Windows, Synchronization, Precedence and Lunch Break constraints (VRPTW-SPLB). The constraints applied in this generalization of the VRP is appropriate to the maximum as it is in real life where most of patients have a preferred time to not precede or exceed, two or more cares can be done by more than one caregiver according to the skills of this latter, the cares can be one after another according to the priority of care and finally caregivers need the replenish their energy to continue the cares in the most efficient way thus the need to have a lunch break granted to the caregiver.

The next sections of this article are structured as follows: a section is devoted to the literature review of past works dealing with the vehicle routing problem in the home health care sector, the next section is dedicated to the problem definition, afterward a mathematical formulation is proposed in Section 4, next the explanation and components of the different metaheuristics used to solve the problem. Section 6 illustrates the computational results, while Section 7 is mainly for the discussion. Finally a conclusion and future perspectives are displayed.

II. Literature Review

The vehicle routing problem is regarded as an NP-Hard problem and considered within the field of Operational Research as one of the most studied in the last two decade. In order to assimilate a valid and explicit explanation of the vehicle routing problem as proposed by [1] is to build a route delivery with the least possible cost from a center (depot) to a series of distributed locations (customers) in accordance to a set of constraints that need to be respected. Both exact and approximate methods or approaches are considered as solution to the problem, only that metaheuristics methods are plainly approved to be effective as illustrated in [2] which is up to find a solution up to 26 variants of the VRP. The main body of this study is to plan and allocate several resources such as vehicles (caregivers), material... to a set of demanding requests made by a group of clients (patients) dispatched in a geographical region to achieve a predefined and specific services. From that perspective, the arousal of many constraints seem natural such as fixed time windows determined by the patient, preferences of patients toward caregivers, synchronization of services either simultaneously or not, lunch break of caregivers... Many approaches set by many authors such as [3-7] have realized and confirmed that the best methods to solve this problem is by using heuristics and metaheuristics. These authors experimented their approaches on qualified and known benchmark of the literature offered by [8,9]. Putting all this aside, it is obligatory to search and evaluate the most common objectives (cost, preferences, waiting time...) considered in those studies. Considering the HHC context, the vehicle routing problem offers as a problem many objectives to optimize; the total travelled distance by vehicles, the number of vehicles required, the total travelled time, patient's preferences toward caregivers are the most studied as objectives to optimize and frequently studied separately, thus these objectives can be considered on together but very rarely.

The idea supported in transportation in the home care as a logistical issue can be clustered into two broad categories, allocation and planning issues and pickup and delivery issues.

In this paper, the rest of the literature proposed in this study will be separated in three subsections, studies and works illustrated in the past talking on the subject of vehicle routing problem in home health care sector are shown in first, the second is an extension of contribution linked to the aspect synchronization and precedence in the vehicle routing problem and the last section is a highlights of the current contribution.

A. VRP in HHC

The logistics factor was first demonstrated in the field of Health Services by [10-12] on door-to-door transportation to support the elderly and/or disabled. The increasing rate of elderly people worldwide has prompted the authorities to introduce a scheme called DARP that translates to Dial-A-Ride in order to assist people with disabilities of any type. From this perspective, [13,14] have established a particular version of DARP that seeks to deploy a special team with appropriate and specific expertise to the home of the patient to assist them with any support they request (bath, dress, help to move...) until the resource that is responsible for transporting them to the hospital or a treatment structure arrives. This scenario has prompted researchers to integrate constraints such as synchronization constraints to ensure that priority between visits is respected, in order to ensure the effectiveness of this procedure, fine teamwork is expected by medical assistants and transporters. In the same category, a new constraint emerges for patient requirements that allow them to select a time window in which they are prepared to receive the care staff in their homes. The author in [15] suggests and incorporates this feature of time windows in the DARP for improved planning, but does not take into account the state of emergency and is instead concerned with minimizing the cost of transporting patients. The Dial-A-Ride problem has been expanded into the HHCP (Home Health Care Problem) conducted by a study well known in the literature [16] in which the key concept is to delegate caregiver staff to provide a certain amount of care needed by clients (patients) in their homes. The study shows two methods of resolution, the first of which was an exact method by suggesting a mathematical model with inter mixed variables (MILP) and the other one is a heuristic method. As for applying these methods on real instances, 10 patients and 4 nurses were considered to do the test, this will be the initial step to various and many papers addressing the same study but by using and testing various optimization resources and tools.

The optimization context in the home health care sector considers many aspects. The logistical aspect as concerned with this study contains two main categories that can be illuminated regarding other categories such as planning and allocation issues and pickup and delivery issues which are considered to be the most illustrated aspect on this field. For this study to be comprehensive and for the sake of simplification of the problem, generally a patient asking for a required care service is allocated a caregiver according to adequate skill and qualification for doing and fulfilling the service. Those skills and qualification are medical qualification, age, gender, language, specialty... The caregiver has a begin time and an end time throughout the day. As for the service required, it is characterized by a duration, skill and time window.

The dilemma of attributing visits to personal caregivers and the preparation of visits was studied in[17], where a MILP methodology and a Tabu search method were implemented to solve the problem and deliver fair and high quality solutions. In the same line, [18] proposed the same approach but considered as a partitioning problem which goal is to maximize the satisfaction of the patients and to minimize the transportation duration. There are studies in the literature that considers a single objective optimization such as [19] where the time windows constraint was introduced therefore the problem became VRPTW (Vehicle Routing Problem with Time Windows) and proposed a metaheuristics based on VNS and a mathematical formulation (MILP) in order to shorten the time that nurses consume in their movements and to maximize and increase the ratio nurses/patients of satisfaction. Another study [20] aims to solve the vehicle routing problem with time windows but this time with another approach and another implementation by introducing the Particle Swarm Optimization (PSO) so that the total travelled distance by nurses and caregivers is reduced with respect of constraints of time windows and vehicle's capacity. The multi-objective optimization is another variant of dealing with optimization in HHC, considering two functions to optimize at the same time leading to a high complexity of the problem.

The study in [21] deals with this multi-objective optimization by applying two metaheuristics based on ACO (Ant colony Optimization) which is an algorithm based on population, a mathematical formulation for the linear optimization and a multi-objective genetic approach.

Coordination of visits between nurses and/or caregivers is considered in the study of [22,23]. This coordination is either simultaneously or in a given order. The authors proposed a Branch and Bound method in order to reduce the travel cost of caregivers and to increase and maximize the satisfaction of patients toward nurses/caregivers. As for waiting time, which is a significant matter to consider was addressed by [24] for the sake of planning and scheduling visits to patients and giving tasks and mission to nurses/caregivers. First the mathematical formulation was established. Some authors tackled this problem by solving it in stages and not fully at once. This was done by [25,26] where a set of routes were generated initially before assigning routes to caregivers then enforce synchronization constraints in order to reach feasible and concrete solutions. A problem of node duplication was addressed by [27] where to run away from this duplication of nodes, an Iterated Local Search (ILS) was implemented and applied on relatively small instances of the literature. As for [28], maximizing the satisfaction of patients was prioritized along with the reduction of the cost of travel by using a Variable Neighborhood Search (VNS).

Synchronization constraints are really important to highlight the VRPTW problem in its totality where a set of request made by patients can have multiple visits and require the intervention of one or several caregivers either simultaneously or in a predefined order. In [7], work on this by formulating at first a mathematical skull of the problem, a GRASP and several metaheuristics for the sake of optimizing the total travel cost and increase the satisfaction of patients.

The pick-up and delivery aspect is studied and highlighted as well in many papers where a patient is to be picked up from a place or dropped off in a place. This was highlighted by [13]. This aspect of picking up and delivering is flexible to the points that authors in [4] proposed a system in order to deliver meals to patient's homes. To do so, patients were clustered and partitioned in zones according to the care structures policy. Meanwhile caregivers were also grouped in teams.

Before jumping on the constraints used in general, detailing synchronization constraints on the home health care sector is significant to the continuity of this study. As for our knowledge, the synchronization aspect dealing with the home care services are either considered in a simultaneous way likewise [27]. Other authors considered both simultaneous and precedence (in a given order) in separate, this is clearly shown in [23] and [14] where [7] and [29] illustrated the synchronization in the same model. In our study, we will consider both synchronization aspect in addition to a set of constraints that were not combined in the same study before.

Table II shows the overall constraints of paper found in literature meanwhile Table I expresses the constraints that are the most common.

TABLE I. CONSTRAINTS COMMONLY USED IN HHC

Abbreviation	Description
TW	Time Windows
Simu	Simultaneous synchronization
Prec	Precedence synchronization
SS	Specific Service
LB	Lunch Break

TABLE II. CONSTRAINTS IN THE LITERATURE

Category	Article	TW	Simu	Prec	SS	LB
	[17]				X	
	[18]	X	X		X	
	[20]	X				
	[23]	X	X	X		
Planning & Scheduling	[29]	X		X	X	
	[25]	X	X			
	[27]	X	X		X	
	[30]	X	X	X		
	[31]	X	X	X	X	
	[16]	X			X	
	[15]	X				
Pick-up & Dekiveries	[32]	X	X	X		
	[33]	X				
	[34]	X				X
Both categories	[14]	X		X		

	[35]	X			X	
Our paper		X	X	X	X	X

As shown in Table II, the temporal aspect is what most of the authors are interested in for the aim to get a better quality of services. In the other hand, the constraints that give a realistic aspect to the problem are neglected in the most of the works such as lunch break of caregivers, precedence constraints and synchronization constraints.

B. *VRP with Synchronization Constraints in HHC*

When talking about timing and temporal constraints in the vehicle routing problem especially in the home health care sector, it is very needy to point at very important notion that enforce the realistic aspect of the problem overall. the study in [36] point in his classification that there is a dependence that takes place in the route taken by caregivers in order to get to patient's homes requesting cares, he defines it as synchronization. This starting point encourage other authors such as [27] to group and cluster this so-called synchronization into two main part, one for the simultaneous synchronization which is a two or more than one care is done at the same time, this type of synchronization require the intervention of more than one caregivers and is given a special attention (like transporting a patient from one place to another, bathing of a disabled patient...), the second type is more a synchronization that prioritize a care over another which means that two or more cares cannot be done simultaneously but need to be successive in a predefined order.

As for the literature and past works dealing with synchronization, we must cite [37] where he made the aircraft routing problem a simultaneous synchronization problem that ensure that a flight possessing an Id have to depart and take off in the same time at the same place. This resulted in an approach based on a decomposition called Dantzig-Wolfe decomposition.

Synchronization constraints proposed by [36] in the operation of collecting raw milk from farms consisted in a problem called Vehicle Routing Problem with Trailer and Transshipment (VRPTT). The problem is decomposed and constituted of two main vehicles trucks that can operate alone (autonomous) and trailers that can operate only with the help of a truck (non-autonomous). The authors conceived a mathematical formulation of the VRPTT and a simple heuristic. The synchronization used in this paper is a precedence one.

Precedence constraints in the home health care problems were addressed by multiple authors such as [38] where in the context of waste collection of a household the minimization of the overall travel duration is tackled. To ensure the fluidity of the work, vehicles were separated into two types, small ones called satellites that can drive on any type of road where the second one is large or compactors that have strict and known road to drive on. A local search procedure was used in order to find feasible solution of this problem.

C. *Contribution Highlights*

In this paper, the points addressed in the literature do not handle a majority of constraints in an efficient manner, where most of the authors push the search area of solutions to grow considerably by duplicating nodes. The second point that found its limitations as the instances grows is the use of models that are based on MILP (Mixed Integer Linear Programming) where only small instances give satisfactory solutions in the contrary of large instances. Authors decided then to use metaheuristic to work on larger instances, but fail in catching the optimal solutions and are trapped in local optimum because most of them don't use local search procedure in their work in order to avoid being trapped in a local optimum. For all these points, this paper aims to contribute in:

- Put a model that engage both simultaneous and precedence constraints, caregiver lunch breaks and time windows in the same design. Afterwards solve this model with a linear solver such as Cplex.
- Manage to work with multiples and several constraints such as time windows of patients, synchronization constraints with their two type (simultaneous/precedence), caregivers skills and qualification, caregivers lunch break and the preference of patients in regard of caregivers. It is very rare to find authors that handle more than three constraints at the same time in their work.
- Propose a hybrid algorithm in order to find feasible solutions of instances considered large on existing benchmark of literatures.

III. Problem Description and Mathematical Formulation

The vehicle routing problem is a problem that considers the determination of routes that must be taken by vehicles from a point i to a point j. The variant of the VRP considered in this study can be considered as a vehicle routing problem with time windows, synchronization, precedence and lunch break constraints (VRPTW-SPLB).

Consider G= (V, E) a graph where N = {1… n} is a set of clients (patients) and V= NUD a set of nodes. c is the initial depot and f the final one where D = {c, f}. Routes are linked by a set of points where m = [i, j] a margin, C_{ij} is the cost linked to the path ij and T_{ij} the duration travel. A health structure offers multiple services S = {1…s} to a set of patients where a patient i require a subset of services to the health structure $S_i = \{s \in S: e_{is} = 1\}$. A duration d_{is} is assigned to each service asked by a patient along with a time windows $[a_i, b_i]$ wherre a service cannot begin before a_i and cannot terminate afterb_i. The set of vehicles required to ensure the task of providing services to patients is denoted by K where each vehicle $k \in K$ have a time windows of availability specified by α_k, β_k. Prf_{ik} a number ensuring the preference of each vehicle k, this number defines the non-preferences of patient i in regard of the caregiver k.

P describes the lunch break of caregivers and nurses and L is the duration and $[w_p, z_p]$ is the time window as defined by [16]. g_{ik} is 1 if the caregiver consumes his/her break at patient i before service termination and 0 otherwise, g'_{ik} is 1 if the caregiver consumes his/her break at patient i after service termination, 0 otherwise. st_{ik} is the service start time of caregiver k at patient i and st_{pk} is the lunch break time start of the caregiver k. For the sake of simplicity, we write $k \in K_s$ if y_{ks}=1 and $s \in S_i$ if $e_{is} = 1$.

For the simplification and generalization of this study, we suppose that $k \in K$ accomplishes a unique service s. y_{ks} equals 1 if vehicle k offers the service s, 0 otherwise ($\forall\ k \in K, \forall\ s \in S$).

For customers (patients) whom are demanding more than one service either in a specific order or simultaneously, we define for each patient $i \in N$ a $slot_{isr}$ as the time separating the time start of service s and service r demanded by patient i where services s and r have to be provided in a given order (service s before service r).

Just to note that lunch breaks are not depending entirely on the time duration of the route, and a lunch break can be taken and utilized if the duration of the route outpaces and exceeds α_p. It is also permissible to take a lunch break at the end of a route even if the time window of the lunch break is not yet covered.

This problem formulation considers the minimization of the sum of non-preferences of patients towards caregivers and total travelling time by shortening to the maximum the total travelled distance overall.

To generalize the constraints of time in this formulation, we presume that all the requested services by patients are asked independently from the initial depot (center) c and their duration is 0 ($D_{cs} = 0$, $\forall\ s \in S$).

The Mixed Integer Linear Programming corresponding to our formulation is as follows:

$$\min \sum_{i \in V\setminus\{f\}} \sum_{j \in V\setminus\{c\}} \sum_{k \in K} C_{ij}\, x_{ijk} + \sum_{i \in N} \sum_{j \in V\setminus\{c\}} \sum_{k \in K} Prf_{ik}\, x_{ijk} \qquad (1)$$

Subject to:

$$\forall k \in K, \sum_{j \in N} x_{cjk} = 1 \qquad (2)$$

$$\forall k \in K, \sum_{j \in N} x_{ifk} = 1 \qquad (3)$$

$$\forall h \in N, \forall k \in K, \ \sum_{i \in V\setminus\{f\}} x_{ihk} = \sum_{j \in V\setminus\{c\}} x_{hjk} \qquad (4)$$

$$\forall i \in N, \forall s \in S, \ \sum_{j \in V\setminus\{c\}} \ \sum_{k \in K_s} x_{ijk} = e_{is} \qquad (5)$$

$$\forall i, j \in V, \forall s \in S: s \in S_i \cup S_j, \forall k \in K_s, st_{ik} + (T_{ij} + D_{is}) x_{ijk} \le st_{jk} + b_i (1 - x_{ijk}) \qquad (6)$$

$$\forall i \in N, \forall s \in S_i, \forall k \in K_s, a_i \sum_{j \in N} x_{ijk} \le st_{ik} \le b_i \sum_{j \in N} x_{ijk} \qquad (7)$$

$$\forall k \in K, \alpha_k \le st_{ck} \le \beta_k \qquad (8)$$

$$\forall k \in K, \alpha_k \le st_{fk} \le \beta_k \qquad (9)$$

$$\forall i, j \in V, \forall k \in K, x_{ijk} \in \{0,1\} \qquad (10)$$

$$\forall i \in V, \forall k \in K, \ g_{ik}, g'_{ik} \in \{0,1\} \qquad (11)$$

$$\forall i \in V, \forall k \in K, \ st_{ik} \in \{0,1\} \qquad (12)$$

The objective function (1) is to minimize the total travelling time and the sum of non-preferences. Constraints (2) and (3) ensure that each vehicle have to leave the initial depot and return to it. Constraint (4) guarantees the continuity and constancy of the routes while constraint (5) guarantees that the demands of customers (patients) are provided and executed. Scheduling and planning permitting the consistency between the durations of visits is maintained by constraint (6). Constraints (7), (8), (9) guarantee the respect of patients, caregivers time windows. Constraints (10), (11) and (12) define the nature of the decision variables.

Just to cite that the caregivers' lunch breaks are not entirely related to the time duration of the route taken, and a lunch break can be used on the condition that the duration of the route outpaces and exceeds α_p. It is also permitted to take a lunch break at the end of a route even if the time window of the lunch break is not yet covered.

IV. Resolution Approach

A metaheuristic is developed due to the challenge of solving large instances in an acceptable amount of time even though the proposed model decreases the graph size in comparison of prior formulations. The metaheuristic proposed is a hybridization of GRASP and ILS.

The GRASP method was firstly presented and dealt with by [39] for a covering set problem. The GRASP method was effective for construction routing problem, when [40] utilize it in order to solve a tour design problem with time windows to minimize the vehicle number used.

The GRASP method consists of iteration which produces in each iteration a new solution in two phases:

- Construct a feasible solution by the means of a greedy randomized heuristic.
- Improve the solution with the help a local search procedure by the iterative exploration of search space in order to improve the current solution.

The Iterated Local Search (ILS) belongs to a class of metaheuristic based on the exploration of a neighboring of near local optimum. The local optimum according to [41] is grouped in a cluster of search space. ILS initiates its search from a starting solution called initial solution of a good quality. This initial solution is created by the help of a greedy heuristic, and then improved by a local search to find a first local optimum. In the contrary of the GRASP, then a perturbation procedure is done on the current solution in each iteration in the hope to escape from the local optimum. When the solution is perturbed and improved by the local search, then this time the new solution becomes another local optimum. This cycle composed of perturbation and local search will not stop until a stopping criterion is met.

In the hybrid version, the ILS procedure replaces the local search of the GRASP method. The general structure of the GRASP combined with ILS is proposed to solve the VRPTW-SPLB is illustrated in the algorithm 1. Even tough, the algorithm is simple; its capacity to solve multiple combinatory problems is proven such as [42] for the localization routing in two echelon and [43] for the periodic vehicle routing problem with time windows.

The principle loop in the algorithm 1 (line 3-32) defines the hybrid method GRASP*ILS composed of a construction phase and an improvement phase done by the iterated local search (ILS). The first solution Sol_1 is generated by a parallel randomized constructive heuristic (PCH). In each iteration of the ILS algorithm (line10-25) calls the perturbation procedure (PP), then the random local search updates the last solution encountered (Sol_1) when this latter is improved. In order to stop the algorithm, stop conditions must be met such as the maximum number of iteration $\mathrm{Iter}_{\mathrm{MaxG}}$ is met. The second stop condition is when an inferior margin LM after a number of iteration $\max_{\mathrm{Failure}}$ counted by the variable NImp where there is no improvement of the current solution or just after a maximum number of iteration $\mathrm{Iter}_{\mathrm{MaxL}}$.

Algorithm: GRASP combined with ILS

```
1:  F* ← +∞
2:  $Iter_G$= 0
3:  **While** (($Iter_G$ < $Iter_{MaxG}$) and (F* > LM ) **Do**
4:      $Sol_1$← PCH($Sol_1$)
5:      **If** ($Sol_1$ $is\ feasible$) **Then**
6:          $Iter_G$ = $Iter_G$ + 1
7:          $Iter_L$= 0
8:          $NImp$ = 0
9:          $Sol_1$← LocalSearch($Sol_1$)
10:         **While** (($Iter_L$ < $Iter_{MaxL}$) and ($NImp$ < $max_{Failure}$) **Do**
11:             Sol ← $Sol_1$
12:             Sol ← PP(Sol)
13:             **If** (Sol is feasible) **Then**
14:                 $Iter_L$ = $Iter_L$ +1
15:                 Sol ← LocalSearch (Sol)
16:                 **If** (F(Sol) < F($Sol_1$)) **Then**
17:                     $Sol_1$← Sol
18:                     $Iter_L$ ← 0
19:                 **Else**
20:                     $NImp$ = $NImp$ + 1
21:                 **End If**
22:             **Else**
23:                 $NImp$ = $NImp$ + 1
24:             **End If**
25:         **End While**
26:         **If** (F($Sol_1$) < F*) **Then**
27:             S* ← $Sol_1$
28:             F* ← F($Sol_1$)
29:             $Iter_G$ ← 0
30:         **End If**
31: **End While**
32: **Return** Sol*
```

Just to note that F(Sol) corresponds to the cost of the solution Sol.

V. Numerical Evaluation

A. Instances and Implementation

The metaheuristics proposed in this paper were coded in Python3.8 and were executed and tested on a machine with a processor Core i7-5500 CPU with 8GB of RAM, functioning under Windows 10 Home as for the mathematical model, the test was done with CPLEX Studio version 12.10.

The benchmark used as instances are from [23], which is known for its features that includes all aspect studies in this paper such as: Time Windows, Synchronization and Preferences.

This benchmark contains three parameters; the number of customers is equivalent to patients in our case, the number of vehicles equivalent to caregivers and are grouped according to this scheme:

- 18 clients (patients) with the use of 4 vehicles (caregivers),
- 45 clients (patients) with the use of 10 vehicles (caregivers),
- 80 clients (patients) with the use of 16 vehicles (caregivers),

To incorporate the notion of synchronization, a range number linked to synchronization from two to four is determined, for the sake of generalization lunch break is always taken after performing successfully a service s. Time windows are grouped by small, medium and large. In this study, a limit was put to not exceed at most two requested services.

B. Algorithms Parameters

The main goal for setting parameters is to obtain the best performances by testing a limited number of parameters combination. For the sake of this study, the logical practice for choosing parameters is to parameter each algorithm used separately according to its suitability. For the hybrid method GRASP*ILS, parameters set to ensure the best testing possible are the number of restart $\mathrm{IterMax}_{\mathrm{G}}$, the iteration number without success $\mathrm{Max}_{\mathrm{fail}}$, the call number of ILS $\mathrm{IterMax}_{\mathrm{L}}$ and finally the perturbation procedure parameter $\mathrm{pert}_{\max}$ related to the number of patients to remove. The average number of local search executed in order to obtain the best solution for the hybrid version is restored and will be given as a stop condition for the hybrid method GRASP*ILS, all this is done by adding a counting variable that is incremented in each call of the local search procedure (AVD). The metaheuristics proposed in this paper are executed a number of 10 times and the best solution along with the average solution are given in comparison. The Table III states the parameters used for testing, each parameter is given a range value of testing and the best value is taken as a fixed parameter.

TABLE III. PARAMETERS FOR GRASP*ILS

Parameters	Notation	Test values Range	Value
Number of restart	$IterMax_G$	[10 .. 100]	45
Iteration number of ILS	$IterMax_L$	[10 .. 100]	45
Number of solution without improvement	Max_{fail},	[5 .. 50]	35
Perturbation level	$pert_{max}$	[1 .. 6]	3

C. Results and Interpretation

The results of experimentation of the 37 instances of each metaheuristic are found in the end of the paper, as for the summary results, they are given in Table IV and Table V. In order to have a better analysis of the obtained results, four comparisons are highlighted, one comparison for the metaheuristic GRASP*ILS with the CPLEX results, the second comparing the results of GRASP with the results of the ILS, the third comparing the GRASP with the CPLEX and finally the ILS with CPLEX. As for the representation of the results, the results are given for each group of instances (Gins_1 , Gins_2 , Gins_3). For CPLEX, instances that give feasible solution are denoted by (G_{cf}) and for all instances ($\mathrm{G}_{\mathrm{all}}$).

To understand the Table IV, the first part are the results obtained by the CPLEX solver, then the second part of the Table IV and the two parts of Table V highlight the results of the proposed methods. The performance indicators used to compare the different results are the average gap percentage of the best found solution of the GRASP*ILS method for the lower bound ($\mathrm{Ind}_{\mathrm{LB}}$), the upper bound ($\mathrm{Ind}_{\mathrm{UB}}$) along with each criteria separately ($\mathrm{Ind}_{\mathrm{Moving}}$) and ($\mathrm{Ind}_{\mathrm{Pref}}$). Afterwards, an exposition in seconds of the time required to obtain the best solution ($\mathrm{Tsec}_{\mathrm{min}}$). The average solution of the different executions are compared with the best solution of the results obtained by the GRASP*ILS method ($\mathrm{CObj}_{\mathrm{Ref}}$, $\mathrm{CMoving}_{\mathrm{Ref}}$) and $\mathrm{CPref}_{\mathrm{Ref}}$). Finally, the average computational time needed to obtain the average solutions is exposed in seconds denoted by ($\mathrm{Tsec}_{\mathrm{average}}$), as for the optimal solutions, we denote it ($\mathrm{Tsec}_{\mathrm{optimal}}$).

In the next tables, note that the results subject to the CPLEX solver which is there are no feasible solution are not included (Max time of execution is 1 hour).

Note that Ref=GRASP*ILS and X= Total cost of the objective function.

The CPLEX studio solver permitted to solve up to 23 optimal solution of the 37 instances tested. Meanwhile the hybrid method has solved all instances in a more reduced time (23,93 seconds for the GRASP*ILS compared to 609,53 for CPLEX).

In general, the overall results obtained show that the GRASP*ILS method is way better than CPLEX results with an average gain of 2,10%, That means that this method attain an average gap of 3,19% for the lower bound in 162,24 seconds. This gap is reduced more than 1% in comparison with the CPLEX results (4,47% of gap in 1414,51 seconds) in a computational time considerably small.

The efficiency of the hybrid method GRASP*ILS is confirmed by observing the results of the third instance group (Gins_3) where the hybrid method improve by 12,95% in 1186,43 seconds compared to the upper bounds of the CPLEX where the gap is nearly 23% in 3600 seconds.

Overall, the solutions presented by the hybrid method GRASP*ILS are better than the upper bound given by the CPLEX solver in 3600 seconds. In addition, the nature of the problem being NP-Hard make the CPLEX solver incapacity of producing feasible solution in a reasonable time nearly impossible for instances of more than 45 patients. In the other hand, GRASP*ILS work perfectly on larger instances.

After summarizing and analyzing the results given by the CPLEX solver and the hybrid method, an evaluation of the efficiency of the GRASP*ILS method is highlighted and compared to both the GRASP method and the ILS method.

TABLE IV. PERFORMANCE INDICATORS FOR THE 37 INSTANCES OF MILP AND GRASP*ILS

Indicators	**MILP**					**GRASP*ILS**				
	$Gins_1$	$Gins_2$	$Gins_3$	G_{cf}	G_{all}	$Gins_1$	$Gins_2$	$Gins_3$	G_{cf}	G_{all}
Ind_{LB}	0,00	2,25	22,90	4,47	4,47	0,00	2,01	11,87	3,19	3,90
Ind_{UB}	-	-	-	-	-	0,00	-0,31	-12,95	-2,10	-2,10
Ind_{Moving}	-	-	-	-	-	0,00	-0,48	-3,13	-0,36	-0,36
Ind_{Pref}	-	-	-	-	-	0,00	0,15	-5,11	-0,84	-0,84
$Tsec_{min}$	86,81	2002,85	3601,96	1414,51	1480,80	1,78	66,41	1186,43	162,24	344,67
CX_{Ref}	-	-	-	-	-	0,03	3,03	2,73	1,93	343,12
$CMoving_{Ref}$	-	-	-	-	-	0,04	-0,38	0,49	0,04	0,04
$CPref_{Ref}$	-	-	-	-	-	-0,11	0,57	0,64	0,32	0,33
$Tsec_{optimal}$	-	-	-	-	-	0,48	102,8	981,71	165,92	302,33
$Tsec_{optimal}$	86,81	1287,59	-	609,53	609,53	0,39	57,39	-	23,93	23,93

TABLE V. PERFORMANCE INDICATORS FOR THE 37 INSTANCES OF ILS AND GRASP

Indicators	MILP					GRASP*ILS				
	$Gins_1$	$Gins_2$	$Gins_3$	G_{cf}	G_{all}	$Gins_1$	$Gins_2$	$Gins_3$	G_{cf}	G_{all}
Ind_{LB}	0,39	6,19	9,83	4,74	4,98	0,00	4,94	17,28	5,53	6,43
Ind_{UB}	0,39	3,87	-15,33	-0,65	-0,65	0,00	2,68	-3,35	0,59	0,59
$\text{Ind}_{\text{Moving}}$	0,23	-0,08	-4,69	-0,67	-0,67	0,00	1,22	-1,53	0,28	0,28
Ind_{Pref}	-0,04	0,67	-1,18	0,02	0,02	0,00	-0,29	0,70	-0,03	-0,03
IndX_{Ref}	0,39	4,13	-2,39	1,44	0,95	0,00	2,93	6,17	2,51	2,72
$\text{IndMoving}_{\text{Ref}}$	0,23	0,33	-0,57	-0,95	0,00	0,00	1,68	2,73	-0,04	1,35
$\text{IndPref}_{\text{Ref}}$	-0,04	0,47	-0,06	0,82	0,15	0,00	-0,50	1,55	0,74	0,26
Tsec_{min}	0,18	64,33	465,32	106,94	148,98	0,21	260,49	1361,91	173,46	460,24
CX_{Ref}	19,78	14,27	2,17	14,93	13,08	0,97	8,73	8,30	5,76	5,68
$\text{CMoving}_{\text{Ref}}$	3,05	0,12	1,02	0,67	1,47	0,15	0,89	4,03	-0,07	1,46
$\text{CPref}_{\text{Ref}}$	1,76	2,38	7,22	4,39	3,45	0,45	1,15	1,62	1,60	1,01
$\text{Tsec}_{\text{optimal}}$	0,27	75,30	479,96	113,85	156,83	0,28	282,10	1331,94	181,53	459,76

By comparing the results of the two methods GRASP and ILS with the CPLEX results, it is clear that ILS is better than GRASP where ILS show a gap of 4,74% on the lower bound in merely 106,94 seconds where the GRASP is only 5,53% in 173,46 seconds. Thus, ILS improve the upper bound of the CPLEX by 0,65%, meanwhile GRASP only obtain a gap of 0,59%.

When comparing the best solutions obtained by the hybrid method GRASP*ILS (considered as references), the results show that the GRASP metaheuristic is less efficient with an approximate average gap of 2,51%. On the other hand, ILS seems more efficient than GRASP with an approximate average gap of 1,44% from the method of reference. By seeing the ILS results, the difference is not that much with the hybrid method considering the overall execution (0,85% of gap considering all instances).

To summarize, the results obtained lead us to say that the hybrid method is more efficient and stable than the standard version of ILS, where ILS is better than GRASP for a given number of local search calls. Finally, the last conclusion for this paper is that GRASP*ILS is better than ILS which is better than GRASP.

VI. CONCLUSION AND PERSPECTIVES

In this paper, three metaheuristics for the VRPTW-SPLB are proposed. A GRASP method and an ILS method which use a constructive heuristic inspired by the nearest neighbor method and a local search which explore in an organized manner all the neighboring in order to find feasible solution optimally. Firstly, an investigation and overview of the problem of the vehicle routing problem with time windows, synchronization, precedence and lunch break constraints in the sector of home health care is presented. Then a mathematical formulation of the problem is proposed by the aid of the mixed integer programming language (MILP). After introducing the algorithms used to solve the problem, a numerical results section along with its analysis is foreshadowed to know which algorithm gives the better solution. By the outcome, it is clear that the GRASP and ILS method gives better results for larger instances than the CPLEX solver, but compared to the lower bounds given by CPLEX, the GRASP method seems less efficient than ILS, this can be explained by the iterations independency realized by GRASP, for this reason the execution times are longer than ILS, but still reasonable for such a problem which is known to be NP-Hard. The combination of these two methods which result in GRASP*ILS method demonstrated that it is better in term of computational time and the quality of the results obtained. In order to complete this study as future research is to conduct statistical tests in order to establish which method is more stable and gives better results. The limitation found in this study are frequent in this problem such as ensuring the satisfaction of the patient by assigning to him a specific caregiver can lead to an increase in the overall cost which is very degrading in term of optimizing either resources and time. In order to escape from this, a multi-objective approach can aid to solve this matter. As a continuity of this study, a multi-objective approach of the same problem which is VRPTW-SPLB in home health care with genetic algorithms in order to optimize the overall movement costs and patient's satisfaction simultaneously will be conducted in a future paper.

REFERENCES

[1] Dantzig, G.B, & Ramser, J.H. 1959a. The truck dispatching problem. Management Science, **6**(1), 80-91.

[2] Toth, P., & Vigo, D. 2014. Vehicle Routing : Problems, Methods, and Applications. Vol. 18. SIAM.

[3] Bräysi, O., Dullaert, W., & Nakari, P. 2009. The potential of optimization in communal routing problems : case studies from finland. Journal of Transport Geography, **17**(6), 484-490.

[4] Brasy, U., Gendreau, M.: vehicle Routing Problem with Time windows, part1: Route construction and local search Algorithm, Transport. Sci., 39(1)(2005), 104–110.

[5] G. A. P. Kindervater and M. W. P Savelsbergh. Vehicle routing : handling edge exchanges in E. H. L. Aarts et J. K. Lenstra (d.), Local search in combinatorial optimization. Wiley, Chichester, 311–336, 1997.

[6] Y. Nagata. Efficient evolutionary algorithm for the vehicle routing problem with time windows: edge assembly crossover for the VRPTW. IEEE Congress on Evolutionary Computation 2007: 1175-1182.

[7] Ait Haddadene, S.R, Labadie, N., & Prodhon, C. 2016. A GRASP × ILS for the vehicle routing problem with time windows, synchronization and precedence constraints. Expert Syst. Appl, Vol. 66, 274-294.

[8] Solomon, M.M. 1987. Agorithms for the vehicle routing and scheduling problems with time window constraints. Operations Research, 2(35), 254-265.

[9] Gehring, H., H¨omberger, J., “A Parallel Hybrid Evolutionary Metaheuristics for Vehicle Routing Problem with Time Windows”, In Proceedings of EUROGEN99. Jyvaskyla: University of Jyvaskyla, 57-64, 1999.

[10] L.D. Bodin, T. Sexton,. The multi-vehicle subscriber dial-a-ride problem. TIMS studies in M.Sc, 2, 73-86 (1986).

[11] J. Desrosiers, Y. Dumas,F. Soumis. A dynamic programming solution of the large-scale single-vehicle dial-a-ride with time windows. American Journal of Mathematical and Management Sciences, 6(3-4), 301-325 (1986).

[12] J. Desrosiers, Y. Dumas, M.M. Solomon, F. Soumis. Time constrained routing and scheduling. Handbooks in OR and M.Sc, 8, 35-139 (1995).

[13] L.M. Rousseau, M. Gendreau, G. Pesant. The synchronized vehicle dispatching problem. Citeseer (2003).

[14] L.M. Rousseau, M. Gendreau, G. Pesant. The Synchronized Dynamic Vehicle Dispatching Problem. INFOR : Information Systems and Operational Research, 51(2), 76-83 (2013).

[15] A. Coppi, P. Detti, J. Rafaelli. A planning and routing model for patient transportation in health care. Electronic Notes in Discrete Mathematics, 41, 125-132 (2013).

[16] E. Cheng, J.L. Rich. A home health care routing and scheduling problem. Technical report CAAM TR98-04, Rice University (1998).

[17] S. Bertels, T. Stefan. A hybrid setup for a hybrid scenario : combining heuristics for the home health care problem. Computers & Operations Research. 33(10), 2866-2890 (2006).

[18] P. Eveborn, P. Flisberg,M. Rönnqvist. Laps Care – an operational system for staff planning of home care. European Journal of Operational Research, 171(3), 962-976 (2006).

[19] A. Trautsamwieser, P.Hirsch. Optimization of daily scheduling for home health care services. Journal of Applied Operational Research, 3(3), 124-136 (2011).

[20] C. Akjiratikarl, P. Yenradee, P.R. Drake. PSO-based algorithm for home care worker scheduling in the UK. Computers & Industrial Engineering, 53(4), 559-583 (2007).

[21] K. Doerner, A. Focke, W.J. Gutjahr. Multicriteria tour planning for mobile healthcare facilities in a developing country. European Journal of Operational Research, 179(3), 1078-1096 (2007).

[22] D. Bredström, M. Rönnqvist. A branch and price algorithm for the combined vehicle routing and scheduling problem with synchronization constraints. NHH Dept. of Finance Management Science Discussion Paper (2007).

[23] D. Bredström, M. Rönnqvist. Combined vehicle routing and scheduling with temporal precedence and synchronization constraints. European Journal of Operational Research, 191(1), 19-31 (2008).

[24] R. Redjem, S. Kharraja, X. Xie, E. Marcon. Routing and scheduling of caregivers in home health care with synchronized visits. In : 9th International Conference on Modeling, Optimization & SIMulation (2012).

[25] S. Afifi, D.C. Dang, A. Moukrim. A simulated annealing algorithm for the vehicle routing problem with time windows and synchronization constraints. Pages 259-265: Learning and Intelligent Optimization. Springer (2013).

[26] R. Redjem, E. Marcon. Operations management in the home health care services : a heuristic for the caregivers’ routing problem. Flexible Services and Manufacturing Journal, 1-24 (2015).

[27] N. Labadie, C. Prins, Y. Yang. Iterated local search for a vehicle routing problem with synchronization constraints. Pages 257-263 of : ICORES 2014-Proceedings of the 3rd International Conference on Operations Research and Enterprise Systems, Angers, Loire Valley, France (2014).

[28] B. Issaoui, I. Zidi, E. Marcon, K. Ghedira. New Multi-Objective Approach for the Home Care Service Problem Based on Scheduling Algorithms and Variable Neighborhood Descent (2014).

[29] M.S Rasmussen, T. Justesen, A. Dohn and J. Larsen. 2012. The Home Care Crew Scheduling Problem: Preference-Based Visit Clustering and Temporal Dependencies. European Journal of Operational Research (3), p. 598-610.

[30] En-nahli, Laila & Afifi, Sohaib & Allaoui, Hamid & Nouaouri, Issam. (2016). Local Search Analysis for a Vehicle Routing Problem with Synchronization and Time Windows Constraints in Home Health Care Services.-IFAC-Papers-OnLine.49.1210-1215. 10.1016/j.ifacol.2016.07.674.

[31] Euchi, J. 2020. Do drones have a realistic place in a pandemic fight for delivering medical supplies in healthcare systems problems? Chin.J.Aeronaut. (https://doi.org/10.1016/j.cja.2020.06.006).

[32] Liu, R., Xie, X., Augusto, V., & Rodriguez, C. 2013. Heuristic algorithms for a vehicle routing problem with simultaneous delivery and pickup and time windows in home health care. European Journal of Operational Research, 230(3), 475-486.

[33] Ceselli, A., Righini, G., & Tresoldi, E. 2014. Combined location and routing problems for the drug distribution. Discrete Applied Mathematics, 165, 130-145.

[34] Leandro C. Coelho, Jacques Renaud & Gilbert Laporte (2016) Road-based goods transportation: a survey of real-world logistics applications from 2000 to 2015, INFOR: Information Systems and Operational Research, 54:2, 79-96, DOI: 10.1080/03155986.2016.1167357.

[35] Y.Kergosien, Ch.Lente, J.C.Billaut and S.Perrin, 2013. Metaheuristic algorithms for solving two interconnected vehicle routing problems in a hospital complex, Computers & Operations Research 40 (10), p. 2508–2518.

[36] Drexl, M. 2012. Synchronization in vehicle routing- A survey of VRPs with multiple synchronization constraints. *Transportation Science*, **46**(3), 297-316.

[37] Ioachim, I., Desrosiers, J., Soumis, F., & Bélanger, N. 1999. Fleet assignment and routing with schedule synchronization constraints. *European Journal of Operational Research*, **119**(1), 75-90.

[38] Dep Pia, A., & Flippi, C. 2006. A variable neighborhood descent algorithm from a real waste collection problem with mobile depots. International Transactions in Operational Research. 13. 125-141.

[39] Feo, T.A., & Resende, M.G.C. 1995. Greedy randomized adaptive search procedures. *Journal of global optimization*, **6**(2), 109-133.

[40] Kontoravdis, George & Bard, Jonathan. (1995). A GRASP for the Vehicle Routing Problem with Time Windows. INFORMS Journal on Computing. 7. 10-23. 10.1287/ijoc.7.1.10.

[41] Glover, F., & Laguna, M. 2013. *Tabu Search*. Springer New York.

[42] Nguyen, V. Prins, C, & Prodhon, C. (2012). A multi-start iterated local search with tabu list and path relinking for the two-echelon location-routing problem. *Engineering Applications of Artificial Intelligence, 219*, 598-610.

[43] J. Michallet, C. Prins, L. Amodeo, F. Yalaoui, G. Vitry Multi-start iterated local search for the periodic vehicle routing problem with time windows and time spread constraints on services Computers & Operations Research, 41 (2014), pp. 196-207.

Development and Usability Testing of a Consultation System for Diabetic Retinopathy Screening

Nurul Najihah A'bas[1], Sarni Suhaila Rahim[2]
Mohamad Lutfi Dolhalit[3], Wan Sazli Nasarudin
Saifudin[4], Nazreen Abdullasim[5], Shahril Parumo[6]
Fakulti Teknologi Maklumat Dan Komunikasi, Universiti
Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia

Raja Norliza Raja Omar[7], Siti Zakiah Md Khair[8]
Khavigpriyaa Kalaichelvam[9]
Syazwan Izzat Noor Izhar[10]
Department of Ophthalmology, Hospital Melaka
Jalan Mufti Haji Khalil, 75400 Melaka, Malaysia

***Abstract*—This study aims to develop a novel web-based decision support system for diabetic retinopathy screening and classification of eye fundus images for medical officers. The research delivers diabetic retinopathy information with a web-based environment according to the needs of the users. The proposed research also intends to evaluate the developed system usability to the target users. The complex characteristics of diabetic retinopathy signs contribute to the difficulty in detecting diabetic retinopathy. Therefore, professional and skilled retinal screeners are required to produce accurate diabetic retinopathy detection and diagnosis. The proposed system assists the communication and consultation among the medical experts in the hospital and the primary health cares located at the health clinics. The agile software development model is the methodology used for the development of this research project. The project collaborates with the Department of Ophthalmology, Hospital Melaka, Malaysia for the medical content expertise and testing. Representative medical officers from Hospital Melaka and all the public health clinics in Melaka were involved in the preliminary study and system testing. This research study consists of a web development producing an interactive web-based application of diabetic retinopathy consultation which comprises image processing and editing features as a core of the system. It is envisaged that this research project will contribute to the management of diabetic retinopathy screening among medical officers.**

***Keywords*—Consultation; diabetic retinopathy; eye screening; image editing; image processing; web development; testing**

I. Introduction

Diabetic retinopathy (DR) is a type of eye disease due to a diabetic condition that damages the retina, leading to blindness or vision loss. In Malaysia, diabetes in the eye is the most common cause of vision loss in working age adults. Therefore, screening of DR is important for early detection and early treatment. A precise retinal screening is required to help retinal screeners to distinguish the retinal images efficiently. This study aims to develop a web-based decision support system to screen and diagnose DR in eye fundus images. Furthermore, the study would examine the information on DR, including screening and diagnosis. This research also aims to determine the usability and needs demand by the target users of the developed system. The detection of DR is challenging due to the complex characteristics of the DR features as illustrated in Fig. 1 which emphasizes the need for highly qualified and experienced retinal screeners to ensure accurate detection and diagnosis of DR. Moreover, this system involves image processing and image editing features to enhance the visibility of fundus images captured by primary health cares. The built web-based system aims to create a new web application to provide online users with innovative services or solutions. Therefore, an enhancement web-based system based on online consultation to overcome the weaknesses of the current method such as many-to-many communication, image upload, image editing and image processing.

In this paper, basic information of DR and related works will be stated in the next section, a brief explanation on research materials and method used will be shown in the third section that includes the image processing and image editing features, while the fourth section will discuss the results of this research.

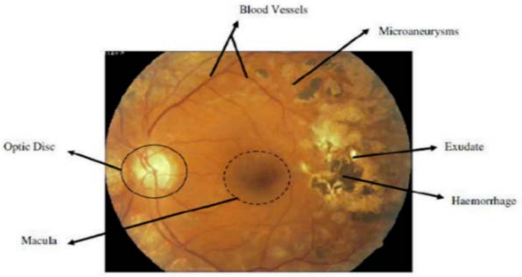


Fig. 1. Fundus Retinal Image with Lesions [1].

II. Literature Review

DR is a diabetic complication caused by elevated blood glucose levels. In order to diagnose DR symptoms, an eye screening is required. Microaneurysms, retinal haemorrhages, hard exudates, abnormal new vessels, cotton wool spots, and venous beadings are some of the signs that must be discovered throughout the DR screening process. Thus, it's essential to have a thorough consultation platform that allows primary health care to identify and classify the DR risk factors based on the patient's fundus image, since the vision impairment due to DR is increasing. Early detection and diagnosis of DR are vital for saving the vision of diabetic persons.

Globally, at least 2.2 billion people experience vision impairment, and of these, at least 1 billion people have vision impairment that could have been prevented or is yet to be addressed [2]. It is revealed that globally, 146 million (34.6%) were diagnosed as DR in 2014 for adults aged over 18 years with diabetes [2]. The largest number of blind and visually impaired people reside in the Asian region, namely, South Asia, followed by East Asia and Southeast Asia [3]. DR caused 1.1% of all cases of blindness and 1.3% of all visual impairment in 2015. The percentage of blindness caused by DR varied in the Asia-Pacific, ranging from less than 1% in South Asia (0.16%), Oceania (0.32%), East Asia (0.51%), and Southeast Asia (0.59%) to more than 3% in Central Asia (3.60%), high-income Asia-Pacific countries (3.87%), and Australasia (4.48%). Prevalence of blindness and visual impairment due to DR decreased between 1990 and 2015 in Oceania and Central and Eastern Europe, yet increased in high-income regions of Asia-Pacific, North America, Australasia, and Asian regions [3].

Malaysia also has the largest number of people with visual impairments, namely, from Kedah, Perlis, Johor, and Perak as studied by the Malaysian [4] as shown in Fig. 2 [5]. The result shows the prevalence of all states in Malaysia. Kedah demonstrates the highest percentage of people with diabetes with a difference of 4.8% in Perlis while a percentage difference of 0.8% in Johor. It is also shown that Johor has a difference of 0.4% of diabetic prevalence with Perak.

Fig. 3 shows the results by the [6] regarding the prevalence of diabetes mellitus in Malaysia with an age group above 30 years old. The percentage of diabetes mellitus in Malaysia is increasing from 1986 to 2011. It increased by about 2% from 1986 to 1996 within the range of 10 years. In 2006, it increased by 6.6% from 1996 which is also within 10 years. It shows that the prevalence between 1996 and 2006 increased by 4.3% higher than from 1986 to 1996. From 2006 to 2011, the percentage of the prevalence is increasing by 5.9% which is lower by 0.7% than between 1996 and 2006. Overall, it shows that the prevalence of this disease is gradually increasing.

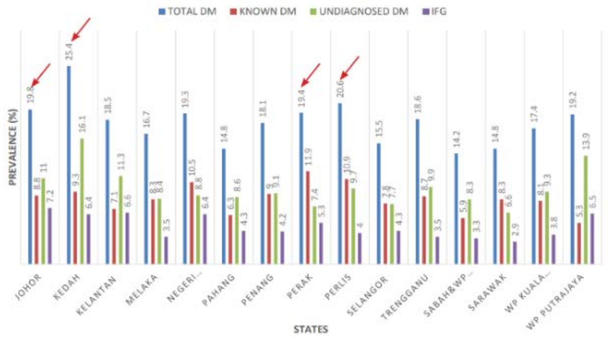


Fig. 2. Prevalence of Diabetes by States [5].

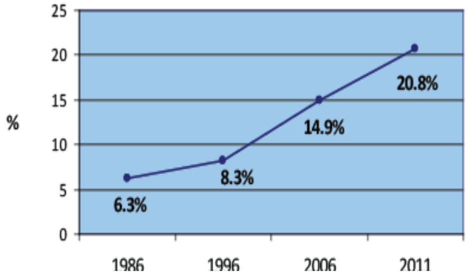


Fig. 3. Prevalence of Diabetes Mellitus in Malaysia (>30 Years Age Group) [6].

The proposed research project is a guideline in the development of a system. Several systems for detection and diagnosis of DR are reported in the literature which are development focused and proposed techniques for detecting certain features of DR. For instance, the existing system of automated DR screening detection as proposed by [7] and [8] that assist in diabetic retinopathy detection. Another existing related system is the teleophthalmology screening system developed by [9] which integrates multiple tools into ForusCare such as software platform where the system secures the image analysis module and stored and forwarded into the Amazon cloud. In addition, there are additional system features such as system installation and workflow that enhance communication and assures quality photos. [10] have developed a mobile system called mTEH that enables employees operating from different remote locations to be connected, securely stores all eye-screening participant data, and encourages improved decision-making of participants on eye health. An online teleophthalmology screening conferencing system was developed by [11] known as TeleOph as the virtual channel for DR screening. Telemedicine is a correlation between medicine and the technology available [12]. Recently, [13] proposed a telemedicine system that offers communication using 5G uRLLC and mMTC. Furthermore, another similar telemedicine that has made tremendous strides and successfully interacts with patients and doctors is HEMAN which is an IoT-based e-health care system for remote telemedicine [14].

III. Materials and Methods

The developed web-system is built using these programming languages and scripting; HTML5, PHP7, JavaScript, jQuery, CSS3, and SQL, with the CMS WordPress platform. These programming languages are chosen since these languages are able to perform responsively in any types of gadgets such as laptop, tab, and mobile [15]. This system also uses cPanel as the web hosting platform to publish the system online. Adobe Dreamweaver CC 2020 is another tool used in writing the codes and designing the interface and theme of the system. PHP7 performs many functions that need to connect to the database such as fundus image upload and retrieve. Other than that, jQuery language plays a big role in running the image editing and image processing plugin.

This section outlines the research process to accomplish the objectives of this research. Overall, the five phases cover in this research are analysis, design, development, testing, and evaluation; the methodology of this research to achieve the objectives of the research study. Fig. 4 presents the graphical illustration of the Agile model.

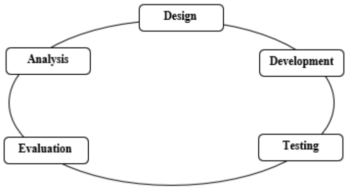


Fig. 4. Web Application Development Process.

A. Analysis

An online survey was created to collect data as the input for developing the system. The survey has been distributed to 202 primary health cares from Hospital Melaka and 32 health clinics around Melaka. This survey covered four main parts which are demographic, computer literacy/skills, current DR screening practice, and expected outcome. The details of the online survey are summarized in Table I. The online survey result clearly indicates that most of the respondents agreed to the development of the system. Moreover, this online survey helps analyze the problems, limitations, suggestions, and expected outcome of the system. Thus, it can be concluded that the proposed web-based consultation system is highly needed by primary health cares to conduct the DR screening process.

TABLE I. SUMMARY OF THE PRELIMINARY ONLINE STUDY

Date/Duration	23/6/2020 – 2/9/2020 (10 weeks)
Type of survey	Online questionnaire via Google Form
No. of respondents	202 respondents
No. of workplace	1 hospital, 32 health clinics
Type of respondents	Consultant, Specialist, Medical Officer and House Officer
No. of survey parts	4 sections - Demographic - Computer Literacy/Skills - Current DR Screening Practice - Expected Outcome

B. Design

In the design phase, considerations such as software needs, how the output of the system will look, what database will be used, and the timeline of development were identified. Fig. 5 shows the flow of the system process. The design phase for this system began with designing the web environment since the primary health cares required an online platform for the system. Therefore, this system needs to have a web domain and hosting for the system to be published and viewed on the web browser.

The next task was to design a database for the system. Designing a database requires the list of functions of the system and information needed to derive from the system to determine the possible tables and fields need to create in the database. Microsoft Visio software was used to design the Entity Relational Diagram and phpMyAdmin to create the database of the system. After the database was designed, creating web interactivity starts with the connection of the system to the database and triggering the buttons in the system. Ensuring that the connection of the database and trigger buttons are working fine will ease the next step which was designing the web-based interface with the bootstrap theme plugin. Once the system was completed with the design of web environment, database, interactivity and interface, the system proceeded to the development process.

C. Development

Once all the requirements and designs were documented, the development phase took place. There are four features that are the core function of the system. This system is able to be accessed through this link; http://drcs.com.my/ which has four types of users which are: Admin, Specialist, Primary Health Care, and Public. Admin is allowed to handle the user and hospital and health clinic data. The specialists can view the list of DR cases uploaded by the primary health cares and specialists can give feedback on the DR cases while using image processing and image editing features in the system. Primary health cares are allowed to fill up DR case form including uploading fundus images from both right and left eyes, view the list of DR cases uploaded by other primary health cares, and view the feedback from the specialists and reply to the feedback given. Besides that, the public type of user can gain information on the DR and DR screening, view an example of DR fundus images, and gain knowledge on the signs of DR from the fundus images. and gain knowledge on the signs of DR from the fundus images. Table II shows the list of the system users and their accessibility.

Fig. 6 depicts the flow of the development of the DRCS. The system development began with the basic functions which were Create, Read, Update, and Delete (CRUD) function for the user, hospital and health clinic, and diabetic retinopathy’s case details using PHP and MySQLi.

The image processing plugin which includes image color conversion and image filtering was developed. Subsequently, the development of image editing plugin that allows to crop, scale, and annotate the fundus image takes place. A forum area for specialists to review and comment on the DR cases with the primary health cares was provided in the system.

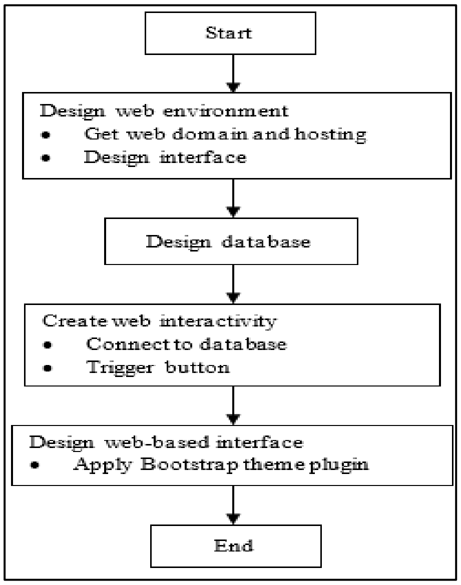


Fig. 5. Flow of the System Design Phase.

TABLE II. SUMMARY OF USER ACCESSIBILITY

Type of User	Accessibility
Admin	User and hospital/health clinic data
Specialist	DR case, discussion area
Primary Health Care	DR case, discussion area
Public	Information on DR and DR screening

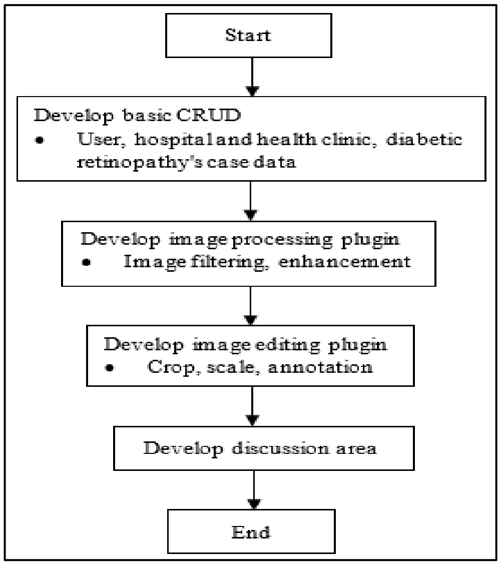
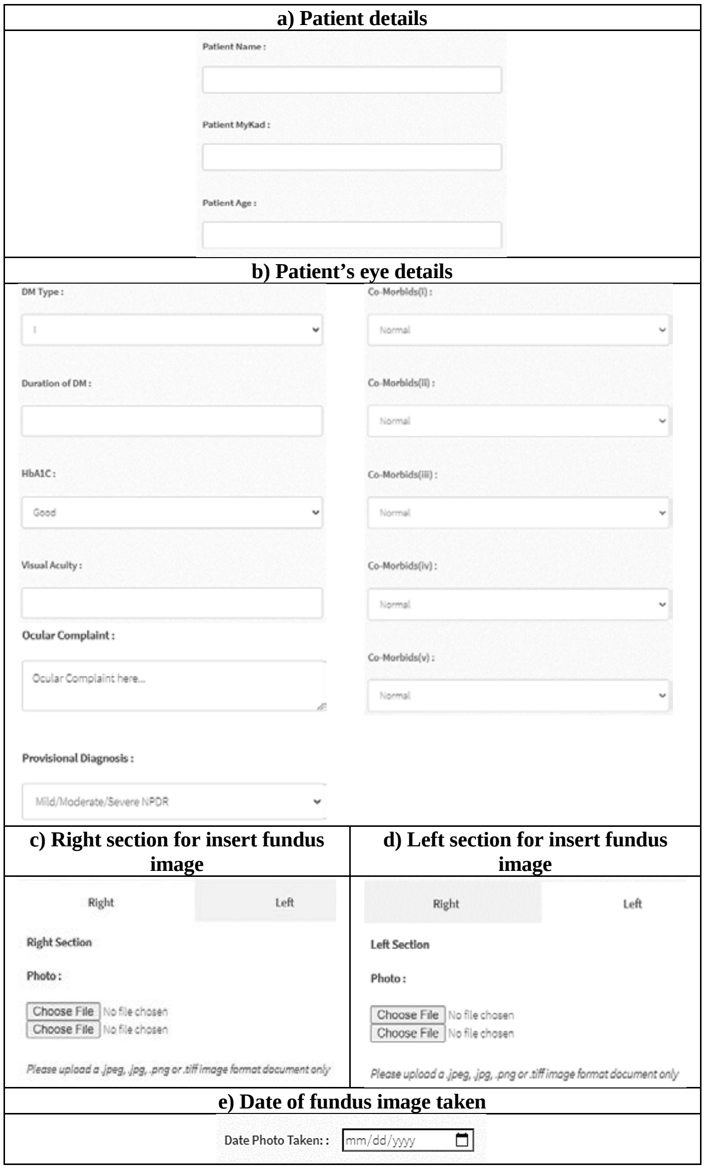

Start
Develop basic CRUD
• User, hospital and health clinic, diabetic retinopathy's case data
Develop image processing plugin
• Image filtering, enhancement
Develop image editing plugin
• Crop, scale, annotation
Develop discussion area
End

Fig. 6. Flow of the System Development Phase.

1) Create, Read, Update, Delete: CRUD is the acronym for create, read, update, and delete which are the four basic types of functionality for constructing a web application. This system development includes inserting user, hospital or health clinic, and DR's case data. Create function will be called when new data need to be added into the database. The new entry is assigned a unique ID (identity document), which can be used to access the data later. Read or view function will view all the data currently exist in the database when it is called. The system will show the list of all the DR's data in a form as shown in Fig. 7. The update function will be called when no data need to be altered. Furthermore, the data will display all information from the selected row and users are able to edit data in each column at once. After the function is called, the corresponding data will be updated into new data. The delete function is also needed in this system to remove data that are no longer needed in the database.

The admin user is only allowed to handle user and hospital and health clinic details. In this system, only the admin can handle the registration of the new system user since the access and data are strictly confidential. Besides that, the admin is able to view the details of the users, such as username, password, email, full name, doctors' level, and workplace. As for hospital and health clinic details, the admin can add, view, edit and delete the details including the type of organization either hospitals or health clinics, name, address, region postcode, phone number, fax number, email, and status of the organization. Any changes in the details can be made by the admin only for security purposes. For instance, changes in phone and fax numbers can only be done by the admin of the system.

As for DR case form, primary health cares need to insert two fundus images from the right side and another two fundus images from the left side. Two views need to be captured for each side of the eye, which are macula center and optic disc center. These requirements for inserting fundus images area are based on the DR practice mentioned by the primary health cares during the preliminary study. Each DR case form includes the patient's detail consists of name, identification number, and age as shown in Fig. 7(a). It also has patient's eye details that include the Diabetes Mellitus (DM) Type, duration of DM, HbA1C, visual acuity, ocular complaint, provisional diagnosis, and types of co-morbids that are visualized in Fig. 7(b). Fig. 7(c) to Fig. 7(e) show the fundus image section required in the DR case form which needs to be uploaded by primary health cares containing both sides of fundus image uploading form and the date of the photo taken. In order to ensure the primary health cares insert the correct image format into the file form, the error function is created if the primary health cares inserted files other than the image file format listed such as .jpeg (or .jpg), .png or .tiff file format as shown in Fig. 7(c) and Fig. 7(d). This is to prevent primary health cares from inserting unnecessary file format such as .pdf, .gif or any other file formats.


a) Patient details

Patient Name :

Patient MyKad :

Patient Age :

b) Patient's eye details

DM Type : I

Duration of DM :

HbA1C : Good

Visual Acuity :

Ocular Complaint : Ocular Complaint here...

Provisional Diagnosis : Mild/Moderate/Severe NPDR

Co-Morbids(i) : Normal

Co-Morbids(ii) : Normal

Co-Morbids(iii) : Normal

Co-Morbids(iv) : Normal

Co-Morbids(v) : Normal

c) Right section for insert fundus image

Right Section

Photo :

Choose File No file chosen
Choose File No file chosen

Please upload a .jpeg, .jpg, .png or .tiff image format document only

d) Left section for insert fundus image

Left Section

Photo :

Choose File No file chosen
Choose File No file chosen

Please upload a .jpeg, .jpg, .png or .tiff image format document only

e) Date of fundus image taken

Date Photo Taken: : mm/dd/yyyy

Fig. 7. DR Case Form.

2) Image processing: Image processing performs some operations on an image to produce an enhanced image and is useful in extracting useful information from it. Image preprocessing techniques involved in the proposed work include color image conversion, contrast enhancement, filtering and segmentation among others. Current fundus images might not be so clear due to the noise during the image capture process. Therefore, this system includes image processing techniques which is important to help in processing the fundus image into a clearer version and able to help the primary health cares to detect the signs precisely. This plugin includes color image conversion such as grayscale, black and white as shown in Fig. 8(a) and Fig. 8(b), respectively. This plugin includes image filtering such as thresholding and edge detection as presented in Fig. 8(c) and Fig. 8(d). Thus, these image processing features help specialists view and locate the complex signs of DR such as microaneurysms and neovascularization.

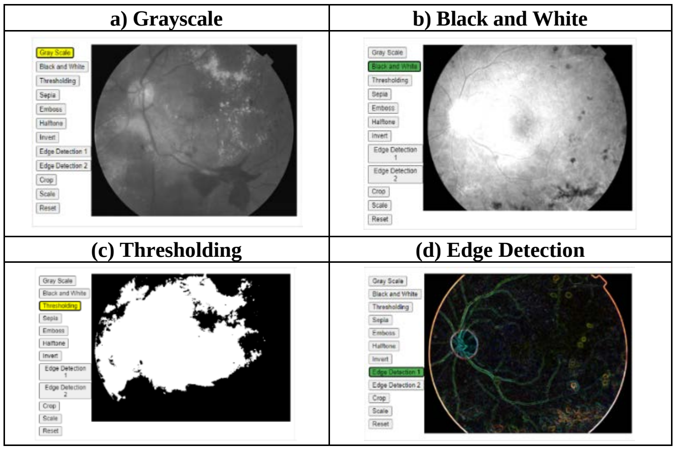

a) Grayscale **b) Black and White** **(c) Thresholding** **(d) Edge Detection**

Fig. 8. Example of Image Processing Plugin.

3) Image editing: Image or photo editor plugin was implemented to support basic editing on the image uploaded such as adding text, crop, and scale. Fig. 9 shows an example of an image editing feature which is annotation of the eye image which could help the specialist to review or respond to the DR case submitted by the primary health cares effectively. This annotation feature is important to facilitate the specialist to label the signs of DR on the fundus image since the specialists are expert in detecting the critical signs of DR. Therefore, specialists are able to use any suitable tools of annotation listed in Table III by labelling or drawings on the fundus images that later can be viewed by the primary health cares. Several annotation styles were provided to the specialists to annotate the fundus image such as adding box, circle, adding text for notes, arrow, and draw. Each annotation tool is effective in labelling the signs of DR which ease the specialist to point out the critical signs for the primary health cares' view. At the same time, primary health cares are also able to see the label created by the specialist clearly. This annotation tool also provides the undo and redo button to erase the last change done to the fundus image.

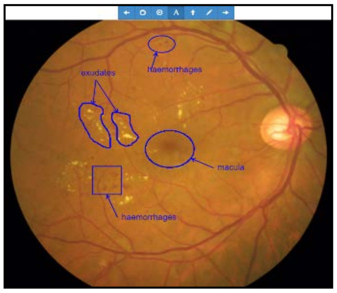
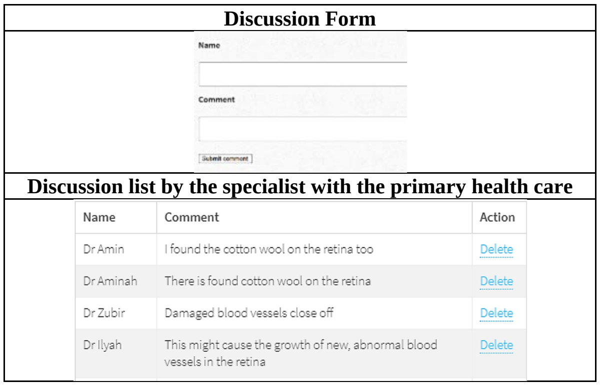


Fig. 9. Example of Applying Annotation Feature of Image Editing on the Fundus Images (Fundus Image Taken from the [4]).

TABLE III. LIST OF ANNOTATION TOOLS

Annotation tools	←	▢	⊙	A	↑	✎	→
Function	Undo	Box	Circle	Text	Arrow	Draw	Redo

4) Forum: Based on the previous preliminary survey conducted, primary health cares mentioned that one-to-one conversation is a limitation in the current communication method. Therefore, developing a platform for discussion would be beneficial for primary health cares to share and learn by discussing with the specialist. In this system, the discussion area is mainly for the consultants and specialists to give their comments and verification to the data provided by the primary health cares in the DR case form. Fig. 10 shows an example of the discussion area of the case details where Fig. 10(a) is a form for the specialist or the primary health care to write the review and Fig. 10(b) is the listed review and comments done by the specialist.


Discussion Form

Name

Comment

Submit comment

Discussion list by the specialist with the primary health care

Name	Comment	Action
Dr Amin	I found the cotton wool on the retina too	Delete
Dr Aminah	There is found cotton wool on the retina	Delete
Dr Zubir	Damaged blood vessels close off	Delete
Dr Ilyah	This might cause the growth of new, abnormal blood vessels in the retina	Delete

Fig. 10. Discussion Area for the Primary Health Cares.

IV. RESULTS AND DISCUSSION

Alpha testing has been conducted to evaluate the usability of the system prototype. This testing was done by the consultants, specialists, and primary health cares from the Department of Ophthalmology, Hospital Melaka. The testing covers the functionality of the features available in the system. Table IV summarizes the alpha testing conducted. There are three different users in this system which are admin, specialist, and primary health care which give different perspectives for each user. Therefore, three different results are reported based on each type of user. Moreover, these results are important for this study to identify the effectiveness of the features for DR screening consultation. This section is divided into three types of user which are admin, specialist, and primary health care.

The first part of the questionnaire; demographic information of the respondents, aims to provide an overview of the general details of the respondents such as gender, age, level of work, current workplace, and years of work experience. These data are useful for strategic planning for consultation web-system. Based on the result summarized in Table V, the respondents are mostly female between 40 and 50 years old. The result also shows that the respondents are currently working in Hospital Melaka with more than 10 years of working experience in the medical sector.

A. Admin

1) Functionality of DRCS: The functionality of the system in the admin site covers the functions of CRUD in user, hospital, and health clinic details. The results of this functionality testing are detailed in Table XI. Based on the overall result for admin site, most of the respondents successfully used the functions in the system which are the basic CRUD functions for user details, hospital, and health clinic details.

TABLE IV. SUMMARY OF THE ALPHA TESTING CONDUCTED

Date/Duration	11/12/2020 (3 hours)
Type of survey	Online questionnaire via Google Forms
No. of respondents	10 respondents
No. of workplace	Hospital Melaka
Type of respondents	Consultant, Specialist, Primary Health Care
No. of survey parts	4 sections - Demographic - System Functionality - System Usability - Suggestion for Improvement

TABLE V. DEMOGRAPHIC

	Doctor's Level			***Total***
	Consultant	***Specialist***	***Primary Health Care***	
Gender • Male • Female	- 1 1	2 6 8	- 1 1	**2** **8** **10**
Age • Less than 20 years old • 20-30 years old • 30-40 years old • 40-50 years old • More than 50 years old	- - - - 1 1	- - 6 2 - 8	- - 1 - - 1	**0** **0** **7** **2** **1** **10**
Current Workplace • Hospital • Health Clinic	1 - 1	8 - 8	1 - 1	**10** **0** **10**
Working Years • Less than 3 year • 3-5 years • 6-8 years • 8-10 years • More than 10 years	- - - - 1 1	- - - 1 7 8	- - 1 - - 1	**0** **0** **1** **1** **8** **10**

2) Usability of DRCS: The usability of the system in admin site was evaluated based on how easy the system is used. From this section, respondents were able to find the design flaws in terms of error messages during the input of data, confirmation message before the data were deleted, and whether the functions applied in fundus images work as it supposed. Results of the system usability for admin are shown in Table VI. Overall, the results from this testing show that the respondents are satisfied and at ease with the design and performance of the system.

3) Suggestion for improvement: Based on the alpha testing with the doctors, the respondents suggested sending an auto-mail to the registered users including the username and password. They also suggested having a mobile-friendly view of the system for a compatible platform of the system.

B. Specialist

1) Functionality of DRCS: The functionality testing for the specialist site covers the functionality of the DR cases and image editing. Table VII shows the results of the testing. Based on the overall result for the specialist site, most of the respondents successfully use the functions in the system which are the DR cases view, functionality of the annotation tools, and image editing tools.

2) Usability of DRCS: In the specialist site, usability is evaluated based on the friendliness of the system to the users. The result of this usability testing is shown in Table VIII. Overall, the results from this testing show that most of the respondents agree with the friendliness of the system design except for information finding, where the result mostly shows disagreement due to the unavailability of the searching function.

3) Suggestion for improvement: Several suggestions were given by the respondents regarding the specialist site. For instance, it is suggested to provide a full scale for viewing the fundus image such as the zooming tool. Besides that, it is better to include the security function for patients and doctors' privacy. Regarding the case form of diabetic retinopathy, the respondents suggested a better arrangement of the data in the form containing the demographic information of the patients. For the discussion area, the respondents suggested that the specialist or primary health cares who are assigned to handle the case can only edit and delete the discussion. It is also suggested to disable the discussion area once the diabetic retinopathy case is completed to avoid any changes in the data.

C. Primary Health Care

1) Functionality of DRCS: In the Primary Health Care site, various parts are available in the system such as DR case details, fundus images processing, and discussion area for the primary health care, and specialist. This section provides the evaluation results on the functionality of the DR case details, fundus images, image processing, and image editing. Based on the overall result, most of the respondents have successfully used the functions, as presented in Table IX.

2) Usability of DRCS: The usability of the system for the primary health care's site was evaluated based on the easiness and efficiency of the system. Table X shows the result of the usability testing of the primary health care's site. Overall, the results from this testing show that the respondents mostly agree with the design and performance of the system, except for the information finding where respondents found it is difficult to search for the information of DR case or patients details when there are many DR cases listed in the system. This is due to the lack of searching feature in the system.

3) Suggestion for improvement: The alpha testing for primary health cares revealed that the respondents suggested more security for patients' details since the data are confidential. In addition, respondents suggested for the primary health cares' discussion area to be anonymous or semi-private.

TABLE VI. ADMIN TESTING ON USABILITY

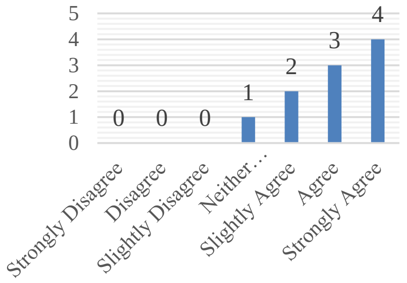
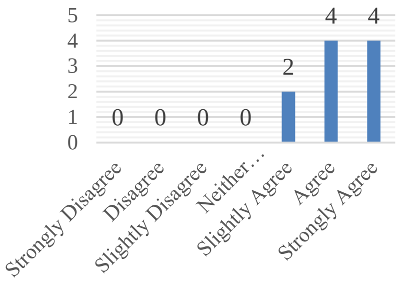
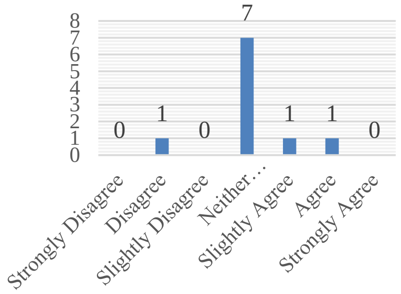
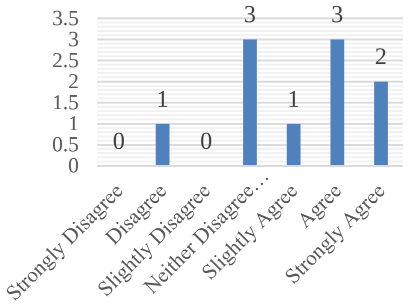
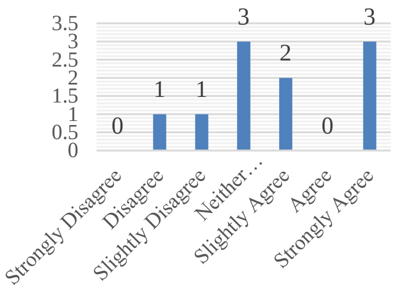
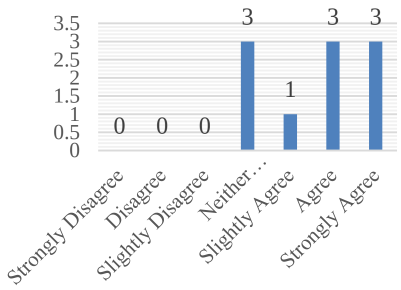
Ease of use	**Ease to learn using the system**	**System gives error messages**
		
Ease in information finding	**Expectation functions and capabilities included in the system**	**Satisfied with overall system**
		

TABLE VII. FUNCTIONALITY TESTING ON SPECIALIST SITE

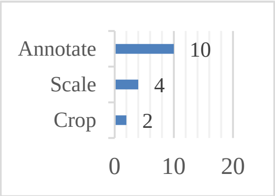
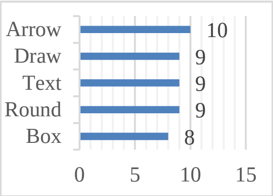
List of DR cases submitted by other Primary Health Cares able to view		**Successfully view DR case details**		**Successfully comment on DR case details**		**Annotation tool working fine**		**Requiring the annotation tool**	
Yes **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	9 1	**Yes** **No**	10 0	**Yes** **No**	10 0
Image editing style required		**Annotation style required**		**Image editing buttons are working fine**		**Image editing features offered are suitable**		**Image editing features offered sufficient**	
				**Yes** **No**	9 1	**Yes** **No**	8 2	**Yes** **No, need magnifiers**	8 2

TABLE VIII. SPECIALIST TESTING ON USABILITY

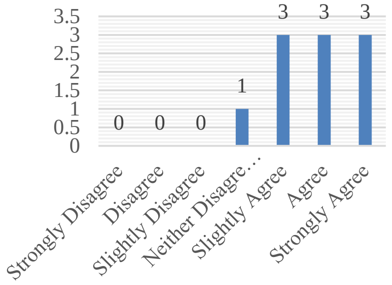
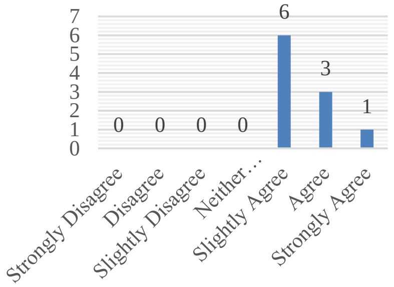
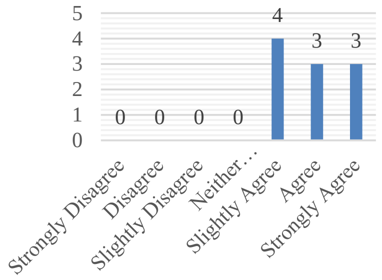
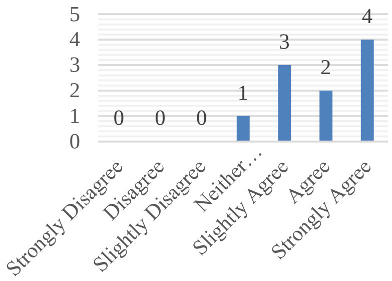
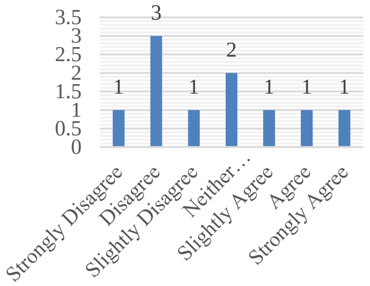
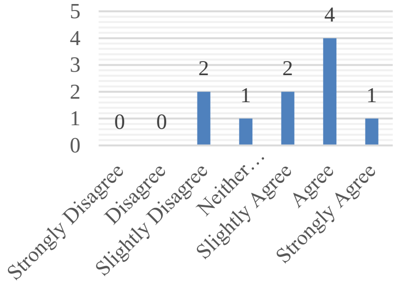
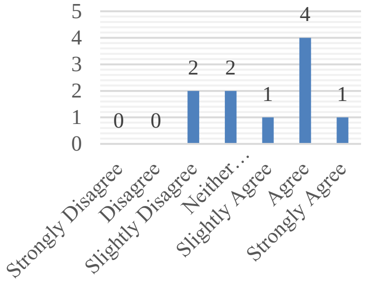
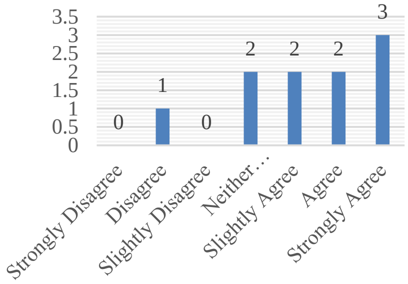
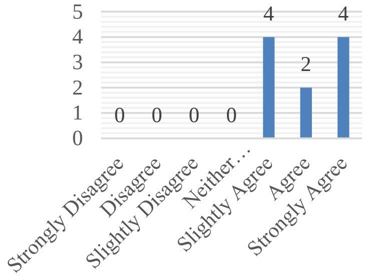
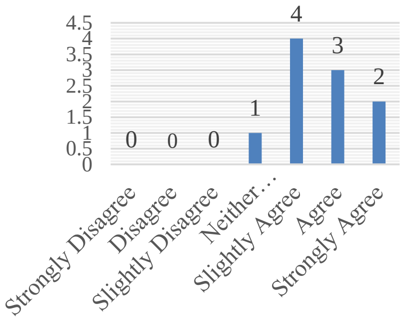
<table>
<tr><td><b>Simple usage of the system</b></td><td><b>System effectively diagnose the signs of DR</b></td><td><b>Annotation features able to use on the fundus efficiently</b></td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td><b>Ease learning to use the system</b></td><td>System gives error messages</td><td>Ease in information finding</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td><b>Expectation functions and capabilities included in the system</b></td><td><b>Ability to consult primary health care on DR through the system</b></td><td><b>Communication medium among specialist and primary health cares can be done in the system</b></td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td colspan="3"><b>Satisfied with overall system</b></td></tr>
<tr><td colspan="3">
</td></tr>
</table>

TABLE IX. FUNCTIONALITY TESTING ON PRIMARY HEALTH CARE SITE

Ability to view DR case details list submitted by Primary health cares		**Successfully add new DR case details**		**Successfully view DR case details**		**Successfully edit DR case details**		**Successfully delete DR details**	
Yes **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	8 2	**Yes** **No**	7 3
Successfully comment on DR case details		**Successfully upload fundus images**		**Able to view a fundus image**		**Successfully change/reupload another fundus image**		**Successfully delete fundus image**	
Yes **No**	10 0	**Yes** **No**	9 1	**Yes** **No**	10 0	**Yes** **No**	9 1	**Yes** **No**	10 0

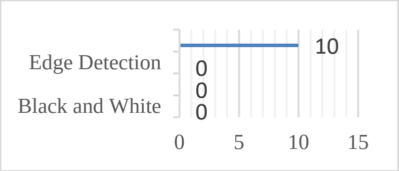
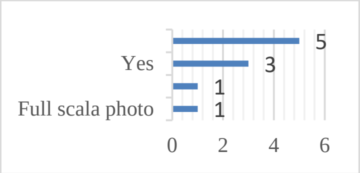
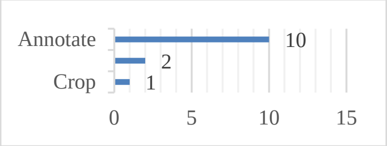
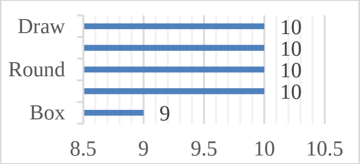
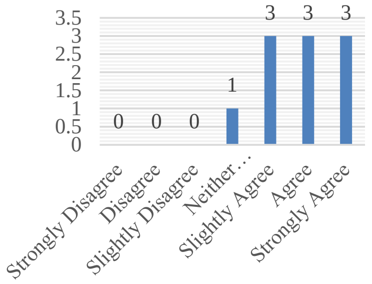
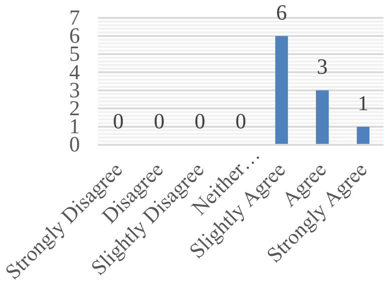
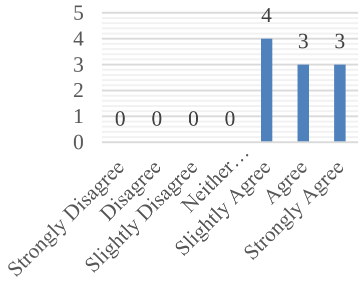
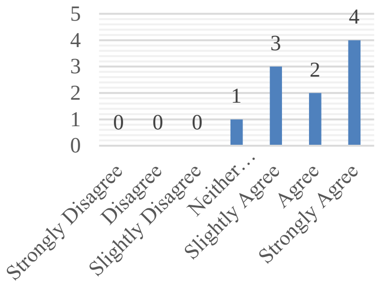
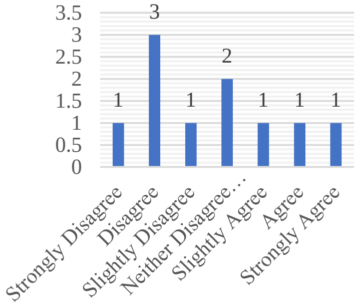
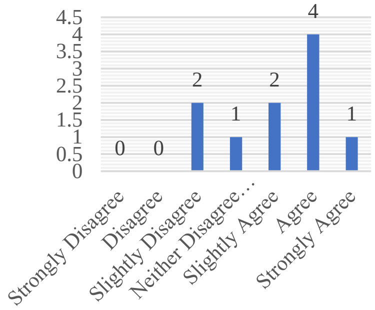
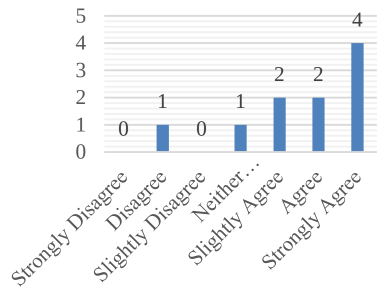
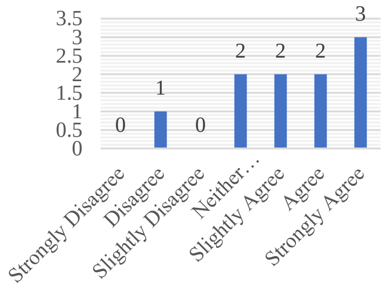
Image processing buttons work fine		**Offered suitable image processing features**		**Required image processing features**	**Offering sufficient image processing features**
Yes **No**	9 1	**Yes** **No**	9 1		
Annotation tool works fine		Requiring annotation tool		Image editing style required	Annotation style required
Yes **No**	10 0	**Yes** **No**	10 0		

Image editing buttons are working fine		**Image editing features offered are suitable**		**Image editing features offered sufficient**	
Yes **No**	10 0	**Yes** **No**	10 0	**Yes** **No, need magnifiers**	9 1

TABLE X. PRIMARY HEALTH CARE TESTING ON USABILITY

Simple usage of the system	**System effectively diagnose the signs of DR**	**Annotation features able to use on the fundus efficiently**
		
Ease learning to use the system	**System gives error messages**	**Ease in information finding**
		

<table>
<tr><th>Expectation functions and capabilities included in the system</th><th>Ability to consult primary health care on DR through the system</th></tr>
<tr><td>


</td><td>


</td></tr>
</table>

TABLE XI. FUNCTIONALITY TESTING ON ADMIN SITE

Ability to view user details list		**Successfully add new user details**		**Successfully edit user details**		**Successfully delete user details**	
Yes **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	8 2
Ability to view hospital / health clinic details list		**Successfully add new hospital / health clinic details**		**Successfully edit hospital / health clinic details**		**Successfully delete hospital / health clinic details**	
Yes **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	10 0	**Yes** **No**	8 2

V. DISCUSSION

In summary, a total of 10 respondents from Hospital Melaka have participated in the alpha testing of the system. The results of this testing enable a further review and evaluation of the functionality and usability of the developed system. Throughout the development of the system, it can be concluded that the system is able to overcome the problems and challenges of DR screening with the existence of the image processing and image editing plugins in the web-based environment. The image processing approach provides image filtering features and other image processing techniques which produce the image improvement and enhancement of several image characteristics or features for the next stage of processing. As for image editing, the annotation feature allows the primary health cares to draw or write on the fundus image, where the specialist is required to scribble some comments on the fundus image. Moreover, with the discussion area provided in the system, it allows the specialist to provide opinions and verify the content of the case details. This function also allows discussion and knowledge sharing of DR signs with other primary health cares through an online platform.

VI. CONCLUSION

The newly proposed system integrates image processing and image editing in a web-based method for DR screening. It improves the communication channel between the specialist and primary health cares. Furthermore, the proposed image processing and editing features reduce the burden in detecting the complex signs of DR in fundus image and create a sharing platform to discuss the DR cases among the users. This paper covers the analysis, design, development and testing phases of DRCS and place more details on the development part of the system. The project contributed to the medical and education fields. Furthermore, it sparks the users' awareness of computer technology that is useful through this application. The findings of this study will positively impact medical and education fields as it helps increase the efficiency among medical experts in treating eye diseases. This project would be a point of reference or benchmark for other eye disorders such as radiology, cornea, skin and hypertensive retinopathy.

ACKNOWLEDGMENT

This research is part of Master research currently being carried out at the Faculty of Information and Communication Technology, Universiti Teknik al Malaysia Melaka (UTeM) This research is fully funded by UTeM through Short Term High Impact Research Grant (PJP/2020/FTMK/HI17/S01715). We are grateful and would like to convey our sincerest gratitude to all primary health cares who have participated in the survey and testing phase.

REFERENCES

[1] Devaraj, D., Suma, R., & Prasanna Kumar, S. C. (2018). A survey on segmentation of exudates and microaneurysms for early detection of diabetic retinopathy. Materials Today: Proceedings, 5(4), 10845–10850. https://doi.org/10.1016/j.matpr.2017.12.372.

[2] World Health Organization. (2019). World report on vision. In World health Organization (Vol. 214, Issue 14).

[3] Chua, J., Lim, C. X. Y., Wong, T. Y., & Sabanayagam, C. (2018). Diabetic retinopathy in the Asia-pacific. Asia-Pacific Journal of Ophthalmology, 7(1), 3–16. https://doi.org/10.22608/APO.2017511.

[4] Ministry of Health Malaysia. (2017). Diabetic Retinopathy Screening Module KKM.

[5] Ministry of Health Malaysia (2015). National Health & Morbidity Survey 2015. In Institute for Public Health, Ministry of Health, Malaysia (Vol. 2).

[6] Malaysian Society of Ophthalmology. (2020). Why Screen for Diabetic Retinopathy? https://www.mso.org.my/index.cfm?&menuid=18.

[7] Rahim, S. S., Palade, V., Shuttleworth, J., & Jayne, C. (2016). Automatic screening and classification of diabetic retinopathy and maculopathy using fuzzy image processing. Brain Informatics, 3(4), 249–267. https://doi.org/10.1007/s40708-016-0045-3.

[8] Kamble, V. V., & Kokate, R. D. (2020). Automated diabetic retinopathy detection using radial basis function. Procedia Computer Science, 167(2019), 799–808. https://doi.org/10.1016/j.procs.2020.03.429.

[9] Larkin, C., Chatra, C., Sreehari, H. P., Kumar, G. R., & Poston, T. (2014). ForusCare: An integrated teleophthalmology screening system.

2014 International Conference on the IMpact of E-Technology on US, IMPETUS 2014, 1–5. https://doi.org/10.1109/IMPETUS.2014.6775869.

[10] Tumpa, J. F., Adib, R., Das, D., Ahamed, S. I., Kim, J., Medic, V., Castro, A., Pacheco, M. S., Rowland, R., & Romant, J. (2019). Poster Abstract: MTEH: A Decision Support System for Tele-Ophthalmology to Improve Eye Health of Wisconsin Population in Community Settings. Proceedings - 4th IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2019, 25–26. https://doi.org/10.1109/CHASE48038.2019.00018.

[11] Wu, Y., Wei, Z., Yao, H., Zhao, Z., Ngoh, L. H., Deng, R. H., & Yu, S. (2010). TeleOph: A secure real-time teleophthalmology system. IEEE Transactions on Information Technology in Biomedicine, 14(5), 1259–1266. https://doi.org/10.1109/TITB.2010.2058124.

[12] Mohammadpour, M., Heidari, Z., Mirghorbani, M., & Hashemi, H. (2017). Smartphones, tele-ophthalmology, and VISION 2020. International Journal of Ophthalmology, 10(12), 1909–1918. https://doi.org/10.18240/ijo.2017.12.19.

[13] Liou, E. C., & Cheng, S. C. (2020). A QoS Benchmark System for Telemedicine Communication over 5G uRLLC and mMTC Scenarios. 2nd IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2020, ECBIOS 2020, 24–26. https://doi.org/10.1109/ECBIOS50299.2020.9203639.

[14] Chanchal, R., Chaman, J., & Wasim, A. (2017). HEMAN : Health Monitoring and Nous. IEEE WiSPNET Conference, 17(2), 2115–2119.

[15] Yang, D. (2020). The 9 Best Programming Languages to Learn in 2020. Fullstack Academy. https://www.fullstackacademy.com/blog/nine-best-programming-languages-to-learn.

Secure Data Transmission Framework for Internet of Things based on Oil Spill Detection Application

Abhijith H V[1]
Department of Information Science and Engineering
Sai Vidya Institute of Technology, affiliated to
Visvesvaraya Technological University
Bangalore, India

Dr. H S Rameshbabu[2]
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, affiliated to
Visvesvaraya Technological University
Bangalore, India

***Abstract*—Internet of Things (IoT) is a leading technology which can interlink anything to Internet and makes everything to intelligent and smart. IoT is not just a single technology. IoT is a combination of various technologies like communication, data analytics, sensors and actuators, cloud computing, artificial intelligence, machine learning, etc. Applications of IoT are spread across various domains. IoT is most suitable for remote applications like underwater networks. One such application is oil spill detection in ocean. Oil spill in an ocean is a critical challenge that causes damages to marine ecosystem. Detection of oil spills in a real time manner helps to resolve the problem quickly to minimize the damage. IoT can be used to detect the oil spill by making use of sensors deployed at various locations of ocean. With a massive amount of sensors deployed and the huge amount of data associated with it, there remain concerns about the data management. Also amount of data generated in IoT based remote sensing network is usually enormous for the servers to process and many times data generated are redundant. Hence there is a need for designing a framework which addresses both aggregation of data and security related issues at various aggregation points. In this paper we are proposing a secure data transmission framework for detecting oil spill through IoT, which avoids redundant data transmission through data aggregation and ensures secure data transmission through authentication and light weight encryption.**

Keywords—Internet of things; wireless sensor networks; sensor nodes; data aggregation; authentication; light weight cryptography

I. INTRODUCTION

Internet of Things refers to connecting various things to the internet. IoT can be used in wide range of applications. IoT applications include remote applications also. Remote IoT applications include military applications such as surveillance, battle field monitoring etc. Underwater applications such as oil spill detection, analysis of aquatic animals lifecycle, etc. smart farming, forest applications such as fire monitoring etc.

The Application considered in this paper is oil spill detection. Oil spills are caused due to accidents in the ocean, leakage of oil, crude oil or gasoline pipelines. Oil spills are also caused due to negligence of offshore refineries or crude oil rigs. Irrespective of the cause of spill or slick, their effects are the same. Oil forms a separate film on the surface of water and stays there unless physically removed. The oil film causes harm to flora and fauna of the ocean. Detection of leakage in pipes or spillage of oil in accidents takes time and combating the pollution takes even longer time. The longer the film of oil stays on water surface, the greater is the damage caused to aquatic eco system. Due to the vastness of oceanic surface, application of Internet of Things to sense oil spills is a very practical technological solution. There are underwater robotic submarines available that can be used to deploy sensor nodes at specific locations. Buoyant nodes sense surface conditions and communicate observations to middle level nodes that in turn communicate to nodes on the bed. Levels of hierarchical node deployment allow us to implement intelligent communication of sensed data. Just knowing that oil spill has occurred is of little practical use. Instead if we also know the extent of damage, area of region polluted, some corrective combating tasks can be initialized.

Oil Spill detection is a remote IoT application, which uses resource constrained networks [22]. It includes deployment of many low cost, low power sensing nodes which has capabilities to sense, process and communicate the data. Each sensor nodes has limited transmission range, limited energy, limited processing capabilities and limited memory.

As Sensor nodes are resource constrained devices many nodes are deployed in shorter region. This leads to the sensing of redundant information. Transmitting redundant data will leads to the wastage of energy and other resources. Redundant data transmission can be controlled by data aggregation techniques. There are different data aggregation approaches: In-network data aggregation, Tree based data aggregation, cluster based data aggregation, Grid based data aggregation and hybrid data aggregation. There can be single level or multiple levels of data aggregation [21].

Along with data aggregation we need to consider the importance of secured data transmission. As Oil spill detection uses wireless communication, there remains concern about security. Fault alarm causes lot of damages as oil spill needs to be addressed in real time manner. There is need for simple and strong authentication scheme to identify the nodes identity. Also there is a need of light weight encryption algorithm to be used to secure the data transmission [19].

In this paper we are proposing a secure data transmission framework for oil spill detection applications which involves intelligent data aggregation technique and security mechanism. Proposed approach aggregates the data as well as provides intelligence to nodes to detect whether to send the data or not.

Also our present work add features to detect the boundary of oil spill which helps in identifying intensity of damage. In this paper we are adding simple ID based authentication mechanism to prove the nodes identity. We can use any of the existing lightweight cryptographic algorithm to secure the data communication.

The paper is organized as follows: Related work is given in Section 2, proposed approach is described in Section 3, Section 4 describes the simulation and results, conclusion is provided in Section 5.

II. Related Work

A lot of work has been carried out in detection of oil spills. Pangilinan et.al, have focused on determining the thickness of oil spill. Their hypothesis is that thicker the spill graver the effects of the pollution. They have determined the pixel intensity concerned with many competing parameters [1].

Reem Alattas has also worked on image analysis in oil spill detection. A threshold method is suggested to detect oil spill. SAR image processing is applied based on minimum cross-entropy with gamma distribution. A major drawback of the work is its intrinsic time lapse involved in data analysis. Oil spill like pollution must be combated as quickly as possible. Real time detection, if possible is most suitable for such problems. Additionally, the method proposed in [2] works with bi-modal images that have two classes of pixels only.

A.Gasul et.al [3] have proposed and tested a new method for SAR image analysis to detect oil spills. Their method works well with low resolution and distorted images as well. In spite of multi stages analysis algorithm implemented, ensuring error free end result is still difficult.

Kruti Vyas et.al in [4] have applied feature extraction of SAR images in oil spill detection. They have recognized three independent features for this purpose. They have performed a battery of experiments while considering types of images.

Mario Monteiro et.al, have made an extensive study of various aspects and challenges of detection of oil spills. Their work is particularly in connection with the seagull project. Their contribution is specifically in application of camera fitted unmanned aerial vehicles in combating water pollution. The author in [5] documents various challenges involved in detection of spills, with emphasis on time lapse between actual spill and its detection.

Unmanned surface vehicle is developed and analysed for performance in [6] by Deqing Liu et.al. Their work concentrates on frequent oil spills that happen at harbors, oil rigs and drilling platforms. They have designed a fluorosensor laser detector to achieve this aim. A feasibility analysis is also carried out by the authors.

Md. Shafi.K.T, et.al have developed a simple resonance based application to detect oil spill using planar microwave [7]. Proposed sensor is developed, deployed and tested in their work. The sensor designed is capable of detecting pollution beyond 5%.

There are different traditional data aggregation mechanisms, they are flat based [8][9][10] and hierarchical approaches.

In flat approach, the base station transmits a query requesting for a data from the sensing nodes within the area. The nodes which have relevant sensed information to the query sent will respond back. In this method base station performs excessive computations and communications. Because of this, if the base station fails, then network connection will be lost with the outer world.

Under hierarchical approach, many techniques have been proposed for energy efficiency and scalability. There are four types of hierarchical approach they are the cluster based data aggregation [11][12], tree based data aggregation [13], the grid based data aggregation [14]., and the chain based data aggregation [15][16].

These traditional data aggregation schemes will just combine the data from multiple sources and forward it to next node. In majority of applications normal data aggregation is not sufficient. We need to add some intelligence to the data aggregation and data transmission process within the limits of constrained resources.

J. Chen, S. Kher, and A. Somani, et al. proposed a scheme in [15] which involves Majority Voting. This approach is a data outlier detection method based on spatial correlation. Here if a reading of local sensor node is different from majority of its neighbouring nodes, then it will be classified as abnormal.

Y. Sun, H. Luo, and S. K. Das et al. proposed a scheme in [16]. Here according to the trustworthiness ranked by comparison with historical data and neighbour data, Weight will be assigned to every sensor data. Then weighted mean value will be calculated at the aggregator. This will be considered as the aggregated data.

S. Din, A. Ahmad, et al. proposed a scheme in [17], where the nodes closer to sink node performs direct communication and remains unclustered, the nodes which are one-hop away from the sink node perform multi-hop communication, and remains clustered.

Bo Yin et. al. in [18] specified a Tree based scheme which Involves construction of aggregation tree for complex queries. This ensures minimum communication cost.

S. Kumar and V. K. Chaurasiya et all in [20] proposes a scheme, where data Mining techniques are used to generate more accurate, consistent and useful information than that generated by any individual sensor node.

There exists different lightweight cryptographic techniques like AES [23], HEIGHT [24], PRESENT [25], DESLX [26], RSA [27] etc.

III. Proposed Work

A. Architectural Setup and Assumptions

Following are the assumptions made in deployment of nodes in the test bed.

1) Grid based deployment of nodes: to specifically determine location of each individual node. This assumption is safe and valid as, there is a commercially available underwater robot capable of deploying nodes at specific coordinates of latitude and longitude.

2) Number of nodes per grid is fixed: this assumption helps us compute statistically significant results after data collection are done and analysis is to be performed.

3) Node density is uniform: non-uniform node deployment does not allow implementation of intelligence.

4) Each node in a grid represents a single unit area: data collected is representative of a fixed area under surveillance.

5) 4 unit squares form a cell: helps in aggregation of data across levels of nodal deployment.

6) Every cell has a cell-sink node: this is an architectural requirement of WSN.

7) Four cells combine to form a level 2-cell: helps implement the second layer of intelligence and additional aggregation.

8) One of the unit cell's cell-sink also works as a sink for four adjacent level-2 cells: based on energy levels, s suitable cell-sink is elected as the sink of adjacent cells.

9) All level-2 cells combine to form the region under observation: highest level of aggregation implemented under our scheme.

*10)*Each node will communicate the observation only if the quantity observed is above a fixed threshold.

B. Node Deployment Phase and Network Establishment

1) Four different levels of nodes are deployed.

- Ground level sensor nodes: They are capable of sensing the oil density at the preliminary level and intensity of light.
- High power node at the ground level: They acts like grid head they can also sense the information.
- Anchor nodes at middle level: They act like second level grid head. They gather the data from multiple grids and transfer it to surface buoyant node.
- Surface buoyant node: They are at the surface level, they can sense the PH level of the water.

2) These nodes are deployed at the required locations using an underwater automatic vehicle (UWAV).

3) The area of interest is divided into equal sized grids. Ground level nodes and high power nodes are deployed at each grid. High power node will be the grid head of that particular grid.

4) Each High power node broadcast a beacon message to the ground level sensor node to indicate its presence in the grid.

5) Each ground level nodes respond back to their grid heads by sending the response beacon. Same Procedure is followed between grid heads and anchor nodes, Anchor nodes and surface buoyant nodes.

Fig. 1 shows the network architecture used in the proposed scheme.

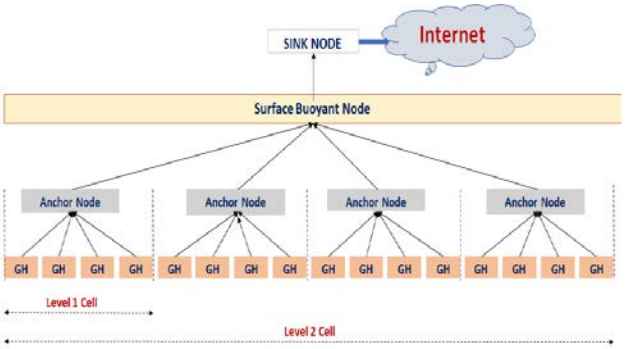


Fig. 1. Network Architecture.

C. Routing and Aggregation Phase

1) All the ground level sensor nodes sense the data and send the sensed data to its Grid head. Grid head aggregates the received data.

2) Grid Head adopts Spatial aggregation: based on the percentage of area generating similar readings. Each unit cell has a designated high power node acting as cell-sink. Communicate reading only if at least three out of four nodes report similar readings.

3) Grid Head sends data to Anchor node. Anchor node aggregates the data. Anchor node also adopts spatial Aggregation.

4) Anchor nodes send the aggregated data to surface buoyant node. Surface buoyant node aggregates and sends the data received as well as PH information to sink.

5) Surface buoyant node adopts Boundary value aggregation: region under observation has a sink node that acts as a data aggregator. Radial survey is made periodically at three lengths of radii- minimum, nominal and maximum. Minimum is close to the sink, maximum is distance from sink to boundary of region under observation, and nominal is an intermediate distance.

D. Boundary Detection

In the proposed system, a coordinate table is created. This table is a list data structure that stores coordinates of each node in the IOT infrastructure. Each entry consists of x, y and x coordinate values stored in an indexed linear linked list. An additional field stored data communicated from the sensor node.

Aggregation as well as intelligence is applied at middle level nodes and at bed nodes respectively. Boundary determination is implemented in the following steps:

1) Search the list of nodes and segregate nodes that have updated sensed data.

2) Perform a sequential search among the segregated nodes to determine nodes with maximum positive and negative coordinates of x, y and x.

3) Step 2 gives probable boundary nodes, compare their observed data with neighbor nodes to determine actual boundary nodes.

4) Communicate boundary coordinates.

5) Rate of spread is determined by calculating (a2-a1)/t, where a2 is the new area of oil spill after elapse of time unit t, a1 is the previously computed area of oil spread, i.e. before the elapse of time t.

E. Security

Assumptions

1) Grid Head, Anchor node, Surface buoyant nodes are considered as aggregator nodes.

2) Ground level sensor nodes are subordinate nodes for Grid heads, Grid heads are subordinate nodes for anchor node and anchor nodes are subordinate nodes for surface buoyant nodes.

3) Every ground level sensor node and other subordinate nodes maintains the following in its memory.

a) A pre shared Secret key shared with its aggregator node (Grid Head, Anchor node, Surface buoyant node).

b) A pre shared secret information (used for Authentication) shared with aggregator node.

4) Every aggregator node maintains the following in its memory.

a) Subordinate node ID.

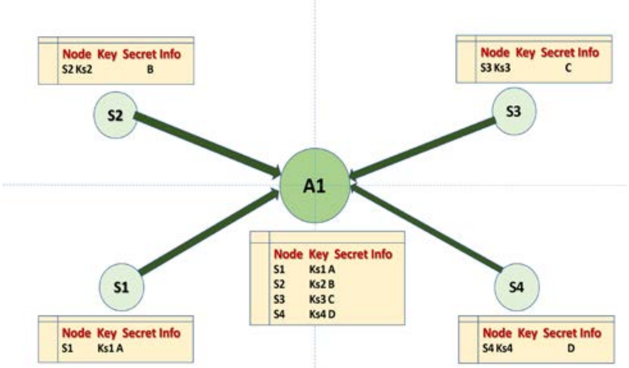
b) Pre Shared Secret Key for encryption and decryption.

c) Pre shared Secret Info for authentication.

Table I shows the Secret information repository format maintained by nodes. Fig. 2 shows the security architectural setup followed in the network.

TABLE I. SECRET INFORMATION REPOSITORY

Node ID	Key (Secret key shared with subordinate node)	Secret Info (Pre Shared Secret Information used for authentication)



Node	Key	Secret Info
S2	Ks2	B

Node	Key	Secret Info
S3	Ks3	C

Node	Key	Secret Info
S1	Ks1	A
S2	Ks2	B
S3	Ks3	C
S4	Ks4	D

Node	Key	Secret Info
S1	Ks1	A

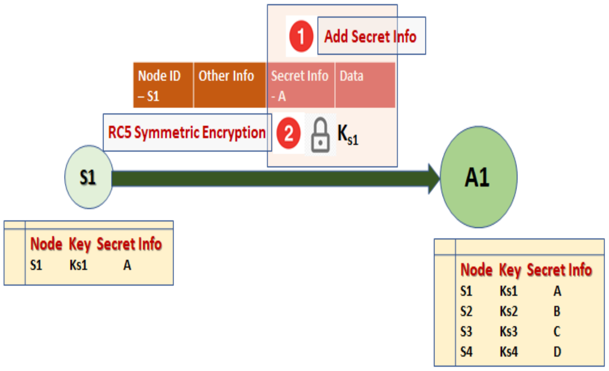
Node	Key	Secret Info
S4	Ks4	D

Fig. 2. Security Architectural Setup.

Authentication Scheme:

Sender Side:

Fig. 3 shows the authentication and encryption procedure adopted by sender node.



Node	Key	Secret Info
S1	Ks1	A

Node	Key	Secret Info
S1	Ks1	A
S2	Ks2	B
S3	Ks3	C
S4	Ks4	D

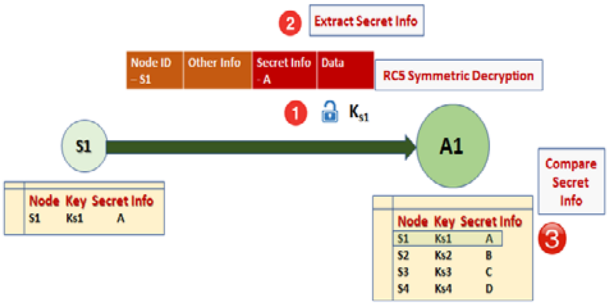
Fig. 3. Sender Side Authentication Procedure.

Pre shared secret info A will be prefixed to the data.

Secret Info and the Data will be encrypted using pre shared secret key.

Receiver Side:

Fig. 4 shows the authentication and encryption procedure adopted by receiver node.



Node	Key	Secret Info
S1	Ks1	A

Node	Key	Secret Info
S1	Ks1	A
S2	Ks2	B
S3	Ks3	C
S4	Ks4	D

Fig. 4. Receiver Side Authentication Procedure.

Received packet will be decrypted using Pre share secret key.

Secret information A will be taken from the header

Node ID and Secret information will be compared with the key repository stored at the aggregator node to prove the authenticity.

Encryption Scheme:

Any Light weight cryptographic scheme can be used to encrypt the data and secret Info at sender side. The list of popular light weight cryptographic algorithms and its comparisons are given in Table II [28]:

TABLE II. LIGHT WEIGHT CRYPTOGRAPHY TECHNIQUES

Light weight Cryptographic scheme	No. of Rounds	Block Size (bits)	Key Size (bits)	Structure
AES	10/ 12/ 14	128	128/192/256	SPN
HEIGHT	32	64	128	GFS
PRESENT	31	64	80/128	SPN
LEA	24/28/32	128	128/192/256	Feistel
TEA	64	64	128	Feistel
RC5	1-255	32/64/128	0-2040	Feistel
DESL	16	64	54	Feistel
Hummingbird	4	16	256	SPN

IV. SIMULATION AND RESULTS

MATLAB is used as the simulation tool to check the efficiency of proposed work. Initially area of interest with dimensions 800m (length) x 800m (breadth) x 800m (depth) is considered. Then the volume under consideration is divided into equal sized grids of size 50m each at ground level. Table III shows the various parameters used in simulation. Fig. 5 depicts the 3 dimension test bed created through simulation and Fig. 6 shows Grid layout in 2 Dimension. Here blue color nodes represent nodes deployed at ground level. Red color nodes represent anchor nodes and the magenta color nodes indicate nodes that are buoyant on oceanic surface.

TABLE III. SIMULATION PARAMETERS

Parameter Type	Parameter Value
Area of application region	800 X 800 m^2
Sea Depth	800 m
Grid range	100 m
Number of Ground level node in each grid	4
Number of High power nodes in each grids	2
Number of Anchor node in each group of grids	1
Number of Surface buoyant nodes in each group of grids	1
Simulation Time	150sec
Payload length	512 Bytes
Parameters of Ground level node	
Initial Energy	4J
Transmission range	20m
Data rate	4kbps
Parameters of High power node	
Initial Energy	7J
Transmission range	25m
Data rate	6kbps
Parameters of Anchor node	
Initial Energy	10J
Transmission range	30m
Data rate	8kbps
Parameters of Surface buoyant node	
Initial Energy	14J
Transmission range	35m
Data rate	12kbps

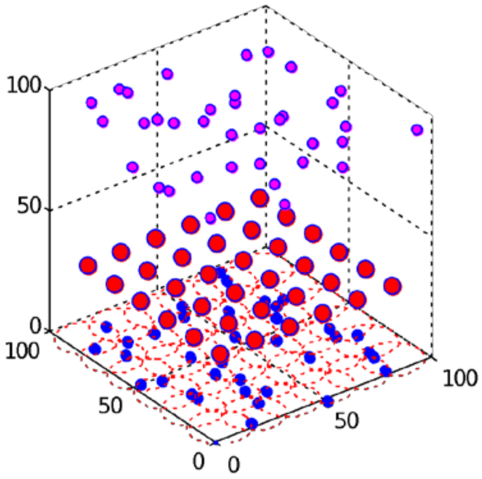


Fig. 5. 3-Dimension Test Bed.

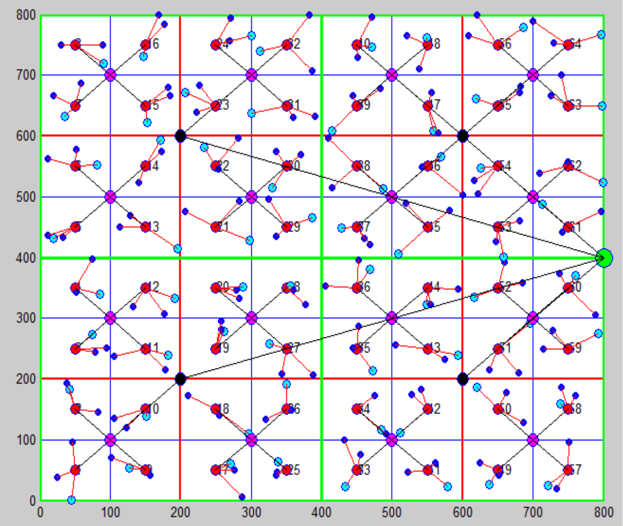


Fig. 6. Grid Layout.

Fig. 7 shows amount of time required to detect the oil spill versus the oil spread area. Time required to detect spills of larger area is evidently greater. From this we can analyse the spreading rate of oil in the water and intensity of oil leakage. As underwater communication is slower than normal environment redundant information transmission can be reduced to ensure the timely delivery of information and to avoid collision in the network Figure 8 helps us understand that the number of redundant transmission in various techniques compared with proposed work are much greater. Proposed method requires lesser number of packet transmissions compared to without aggregation deployment, in-network architecture and cluster based approach. It is clear that proposed work reduces redundant data transmission significantly.

The performance of various lightweight cryptographic techniques is shown in Table IV. Based on this suitable technique can be adopted [28].

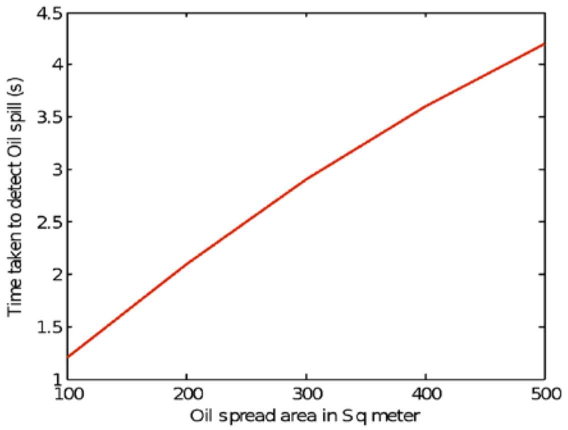


Fig. 7. Time Required for Detecting the Spreading of Oil Leakage.

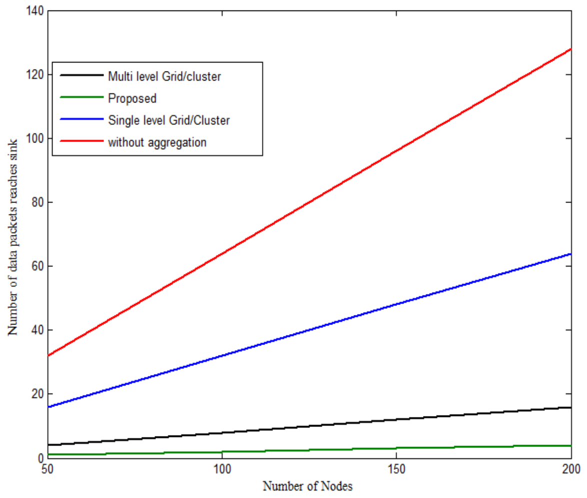


Fig. 8. Number of Redundant Transmission in Various Approaches.

TABLE IV. PERFORMANCE OF VARIOUS LIGHT WEIGHT CRYPTOGRAPHY TECHNIQUES

Light weight Cryptographic scheme	**Flash Memory (in Bytes)**	**Cycles (Encryption)**	**Cycles (Decryption)**
AES	3410	3766	4558
HEIGHT	5672	2449	2449
TEA	1140	6271	6299
DESL	3098	8365	7885

V. CONCLUSION

Adding intelligence to oil spill detection makes detection real-time and provides insight in to combating the pollution in the future. Collecting raw data triggered by an event gives minimal insight into crucial facts like thickness of oil film, expanse of oil spill and also reduces number of redundant transmissions. By detecting the boundary of oil spill we can easily predict the intensity of oil spread in the sea. This enables faster remedial actions. Also proposed methods eliminate the redundant data transmission through the spatial and boundary value aggregation. Proposed method secures the data transmission by adopting simple ID based authentication and existing light weight cryptography techniques. Adding intelligence to sensor nodes to make decision on oil spill can be considered for future enhancement.

REFERENCES

[1] M. N. Pangilinan, R. Anacan and R. Garcia, "Design and Development of an Oil Spill detection and Transmission System Using Artificial Illumination Using LEDs," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), Bangkok, 2016, pp. 407-412, doi: 10.1109/ISMS.2016.61.

[2] Reem Alattas, "Oil Spill Detection in SAR Images Using Minimum Cross-Entropy Thresholding", 7th International Congress on Image and Signal Processing, IEEE 2017.

[3] A. Gasull, X. Fábregas, J. Jiménez, F. Marqués, V. Moreno and M. A. Herrero, "Oil spills detection in SAR images using mathematical morphology," 2002 11th European Signal Processing Conference, Toulouse, France, 2002, pp. 1-4.

[4] Kruti Vyas, Pooja Shah, Usha Patel, Tanish Zaveri, Rajkumar, "Oil spill detection from SAR image data for remote monitoring of marine pollution using light weight imageJ implementation", 5th Nirma University International Conference on Engineering (NUiCONE), IEEE 2015.

[5] M. M. Marques et al., "Oil spills detection: Challenges addressed in the scope of the SEAGULL project," OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 2016, pp. 1-6, doi: 10.1109/OCEANS.2016.7761019.

[6] Deqing Liu, Xiaoning Luan, Feng Zhang, Jiucai Jin, Jinjia Guo, Ronger Zheng, "An USV-based Laser Fluorosensor for Oil Spill Detection", Tenth International Conference on Sensing Technology, IEEE 2016.

[7] Muhammed Shafi K. T., Nilesh Kumar Tiwari, Abhishek Kumar Jha, and M. Jaleel Akhtar, " Microwave Planar Resonant Sensor for Detection of Oil spills", IEEE, 2016J.Kulik.W.R.Heinzlman & H.Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks", WirelessNetwork, Vol 8, Mar-2002, pp.169-185.

[8] C.Intanagonwiwat,R.Govindan and D.Estrin, "Directed Diffusion: A Scalable and robust communication paradigm for sensor networks", 6th annual international conference on mobile-computing and networking, Aug-2000.

[9] B.Krishnamachari and J.Heidemann, "Application specific modeling of information routing in wireless sensor networks", IEEE international performance, computing, communications conference, Vol-23, pp.717-722, 2004.

[10] W.R Heinzelman, "Application-specific protocol architectures for wireless networks", PhD-Thesis, MIT, June-2000.

[11] O.Younis and S.Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks" IEEE transactions on mobile computing, Vol-3, Dec-2004,pp.366-79.

[12] Lindsey.S, Raghavendra.C, "PEGASIS Power-efficient Gathering in sensor Information system", IEEE Aerospace Conference 2002, pp-1125- 1130.

[13] Y. Sun, H. Luo, and S. K. Das, "A Trust-based Framework for FaultTolerant Data Aggregation in Wireless Multimedia Sensor Networks," IEEE Trans. Depend. Sec. Comput., 2017.

[14] J. Chen, S. Kher, and A. Somani, "Distributed Fault Detection of Wireless Sensor Networks" ACM Workshop on Dependability Issues in Wireless Ad-hoc Sensor Netw. 2017.

[15] Kuong-Ho-Chen, Jyh-Ming-Huang, Chieh-Chuan-Hsiao "CHIRON: An Energy-Efficient Chain-Based Hierarchical Routing Protocol in Wireless Sensor Networks", IEEE. 2009.

[16] Liyang-Yu, Neng-Wang, Wei-Zhang, Chunlei-Zheng, "GROUP: a Gridclustering Routing Protocol for Wireless Sensor Networks", 2006 IEEE International conference on wireless-communication networking & mobile-computing, china 2006.

[17] S. Din, A. Ahmad, A. Paul, M. M. U. Rathore, and J. Gwanggil, ``A cluster based data fusion technique to analyze big data in wireless multi-sensor system," IEEE Access, vol. 5, pp. 50690 5083, 2017.

[18] Bo Yin et. all. “Communication-Efficient Data Aggregation Tree Construction for Complex Queries in IoT Applications”, IEEE INTERNET OF THINGS JOURNAL, 2018.

[19] Xiong Li et. all. “Privacy Preserving Data Aggregation Scheme for Mobile Edge Computing Assisted IoT Applications”, IEEE INTERNET OF THINGS JOURNAL, 2018.

[20] S. Kumar and V. K. Chaurasiya, "A Strategy for Elimination of Data Redundancy in Internet of Things (IoT) Based Wireless Sensor Network (WSN)," in IEEE Systems Journal 2018.

[21] Abhijith H V, & Sindhu M P. (2015). Energy efficient multilevel hierarchical data aggregation mechanism for wireless sensor networks. 2015 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2015.7154688.

[22] H.V., Abhijith and Raj, S. Deepak and Babu, H. S. Ramesh, Intelligent Boundary Determination of Oil Spill Detection Using IOT (April 21, 2018). Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018 held at Malaviya National Institute of Technology, Jaipur (India) on March 26-27, 2018, Available at SSRN: https://ssrn.com/abstract=3167315 or http://dx.doi.org/10.2139/ssrn.3167315.

[23] Ahmed, E.G., Shaaban, E. and Hashem, M. "Lightweight mix columns implementation for AES." In Proceedings of the 9th WSEAS international conference on Applied informatics and communications, pp. 253- 258. 2009.

[24] Leander, G., Paar, C., Poschmann, A. and Schramm, K. "New lightweight DES variants." In International Workshop on Fast Software Encryption, pp. 196-210. Springer, Berlin, Heidelberg, 2007.

[25] Jana, S., Bhaumik, J. and Maiti, M.K. "Survey on lightweight block cipher." International Journal of Soft Computing and Engineering 3 (2013): 183-187.

[26] Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A., Robshaw, M.J., Seurin, Y. and Vikkelsoe, C. "PRESENT: An ultra-lightweight.

[27] Tawalbeh, L. A. A., & Sweidan, S. (2010). Hardware design and implementation of ElGamal public-key cryptography algorithm. Information Security Journal: A Global Perspective, 19(5), 243-252.

[28] Sohel Rana, Saddam Hossain, Hasan Imam Shoun and Dr. Mohammod Abul Kashem, “An Effective Lightweight Cryptographic Algorithm to Secure Resource-Constrained Devices” International Journal of Advanced Computer Science and Applications(IJACSA), 9(11), 2018. http://dx.doi.org/10.14569/IJACSA.2018.091137.

A Contemporary Ensemble Aspect-based Opinion Mining Approach for Twitter Data

Satvika[1]
PhD Scholar, CSE Department
Amity University
Gurugram, India

Dr. Vikas Thada[2]
Associate Professor, CSE
Department, Amity University
Gurugram, India

Dr. Jaswinder Singh[3]
Associate Professor, CSE
Department, GJUS and T
Hisar, India

***Abstract*—Aspect-based opinion mining is one among the thought-provoking research field which focuses on the extraction of vivacious aspects from opinionated texts and polarity value associated with these. The principal aim here is to identify user sentiments about specific features of a product or service rather than overall polarity. This fine-grained polarity identification about myriad aspects of an entity is highly beneficial for individuals or business organizations. Extricating these implicit or explicit aspects can be very challenging and this paper elaborates copious aspect extraction techniques, which is decisive for aspect-based sentiment analysis. This paper presents a novel idea of combining several approaches like Part of Speech tagging, dependency parsing, word embedding, and deep learning to enrich the aspect-based sentiment analysis specially designed for Twitter data. The results show that combining deep learning with traditional techniques can produce excellent results than lexicon-based methods.**

Keywords*—*Aspect-based sentiment analysis; dependency parsing; long short-term memory (LSTM); part of speech (POS) tagging; term frequency-inverse document frequency (TF-IDF)

I. Introduction

The past decade is undoubtedly dominated by data and its analytics. Now, anyone with an immense amount of data related to a domain and with the right tools to mine this colossal mountain of data is considered powerful. Surveys (either offline or online) are obsolete now as now more and more user is giving their opinions about myriad products and services on various social networking websites like Facebook, Instagram, LinkedIn, QQ, Telegram, Twitter, WhatsApp, WeChat, etc.[1]. All these social media platforms have been quite popular for exchanging opinions and sentiments and hence provide much-needed feedback about the specific products, services, vital events, organizations, or persons with the active participation of users. A lot of time, effort, and money has been put into analyzing user sentiments from a plethora of social sites especially Twitter due to the reason that tweets are limited to 140 characters[2]. This character limitation makes Twitter an easy and efficient tool for sentiment analysis, which can be pretty useful for organizations, who want to analyze related documents and make optimum changes to better suit their targeted customers. However, the sentiment analysis generally determines the overall opinion and hence may not be able to extract the precise essence needed to review that particular product or service.

This could be easily understood by the following example from a review about Cars: “Ford Mustang is awesome to drive, but the price is too high”. This particular sentence basically talks about two different features of the subject (Ford Mustang) here i.e., “Driving Comfort” and “Price”. Another notable point to comprehend is that a sentence may contain more than one aspect and the polarity of each aspect can be diverse. In the example shown above, “Driving comfort” is an undoubtedly positive polarity aspect while the “Price” aspect is declined more towards negative polarization. Hence to properly analyze the available data, a more detailed-level approach is needed, which is stated as the aspect-centered sentiment analysis. This methodology provides better insight for mining user opinions about the data under review. Overall, the sentiment analysis can be done at manifold stages that are document level, sentence phase, and aspect level[3]. Thus, in a nutshell, it can be stated that aspect-based sentiment analysis produces more fine-grained data analysis of user data.

In this research, an aspect-based sentiment mining problem is discussed on Twitter data and an efficient hybrid approach for the same has been proposed. The rest of this paper is organized as follows: Section II describes the numerous techniques used for aspect terms extraction. Section III gives an understanding of the projected framework. Section IV presents the outcomes of the proposed framework and evaluates them. Finally, Section V entails the conclusion and future work in the same direction.

II. Aspect Extraction Techniques

Sentiment analysis is usually defined as the computational study of a person’s opinions or emotions or views about a particular entity. Nevertheless, an entity has certain features associated with it, which must be considered with utmost precaution as these characteristics define opinions at the atomic level. Aspects can be briefly stated as the attributes or traits of a product or service [4]. An Aspect extraction is a primary process of identifying these significant attributes and a considerate argument here is that aspect extraction aiming at no particular aspect or target is of circumscribed usage [5]. Hence aspect extraction is the most imperative step for pulverized opinion mining. There are two sorts of aspects, namely: implicit and explicit aspects. The explicit aspects are mentioned overtly in the opinionated sentence whereas the implicit aspects are expressed indirectly and that is why more challenging to discover. For example: “Mustang is a classic car from Ford, renowned as being the muscle car”. In the above sentence, “Ford” as the “brand” of the car is an explicit aspect, while “muscle car” refers to the implicit aspect “performance”. The subsequent step to aspect extraction is finding the sentiment polarity of recognized features; like in the above example “brand” aspect is neutral and the “performance” aspect is positive.

Some of the key techniques employed for finding aspects are shown in Fig. 1.

1) Frequency-based approaches: Considered as the most traditional approach, it chiefly employees tracing out the most commonly occurring words. One such technique is Part-of-Speech (POS) tagging which basically finds out the classification of words based on the grammar within a text like a noun, pronoun, verb, adjective or adverb, etc. POS tagging is not only the simplified syntactically tagging taught in school education, but it also tries to find the relationship between the word under consideration and the adjacent words. Thus, POS tagging is not generic but depends upon the sentence to sentence and finally, it can be said that it tries to extract speech tags based on the sentence context. Also, an imperative point is that POS tagging does not actually do the Natural Language Processing (NLP), but it is the precondition to handle a lot of NLP tasks. The main characteristic of POS tagging is that it finds the recurrent noun and noun phrases from the reviews as they are usually the aspects [7]. POS tagging on a review is demonstrated in Fig. 2.

However, it must be understood that some of the aspects identified by this approach may not be the key aspects and must be rendered. Another frequency-based tactic is Pointwise Mutual Information (PMI), which is figured as the variance of the common information between the feature and the related word [8]. Last but not the least is Term Frequency-Inverse Document Frequency (TF-IDF) methodology which evaluates how significant a word is to a document in a group of the corpus by calculating two terms. The first term (TF) is the frequency of a word appearing in a document, divided by the overall words’ tally in that document. The second term (IDF) is computed as the logarithm of the whole documents in the corpus divided by the number of documents where the particular word appears [9]. In simple words, TF-IDF is the statistical tool that keenly discovers the most vital words in the articles. An important feature of TF-IDF is that some words seem noteworthy due to their huge frequency in the text, but actually, they are insignificant like the, this, that, etc. [8]. Hence, TF-IDF can be described as the amalgamation of two algorithms that diminishes the weight of stopwords and recognizes the high opinionated words, which actually influence the sentiments contained in the documents.

2) Relation-based approaches: These techniques are based on discovering relations between features and the opinion words to identify aspects clearly. One such approach is Dependency Parsing – which is based on extracting the grammatical (or syntactic) structure of a sentence and finding the relation between vital words and the words which modify these vital words [10]. The second practice is called Double Propagation, which disseminates information among opinion words and targets back and forth. This method comes under the semi-supervised category as it takes opinion word as the seed to start the process [11].

3) Supervised machine learning: The two most prominent aspect mining techniques Hidden Markov Model (HMM) and Conditional Random Fields (CRF) come under this category. HMM are a set of probabilistic visualization model that allows forecasting a sequence of hidden variables from a set of perceived variables. It is a great approach for tracing the aspects like it may predict the type of weather by observing the type of clothes worn by a person [12], as depicted in Fig. 3.

On the other hand, CRF is a discriminative model employed for predicting sequences where multiple variables are dependent on each other by imputing contextual information from previous labels [14]. Its main applications are in POS tagging and Named Entity Recognition (NER).

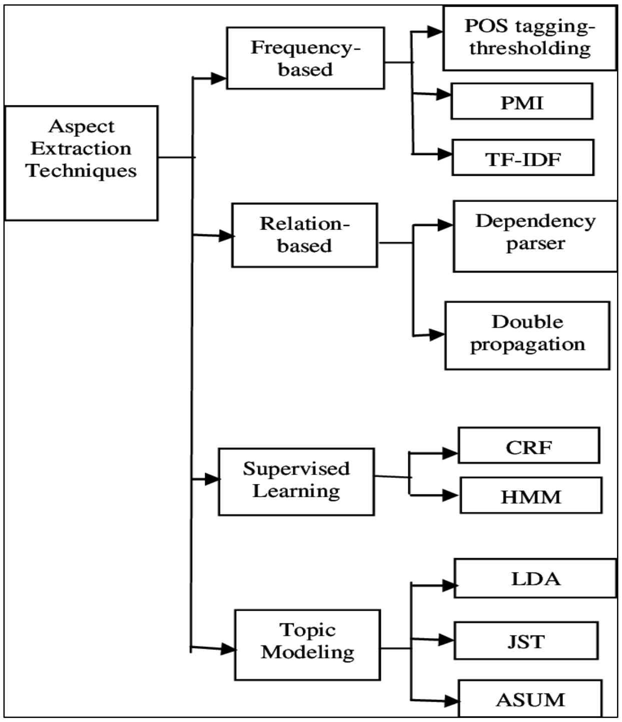
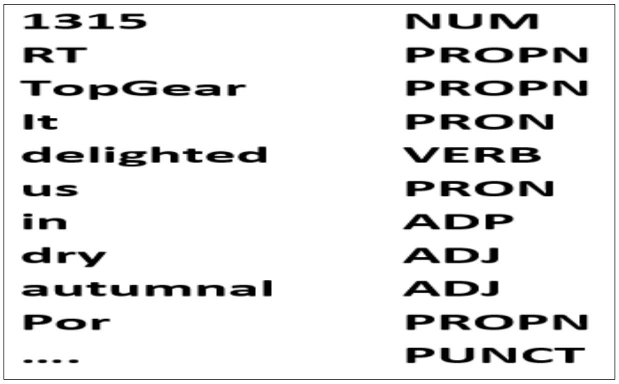


Fig. 1. Various Aspect-Extraction Approaches [6].



```
1315          NUM
RT            PROPN
TopGear       PROPN
It            PRON
delighted     VERB
us            PRON
in            ADP
dry           ADJ
autumnal      ADJ
Por           PROPN
....          PUNCT
```

Fig. 2. POS Tagging Output.

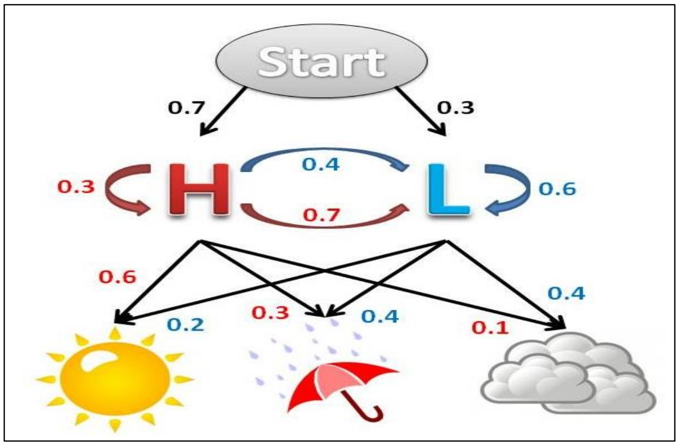


Fig. 3. Working of Hidden Markov Model [13].

4) Topic modeling: As the name suggests, it is a category of techniques used to automatically detect topics present in the text corpus. It comes under unsupervised learning for observing topics (a group of words) in a colossal amount of text clusters. Latent Dirichlet Allocation (LDA) is used to categorize text in a document to a certain topic by using Bag-of-Word (BOW) approach[15]. The prime objective of LDA is to match all the documents to the topics such that words in every document are customarily classified by those imagined topics. The next approach is Joint Sentiment Topic (JST) is the exquisite unsupervised technique that detects sentiments as well as the topics simultaneously from the text corpora by considering their mutual relations[16]. The next methodology categorized under topic modeling is the Aspect and Sentiment Unification (ASUM) model which automatically determines aspects and different sentiments towards these recognized aspects in one go. This model ascertains aspect-sentiment pair called 'senti-aspects' in an unsupervised way, explaining how much a particular word is related to aspect and sentiment [17].

III. Proposed Framework

The focal emphasis is to conduct an exploration of aspect-based sentiment analysis conducted on Twitter data. The planned work consists of the following steps:

1) Data collection from Twitter.

2) Data pre-processing and preparation.

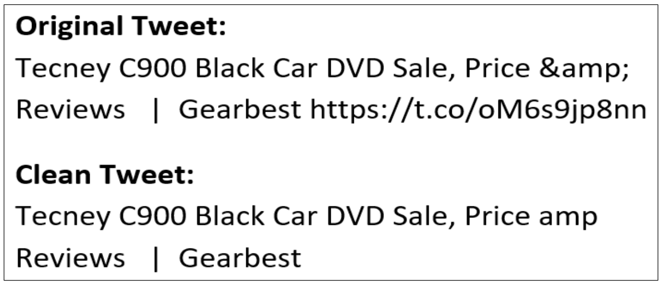
3) Aspect extraction from the data.

4) Aspect selection & polarity detection.

5) Tweet level Sentiment classification & Performance evaluation.

The first and foremost step is collecting data from Twitter, which is a social media networking site. There are numerous ways to extract data from Twitter; among which the most popular one is Twitter API Search, which allows us to retrieve the latest tweets about any topic. For doing so, a Twitter development account must be created first. Earlier, this method had a limitation on the number of tweets and also on the number of requests received per hour from a particular IP address [18].

After the data collection process from Twitter, the next work is to prepare the data for analysis and this process involves cleaning and pre-processing of data. The data from Twitter is not in a particular format and it must be cleansed beforehand like removing the links, emoticons without clear sentiment, new lines, hashtags and symbols like @, punctuation marks, etc. For effective aspect discovery, tweets, users who tweeted them, followers of users, count of retweets, date of the tweet, etc. are chosen from the whole data. After data collection and pruning, the next step is to pre-process it. The data pre-processing starts with the removal of stopwords i.e. removing the commonly occurring words like a, an, the, and, or, is, etc. as these words do not convey meaningful opinions. Then, all the words in the corpus are changed to lowercase for enhanced understanding and investigation. Afterward, the corpus is subjected to stemming and lemmatization i.e., words are reduced to their basic or root form. It is a highly recommended part because it reduces derivationally related forms of a word. For example: "programmable" is changed to "program" and "running" is converted to "run". One example of text before and after pre-processing is shown in Fig. 4.



Original Tweet:

Tecney C900 Black Car DVD Sale, Price &amp;
Reviews | Gearbest https://t.co/oM6s9jp8nn

Clean Tweet:

Tecney C900 Black Car DVD Sale, Price amp
Reviews | Gearbest

Fig. 4. The Output obtained after Applying Pre-processing on Data.

Now, the data is ready for opinion mining and the foremost thing to do is finding aspects in these tweets. It can also be done in myriad ways as discussed in the above section. The aspect-extraction followed here is a hybrid one as multiple methodologies are combined for doing so including POS tagging, Dependency, etc. POS tagging is used to find the frequent nouns and noun phrases which tend to be the explicit aspects, while depending on parsing is brilliant for finding the implicit tags. For bigrams and trigrams, tokens are identified by checking words on the left side for 'proper Nouns' identified earlier. If the dependency is a compound or adjective modifier, then it is regarded as the aspect term. If the token is an 'adjective' or 'verb', the words on both sides (left and right) are examined for adverbial modifiers, open-clausal components, and 'auxiliary' dependencies and if it happens, they are regarded as aspect terms. For negation modification, words to the left of verbs and words on both left and right are diagnosed for adjectives and auxiliaries and in this case, the polarity of aspect is changed. Thus an ensemble of techniques is employed here to give the best output and providing every key aspect from the text corpora obtained after data preprocessing. The aspects can be ranked in decreasing order of frequency and selected on this basis. The next phase is detecting the polarity of each aspect and there are three types of polarities: positive, negative, and neutral. Positive polarity refers to the constructive and affirmative comments about the aspect under consideration while negative polarity denotes the adverse and undesirable user views regarding that precise aspect and if the opinions do not fall under any of the above categories, it is considered as neutral polarity. Also, an imperative deliberation is that if an aspect has an equal number of positive and negative opinions, then it can be ignored. The algorithm for aspect-terms extraction can be described as the following:

1) Apply Tokenization first.

2) Prepare a list of stopwords, so that aspects, if identified on the stopword list, can be ignored.

3) Apply POS tagging for finding aspects.

4) The prime candidates for aspects are 'Nouns' or 'Proper Nouns' individually or in pair with Verbs/ Adverbs or Adjectives (explicit aspects).

5) For finding implicit aspects, dependency parsing is employed, along with bigrams and trigrams:

a) For a noun, check its dependency on its L.H.S. and if it is found to be 'compound', with an adjective or adjective modifier, then it should be an aspect-term.

b) For an adjective, check its dependency parsing on both L.H.S & R.H.S., if its dependency is auxiliary on L.H.S and 'negation modifier' on R.H.S, then is also identified as an aspect-term.

c) For a verb, if an adjective appears to its left or right and their dependency is 'adverbial modifier', then also it should be an aspect. Also, if an adjective appears on its L.H.S and dependency relation is negation modifier, the aspect is recognized.

Finally, the tweet level sentiments are analyzed using Vader Sentiment and Recurrent Neural Networks, and the results determined are verified in regard to state-of-art technologies. The overall process can be described by the Fig. 5 underneath.

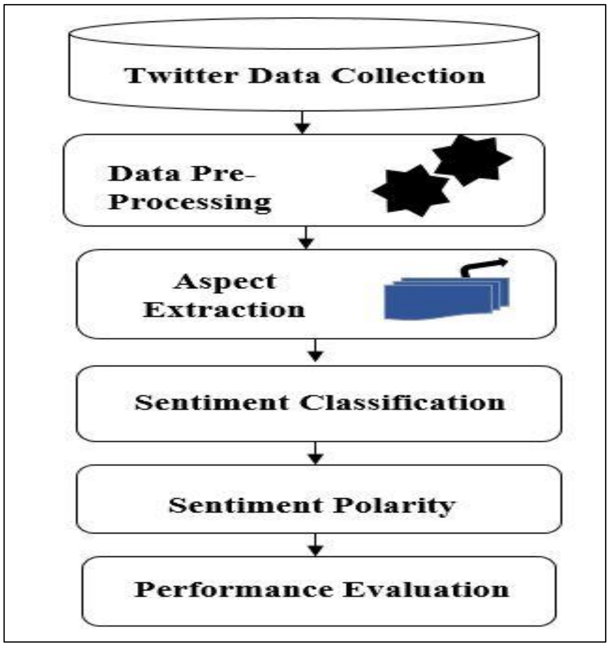


Fig. 5. Aspect-Based Sentiment Analysis Process.

IV. Results and Evaluation

The ensemble methodology for aspect-based sentiment analysis is a novel approach that does excellent work for up to 49493 tweets data corpus. However, more data means more training and better results. The new algorithm is also employed on 'Scraped_Car_Review_ford.csv', which can be downloaded from Kaggle for verification. This dataset has 49493 reviews about ford cars. Fig. 6 shows the first five tweets downloaded from the Twitter site.

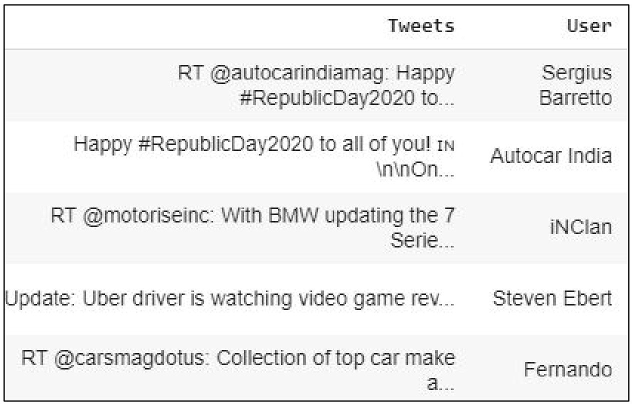
The above image Fig. 7 shows the comparison between tweet frequency and the number of followers for verified Twitter users. The data downloaded from Twitter is pre-processed as discussed in the above section and sample output is shown in Fig. 4 above. After pre-processing of data including tokenization and stemming, POS tagging is applied. After obtaining POS tags for each token, some special words with peculiar tags are selected and these are Nouns, Noun Phrases (like NNP, NNS & NNPS), Adjectives, and Adverbs because these tags are extremely helpful while identifying aspects.

The wordcloud is primarily an innovative tool that generates an aesthetic amalgamation of frequently occurring words. It puts together recurrent textual data into a beguiling visual representation and shows them in bigger and bolder fonts according to their frequency [19]. Fig. 8 shows the potential aspects identified by POS tagging and frequency distribution:

Fig. 9 demonstrates the top 20 frequent words occurring in the tweets according to their frequencies arranged in descending order.

While Noun and Noun phrases are the potential aspects, adjectives and adverbs support their claim when the dependency is checked between them. This is identified by applying dependency parsing between them, as illustrated below in Fig. 10.

After all the aspects are extracted, their polarity needs to be checked and for that Sentiwordnet can be used to check the opinion sentiment for all the mined aspects according to the tweets they are appearing in. Finally, the overall sentiment classification can be calculated using Vader Sentiment, which is a proficient tool for mining the sentiments. Fig. 11 shows the overall sentiment distribution of the downloaded tweets.



Tweets	User
RT @autocarindiamag: Happy #RepublicDay2020 to...	Sergius Barretto
Happy #RepublicDay2020 to all of you! IN \n\nOn...	Autocar India
RT @motoriseinc: With BMW updating the 7 Serie...	iNClan
Update: Uber driver is watching video game rev...	Steven Ebert
RT @carsmagdotus: Collection of top car make a...	Fernando

Fig. 6. Sample of Tweets Downloaded from Twitter.

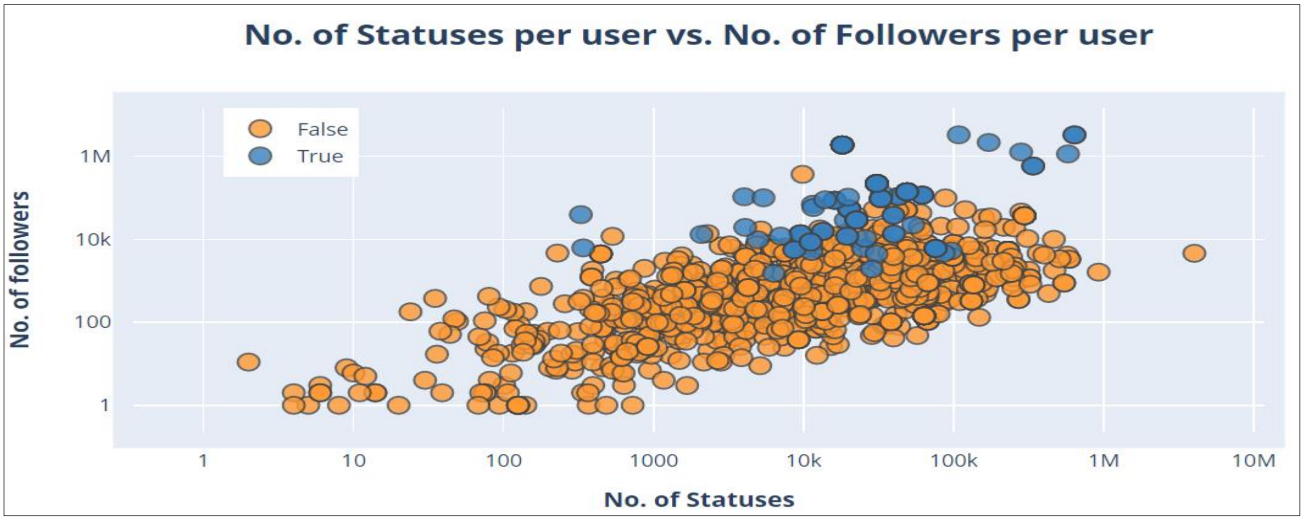


Fig. 7. No. of Statuses vs. No. of Followers.

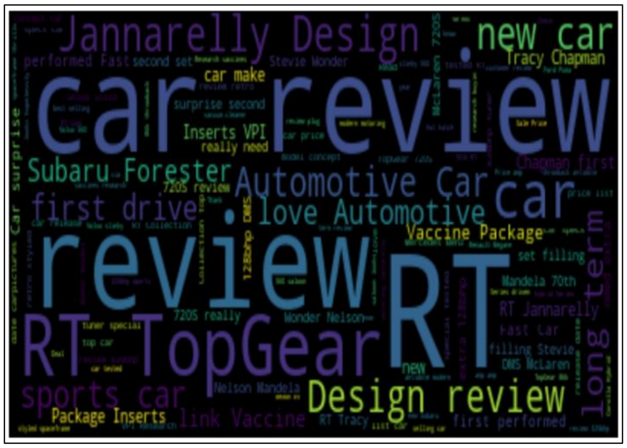


Fig. 8. Wordcloud Showing Potential Aspects.

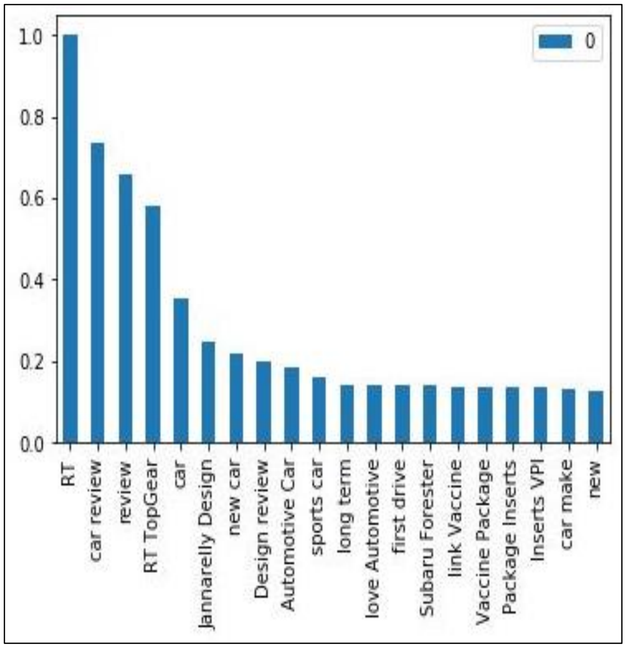


Fig. 9. Frequency Distribution of Potential Aspects.

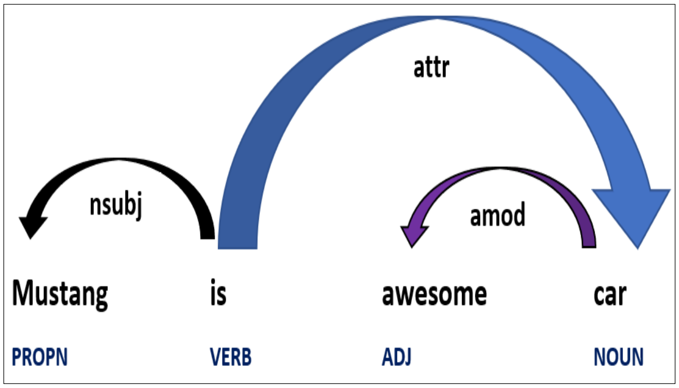


Fig. 10. Dependency Graph Example.

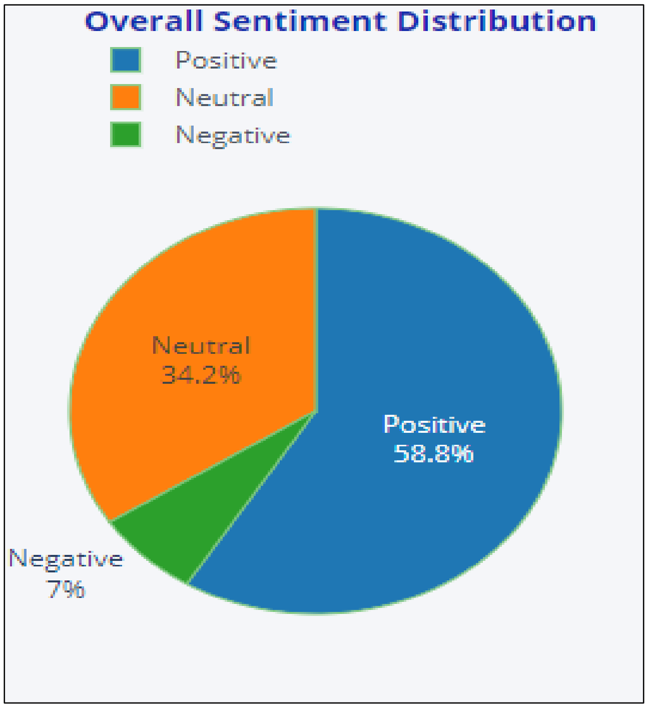


Fig. 11. Overall Sentiment Distribution.

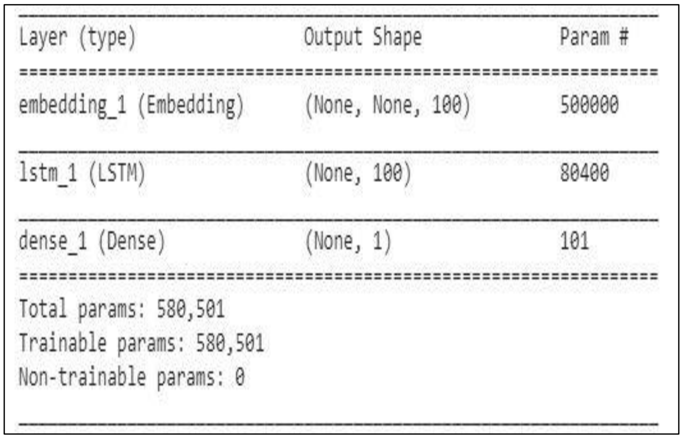
An alternate sentiment categorization can be done using deep neural networks with the aid of TensorFlow and Keras. The model proposed in this paper is quite simple and contains three layers: embedding layer, a layer with LSTM functionality, and last but not least an output layer. The embedding layer will learn word embeddings for each, and every word contained in the corpus through pre-trained word embeddings provided by Google's word2vec. The second layer is of Long Short-Term Memory Networks (LSTM), which is an exceptional branch of Recurrent Neural Networks that are skilled in finding and learning relationships between features of an input sequence. An LSTM layer with 100 memory units is engaged in the model. Lastly, the dense layer provides uses a sigmoid activation function as it needs to classify the polarity of a tweet as either positive or negative. The summary of LSTM model used in this study is displayed in the Fig. 12.

The model once built, must be tested for accuracy to check if it is doing the intended work with efficacy or not. Table I lists the results of the proposed model, primarily the model score and its accuracy on the train and test data.

The validation loss and validation accuracy can also be identified for checking the model's performance. As evident from the Fig. 13, the model gave the best output till 5 epochs and then its performance decreases abruptly till 15 epochs. Hence, the epochs can be decreased to give a superior result.

As apparent from Fig. 14 beneath, the validation loss is quite high and hence the model overfits.

To conclude, the proposed model is evaluated for train and test data. The model shows 82% accuracy for training data but 99% for test data, which shows overfitting of the model.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 100)         500000
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 101
=================================================================
Total params: 580,501
Trainable params: 580,501
Non-trainable params: 0
_________________________________________________________________
```

Fig. 12. Proposed Model's Summary.

TABLE I. RESULTS OF MODEL EVALUATION

Model Score	1.014088016230464
Model Accuracy (Train Data)	82.11%
Model Accuracy (Test Data)	99.95%

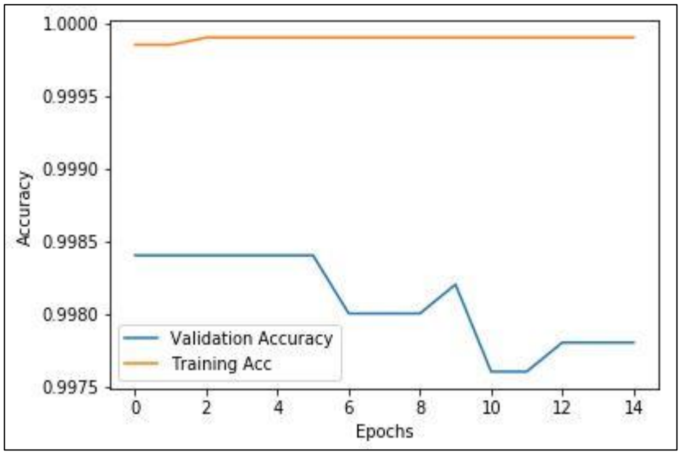


Fig. 13. Validation Accuracy of Proposed Model.

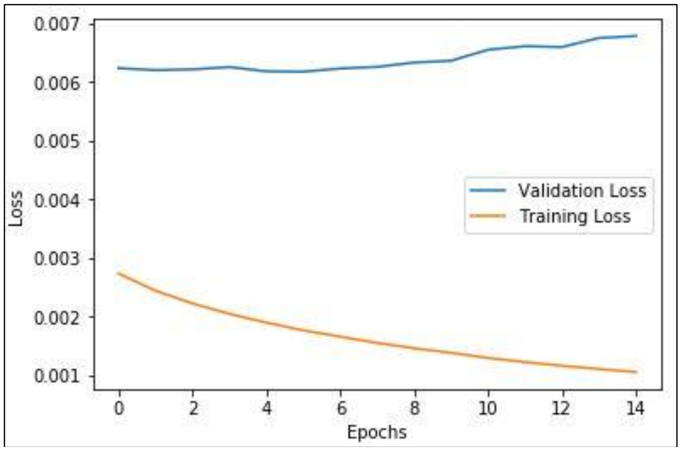


Fig. 14. Validation Loss of Proposed Model.

V. Conclusion and Future Work

The opinion mining of user reviews about a specific product or service has been under the researcher's hammer for quite a long time now. Most of the work revolves around finding sentence-level sentiment analysis, which gives a bigger but not so clear picture of user's opinions. Hence a fine-grained approach that focuses more on the aspect-level sentiment classification is needed. This paper proposes an ensemble aspect extraction methodology based on POS tagging, dependency parsing, and dense neural networks which yields better outputs for sentiment analysis instead of using solo traditional methods. This research will be valuable for the individuals or commercial entities that may employ it for fine graining the existing users' opinion about specific features of their product or service and also targeting the potential customers. On the flip side, the model overfits the Twitter dataset under consideration. The future work of this research will be manifold.

Firstly, some parameters of the model must be modified to avoid overfitting. Secondly, several machine learning and deep learning classifiers can be employed for enhanced sentiment analysis. Finally, the most important prospect will be deploying it for more human languages like French, Spanish, Hindi or Tamil, etc.

REFERENCES

[1] A. Gural, B. B. Cambazoglu, and P. Karagoz, Sentiment Focused Web Crawling, *CIKM'12, ACM Transactions on the web*, Maui, HI, USA, 2012.

[2] O. Alqaryouti, N. Siyam, A. A. Monem, and K. Shaalan, Aspect-Based Sentiment Analysis Using Smart Government Review Data, *Applied Computing and Informatics*, 2019.

[3] K. Bafna, and D. Toshniwal, Feature based summarization of customers' reviews of online products, In: *Proc. Comput Sci 22*, 2013, pp. 142–151.

[4] Y. Zhang, and W. Zhu, Extracting implicit features in online customer reviews for opinion mining, In*: Proc. 22nd international conference on World Wide Web companion. International World Wide Web Conferences steering committee,* 2013, pp. 103–104.

[5] R. Kumar, and R. Vadlamani, A Survey on Opinion Mining and Sentiment Analysis:Tasks, Approaches and Applications, *Knowledge Based Systems*, Vol. 89, No. 1, pp. 14-46, 2015.

[6] P. More, and A. Ghotkar, A Study of Different Approaches to Aspect-based Opinion Mining, 2016.

[7] B. Liu, and L. Zhang, A survey of opinion mining and sentiment analysis, *In: Proc. Mining Text Data. Springer US*, 2012, pp. 415–463.

[8] A. M. Popescu, and O. Etzioni, Extracting product features and opinions from reviews, *Natural language processing and text mining. Springer,* pp. 9–28, 2007.

[9] C. Quan, and F. Ren, Unsupervised product feature extraction for feature-oriented opinion determination, *Inf. Sci. 272*, pp. 16–28, 2014.

[10] K. Liu, L. Xu, and J. Zhao, Co-extracting opinion targets and opinion words from online reviews based on the word alignment model, *IEEE Trans Knowl Data Eng.*, Vol. 27, No. 3, pp. 636–650, 2015.

[11] G. Qiu, B. Liu, J. Bu, and C. Chen, Opinion word expansion and target extraction through double propagation, *Comput. Linguist*, Vol. 37, No. 1, pp. 9–27, 2011.

[12] W. Jin, and H. H. Ho, A novel lexicalized hmm-based learning framework for web opinion mining, In: *Proc. 26th annual International Conference on Machine Learning,* 2009.

[13] L. Rabiner, Predicting The Weather with Hidden Markov Models, [Blog], Retrieved from URL *http://guizzetti.ca/blogs/lenny/2012/04/predicting-the-weather-with-hidden-markov-models*, 2012.

[14] L. Chen, L. Qi, and F. Wang, Comparison of feature-level learning methods for mining online consumer reviews. *Expert Syst Appl.*, Vol. 39, No. 10, pp. 9588–9601, 2012.

[15] B. Ma, D. Zhang, Z. Yan, and T. Kim, An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews, *J Electron Commer Res*,. Vol. 14, No. 4, pp. 304–314, 2013.

[16] F. Li, M. Huang, and X. Zhu, Sentiment analysis with global topics and local dependency, In: *Proc. Twenty-Fourth AAAI Conference on Artificial Intelligence,* 2010.

[17] B. Liu, and L. Zhang, "A survey of opinion mining and sentiment analysis", *Mining text data. Springer, Boston, MA*, pp. 415-463, 2012.

[18] N. Zainuddin, A. Selamat, and R. Ibrahim, Hybrid sentiment classification on twitter aspect-based sentiment analysis, *Applied Intelligence*, Vol. 48, No. 5, pp. 1218-1232. 2018.

[19] Y. Jin, Development of word cloud generator software based on python, *Procedia engineering*, Vol. 174, No. 1, pp. 788-792, 2017.

Ultra-key Space Domain for Image Encryption using Chaos-based Approach with DNA Sequence

Ibrahim AlBidewi[1]
Department of Information System,
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Nashwan Alromema[2]
Department of Computer Science
Faculty of Computing and Information Technology-Rabigh
King Abdulaziz University, Rabigh, Saudi Arabia

***Abstract*—Recently, image encryption has taken an importance especially after the dramatic evolution of the internet and network communication. The importance of securing the images contents is due to the simplicity of capturing and transferring of digital images in various communication media. Although there are many approaches for image encryptions, chaos-based image encryption approach is considered one of the most appropriate approaches because of its simplicity, security, and sensitivity to the input parameter. This research paper presents a new technique for encrypting RGB image components using nonlinear chaotic function and DNA sequence. A new image with the same dimensions of the plain-image is used as a key for confusions and diffusion process for each RGB components of plain-image. Experimental results show the efficiency of the proposed technique, simplicity, and high level of resistant against several cryptanalyst.**

Keywords—Chaos-based; image encryption; confusion; diffusion; color image; RGB components; DNA sequence

I. Introduction

Images are the most common and popular multimedia data type now a days, due to the simple and easy way of capturing and transmitting [1]. The demand to transfer these images in a secure manner has also increased, and encryption is the preferred method to securely transfer image data. In addition, there are several drawbacks and weakness such as the requirement for powerful computing system and highly computational time, thus the implementation of these techniques cause low level of efficiency and it cannot guarantee data confidentiality and security [2]. Encrypting images that are sent over various communications channels has become one of the most important processes of most of the current applications and fields, such as military, educational, medical, industrial and social media [3], [4]. Image encryption can be defined as the use of set of process called algorithm to convert the plain image into a ciphered image in such a way that no one can recover it except for the sender and the intended recipient [2] [5]. A numerous image encryption techniques have been proposed, one of the most efficient techniques is the Chaos-based encryption [6]. Chaos-based encryption methods first introduced by Fridrich [7], [8], these methods are popular due to their randomness, unpredictability, sensitivity and topological transitivity. Fridrich suggested two processes for chaos-based image encryption: confusion and diffusion. The most important stages in image encryption is confusion which is considered about pixels position in plain image. Many efforts tried to kill the pixels neighbors' dependence by exchange the positions under certain condition to maintain the correlation of the pixels and the predictive of new pixels position [9]. We list in this section some of the related works starting with Choi et al. who proposed a framework by the using addition, rotation and XOR to achieve confusion and diffusion for the plain image. In the proposed framework the confusion-diffusion process is done by implement the rotation and XOR operation with chaotic sequences which are generated by the using of two logistic maps [10]. Another study by Bashir et al. proposed image encryption framework, this framework a 4-D chaotic image encryption technique based on a mechanism of dynamic state variables to increase the security and effectiveness of the chaos-based image encryption methods [11]. Kulsoom et al. proposed algorithm that is based on stream cryptography and it use DNA complementary rules in addition to one dimensional chaotic maps [12]. Kar et al. proposed bit-plane image encryption method chaotic, cubic, and quadratic maps. The proposed method is based on permutation, diffusion, and pixel randomization process, at first the proposed method generates chaotic two sequences by the using of the quadratic and the cubic map, then the generated two sequences will be used to shuffle the plain image. the shuffled image will decomposed into its bit-plain components to be encrypted later by the using of confusion and diffusion process [13]. Gu et al. proposed encryption algorithm for JPEG2000 images. The method suggested the use of bitwise XOR and cyclic rotation operation for 2-byte block encryption process also the repeating of encryption process is adopted to avoid an unwanted encryption marker code. The repeating of encryption process can neglect unnecessary computations [14]. Enayatifar et al. proposed a method for image encryption that is based on chaotic map and deoxyribonucleic acid (DNA). The process is started by convert two dimensional plain image into one dimensional array, then the process of pixel permutation is implemented by the using of chaotic map and deoxyribonucleic acid (DNA) while the diffusion is implemented by the using of DNA sequence and DNA operator both of permutation and diffusion of image pixels are done at the same time to reduce the sending time [15]. The remaining part of this article is structured as it follows: Section 2 introduces the background of image cryptography;

This project was funded by the Deanship of Scientific Research (DSR) under grant no. G- 526-611-1431.

The authors, therefore, acknowledge with thanks to DSR's technical and financial support representations Methodology

Section 3 introduces the proposed image encryption scheme with the key space. Section 4 describes the experimental results. Section 5 introduces the security resistance and the ultra key space of the proposed encryption scheme. Lastly, Section 6 draws the conclusion and future work.

II. Image Cryptography

Image encryption like the encryption of any other kind of digital data, it goes through the encryption process using specific encryption algorithm and a key [1]. Steganography is one of the encryptions methods that have been used in image cryptography in the last two (2) decades and in particular on watermarking [16]. The classical image encryption such as Data Encryption Standard (DES), RSA, and ADS are designed with good confusion and diffusion [17], but it has the weakness of low-level of efficiency because of the huge size and noticeable redundancy of image data [3],[18]. Due to the big size of image data, the implementation of images cryptosystems is carried out in spatial domain, frequency domain or hybrid domains of the plain images. Each of these domains has its own methodology. For example, in spatial domain, the pixel value and location in plain image can be directly encrypted using the general encryption function as shown in Equation 1.

$$E(x,y) = f[I(x,y)] \qquad (1)$$

Where *E(x,y)* is the output encrypted pixel C_i in the ciphered image, *I(x,y)* is the input plain pixel P_i in the plain image, and *f* is the encryption function. In frequency domain, image is analyzed mathematically to series of frequencies, each of these frequencies has two main components which are the amplitude and the phase shift [1]. Any changes in spatial domain image produce indirect change on its frequency domain representation. The information of frequency domain divided into two main components, and these components are high frequency components which represent sharp edges and noise of the plain image while the low frequency component corresponds to the smooth area [1]. In our proposed encryption scheme, we concentrate on spatial domain, frequency domain is out of the scope of our work. Classical and modern ciphers algorithms have all been developed for the simplest form of multimedia data. Chaos-based image encryption scheme is one of the encryption algorithms that works on spatial domain and that have suggested an efficient way to deal with fast and highly secure image encryption [18]. Encrypting images, using chaos-based image encryption scheme, considers the image as 2-dimensional array of pixels [5].

III. Proposed Image Encryption Scheme

The proposed image encryption algorithm in this research work is based on chaos-based image encryption scheme that has been proposed first by Fridrich [7], [8]. The scheme of the chaos-based image encryption is depicted in Fig. 1, as it is shown in Fig. 1 the plain image goes through two processes, namely, confusion process and diffusion process. At each stage there is a key input which is a single value starts with a seed value (easy to be predicted by cryptanalyst) and from this value the confusion key and key diffusion are created. In contrast, our proposed scheme follows Fridrich's scheme except the chosen key, in our proposed scheme the key is considered as another image (or on another say, a matrix of random values) with the same dimension (or bigger) the plain image.

The operations needed in encryption processes are confusion and diffusions as in. Both confusion and diffusion processes will be executed only one time for the purpose of reducing the computational time with the grantee of high level of security. Fig. 2 shows the general framework diagram of the proposed chaotic encryption algorithm. Stage (1), the plain-image and key image are gone through preparation process. The preparation process is responsible for analyse the chosen plain image and key image before passing them to the confusion process. The first check will be the size of plain image and key image, the key image should be equal or greater than the plain image, otherwise the process will fail. The second check will look for the plain image and key image, if they are in 8-bit grayscale the process will run and will deal with the images directly, while in 24-bits RGB images a new mechanism to extract the images (plain and key) RGB channels should be implemented to deal with each channel separately.

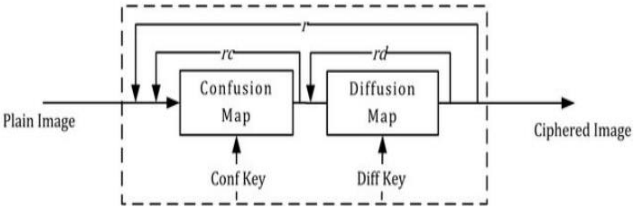


Fig. 1. Fridrich Image Encryption Scheme [8].

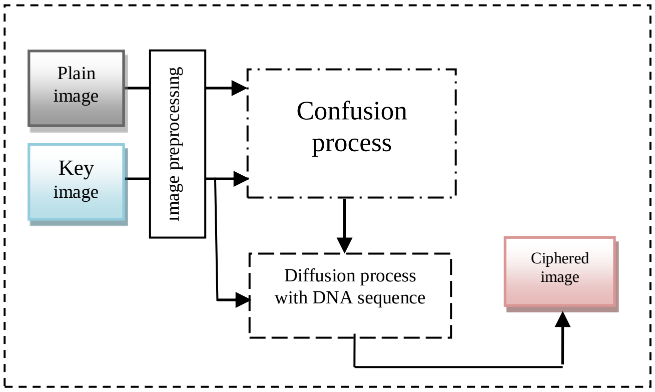


Fig. 2. Proposed Framework of Image Encryption Scheme.

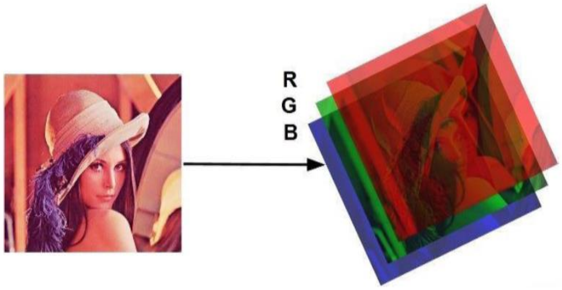


Fig. 3. 24-bit RGB Image Channels Extractions.

The extraction of the RGB channels of the plain and key images will store each image channel in a single 8-bits matrix with dimension equal to the dimension of the original image. Fig. 3 illustrate the extraction process that produces three channels image for plain and key image and each of these channels is 8-bit image. In the next subsections each process will be explained in detail.

A. Confusion Process

Confusion process goes through two processes, the RGB channels extraction (of both the plain image and the key image) and the confusion function that will be applied to each corresponding pixel in plain and key image as shown in Fig. 4. In confusion process, firstly, the chosen key image should be larger or of the same size of the plain image. Secondly, the confusion function which can be defined as any arithmetic, geometric, or bitwise function that will be performed on the corresponding pixels in RGB channels of the plain and key image as shown in Fig. 4. The chosen function should be strictly monotonically function (reversible function) [19], i.e., the encrypted pixel can be recovered back in the decryption process. For clarity, let RGB channels of the plain image represented by P_R, P_G, and P_B respectively. And for the key image, the RGB channels represented by K_R, K_G, and K_B respectively. For a plain image and key image with $n\,x\,m$ dimension, there are $n\,x\,m$ pixels, for simplicity we assume $n = m$, therefore number of pixels are n^2. In the confusion process, as explained previously, the corresponding pixels in plain and key images are operated by the confusion function (f) and result (an intermediate) new RGB channels with the same dimension of the original plain image.

Let us name these new channels as C_R, C_G, and C_B. The equations 2, 3, and 4 illustrate the operations for getting these channels (ciphered RGB), such that i and j are the indexes of the pixels and it ranges form $0 < i < n$ and $0 < j < n$. The function (*f*) is considered a bitwise XOR binary function in this study.

$$C_R(i,j) = f(P_R(i,j), K_R(i,j)) \quad (2)$$

$$C_G(i,j) = f(P_G(i,j), K_G(i,j)) \quad (3)$$

$$C_B(i,j) = f(P_B(i,j), K_B(i,j)) \quad (4)$$

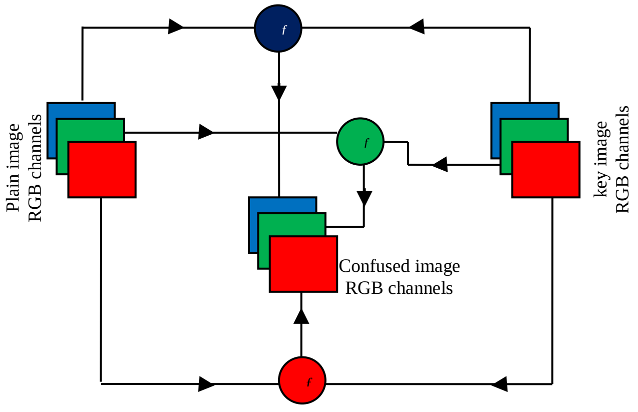


Fig. 4. Confusion Process for 24-Bit RGB Channels.

Confusion process is responsible for reducing the correlation between adjacent pixels by dissolving the pixels [9]. The method for dissolve this correlation is by applying confusion function of pixels in plain image and key image [3]. The output of this process is the RGB channels that will be input into the diffusion process.

B. Diffusion Process with DNA Sequence

Diffusion process starts at the end of confusion process and the output of confusion process will be the input of diffusion process. In this operation the output of confusion process will be encoded using DNA sequence computing as it will be explained in this section. DNA computing becomes important in many fields of research in the last four decades, in this research paper we employ DNA computing for the diffusion process. DNA sequence can be encoded and stored as a binary code of four pairs of two bits. The four chemical bases Adenine, Cytosine, Guanine, and Thymine are pair up with each other, A with T and C with G, each base-pair are the complement of each other. If we assume the binary coding of A is 00 therefore the binary coding of T is 11, and if the binary coding of C is 01 therefore the binary coding of G is 10. Therefore, for each pixel in the RGB channels (which come from confusion process) a new coding can be applied, for example, the pixel with binary 00101101 can be encoded using DNA sequence as AGTC as done in some related works [17], [20], [21]. The key image also will be encoded using the DNA sequence. The RGB channels which are the output of the confusion process will be operated with the RGB channels of the original key image using the logical addition and subtraction of the chemical bases as shown in Table I. The logical addition in this research work is employed for encryption process, the reverse operation which is the subtraction operation is employed in the decryption operation with the rules shown in in Table II.

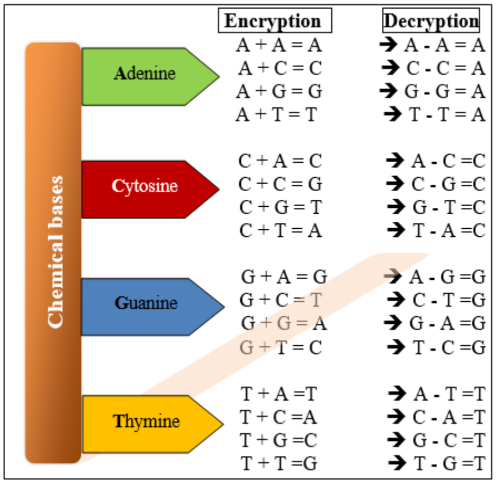
As shown in Table II below the subtraction operation is reverse function of the addition function, therefore using one logical operation for encryption process enforces us to use the other operation for decryption.

TABLE I. DNA LOGICAL ADDITION OPERATION

+	A	C	G	T
A	A	C	G	T
C	C	G	T	A
G	G	T	A	C
T	T	A	C	G

TABLE II. DNA LOGICAL SUBTRACTION OPERATION

-	A	C	G	T
A	A	T	G	C
C	C	A	T	G
G	G	C	A	T
T	T	G	C	A



Chemical bases	Encryption	Decryption
Adenine	A + A = A A + C = C A + G = G A + T = T	➔ A - A = A ➔ C - C = A ➔ G - G = A ➔ T - T = A
Cytosine	C + A = C C + C = G C + G = T C + T = A	➔ A - C =C ➔ C - G =C ➔ G - T =C ➔ T - A =C
Guanine	G + A = G G + C = T G + G = A G + T = C	➔ A - G =G ➔ C - T =G ➔ G - A =G ➔ T - C =G
Thymine	T + A =T T + C =A T + G =C T + T =G	➔ A - T =T ➔ C - A =T ➔ G - C =T ➔ T - G =T

Fig. 5. The Encryption and Decryption of the Chemical bases **A**denine, **C**ytosine, **G**uanine, and **T**hymine using Logical Addtion and Subtraction Operations.

he four chemical bases for DNA sequence that are using in the diffusion are operated using logical addition for encryption and logical subtraction for decryption. Fig. 5 depicts the 16 possible scenarios for additions and the 16 possible scenarios for subtractions [15].

Finally, the Encryption and Decryption techniques are reverse to each other, therefore the last operation in encryption will be the first operation in decryption.

IV. Result Analysis

In this section, the detailed results to verify the efficiency, robustness, and high level of security of the proposed encryption scheme are outlined. A standard Lena image with 256 x 256 gray scale and a key image of the same size was used to verify the proposed methods performance as shown in Fig. 6. The experiment has been performed using visual studio package and implemented in windows environment. The first step the key image in Fig. 6(b) is prepared using the same proposed encryption scheme, second step2, the prepared key image together with the plain image (Lena image) in Fig. 6(a) are input to the encryption process, the result ciphered image is shown in Fig. 6(c). The Lena image is loaded to the encryption software, the key image is prepared using the same encryption algorithm to add more noise to the key image (any image can be used as a key without encrypt it) and to improve the encryption of the original image and the pixels distributions of the encrypted image that returns noisy images as shown in Fig. 6(c). These aspects indicate high level of security against the attacks.

The second experiment has been performed for Firise's image (293x203) as shown in Fig. 7. The chosen encrypted key image is in the same dimension (293x203) as the plain image. The same experiment with this image is performed in the previous studies with the contradiction that in our study we have less computation time and high performance. The same process for encrypting Lena image is applied for Firise image, whereas Fig. 7(a) represent the plain image, Fig. 7(b) represents the encrypted key-image, and Fig. 7(c) represents the encrypted (ciphered) image.

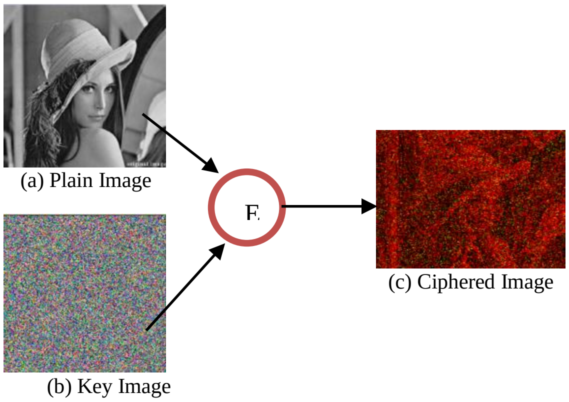


Fig. 6. Encryption Process for Lena Image (225x227), (a) Plian Image, (b) key Image, (c) the Encrypted Image.

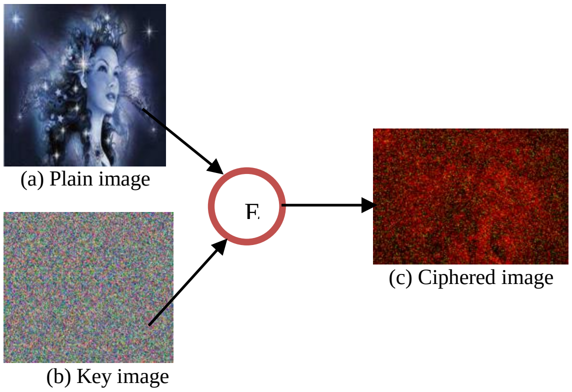


Fig. 7. Encryption Process for Firise Image (293x203), (a) Plian Image, (b) Key Image, (c) the Encrypteds Image.

The same methodology for image encryption using chaos-based approach has been conducted in our previous work [3]. Fig. 8 shows some of the experiments for encryption and decryption process for Lena image with 225x227 gray scale. The encryption process shown in Fig. 8 follows the same approach for conducting the proposed encryption scheme.

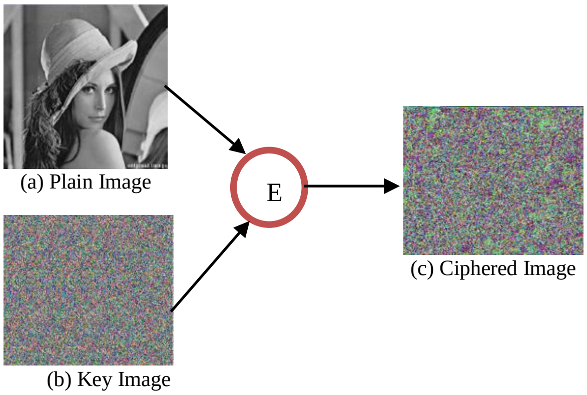
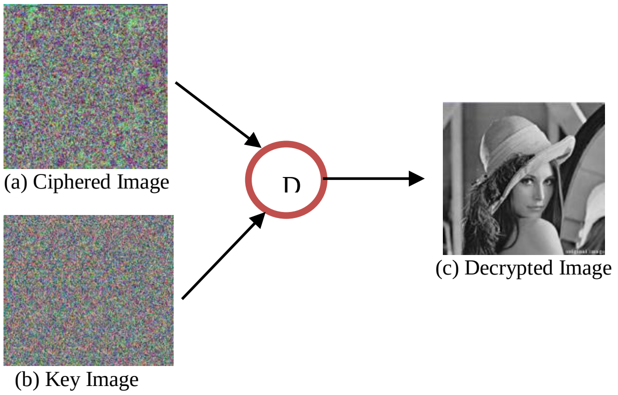


Fig. 8. Chaos-based Image Encryption Scheme[3], (a) Plian Image, (b) Key Image, (c) the Encrypted Image.



(a) Ciphered Image

(b) Key Image

(c) Decrypted Image

Fig. 9. Chaos-based Image Decryption Scheme [3], Lena Image (225x227), (a) Ciphered Image, (b) Key Image, (c) the Decrypted Image.

The reverse operation for image encryption is the decryption process. Fig. 9 shows the decrypted Lena image using the correct key image. Choosing wrong key image will not produce the correct plain image. For the key sensitivity and security analysis, next section investigates these issues.

V. Security Analysis

The proposed chaos-based image encryption scheme with DNA sequence coding has satisfied high-level of security, robustness, and efficiency, these features due to the high domain space of the key as well as the chaos features [22]. In this research work as well as our previous work, in [3], we use a key as an image for the encryption/decryption process in order to prevent the encrypted image against brute-force attack, statistical attack, known-plaintext attack and select plain text attack, these attacks are well known attacks listed in [3]. In the proposed cryptosystem, a confusion process is run, and the output goes to the diffusion process. If we assume we have $M \times N$-sized plain image, for simplicity put $M = N$ therefore the plain image will be N^2-sized image. For the encryption and based on the methodology of our encryption scheme the same key image dimension should be utilized i.e., we will need N^2-sized key image. Since the encryption process is working in bitwise therefore for gray scale image with 8 bits in each pixel, we will have $8*N^2$-sized key image with binary values 0 and 1. The possible combinations of a matrixes with $8*N^2$ binary values are in the domain 2^{8*N^2} , which increase exponentially as N increases. For the RGB channels the key space became ultra, since we will have 3 matrixes for R channel, G channel, and B channel, respectively and the key space become in $2^{8*N^2} * 2^{8*N^2} * 2^{8*N^2} = 2^{24*N^2}$. For example, a plain image of size 225x227 gray scale and key image with the same size, the brute-force attack needs to search the key in $2^{8*225*227} ==> 2^{51525}$ which is a very (ultra) large key space domain. This is the key point of our proposed algorithm, impossible for the key image to be predicated.

VI. Conclusion

The proposed study shows that the achieving of ideal secrecy system in cryptography can be performed with existence of confusion and diffusion processes. Despite the differential analysis is very interested attack but the proposed framework resolves this issue by the proposing of new technique for choosing key method. In general the proposed encryption framework consists of two processes (confusion and diffusion with DNA sequence) and one round of the two operations is grantee to produce high level of image security instead of having (n) rounds for confusion and (m) round in diffusion as in the conventional image encryption frameworks [7], [17], [23], [24]. This feature will reduce the computational time and the system complexity. The proposed confusion process (using confusion function) is responsible on the breach of high correlations of the adjacent pixels, while the diffusion process with DNA sequence logical operations increases withstanding against attack. Comparing to the several studies in literature our proposed algorithm's distinguished in terms of choosing the key as an image (instead of having single value) that contribute to high level of resistance of all kinds of attacks due to ultra-key space domain which is in the range of 2^{8*N^2} in the gray scale domain and 2^{24*N^2} in the RGB channels.

Acknowledgment

This project was funded by the Deanship of Scientific Research (DSR) under grant no. G- 526-611-1431. The authors, therefore, acknowledge with thanks to DSR's technical and financial support.

References

[1] H. Borrebach, "IMAGE ENCRYPTION FRAMEWORK BASED ON MULTI-CHAOTIC MAPS AND EQUAL PIXEL VALUES QUANTIZATION," no. Fb 14, p. 2018, 2018.

[2] M. Salleh, S. Ibrahim, and I. F. Isnin, "IMAGE ENCRYPTION ALGORITHM BASED ON CHAOTIC MAPPING The requirements of information security within an organization have undergone tremendous changes . Before the widespread use of data processing equi pment , the security of sensitive documents depends o," Image (Rochester, N.Y.), vol. 39, no. D, pp. 1–12, 2003.

[3] N. A. Al-Romema, A. S. Mashat, and I. AlBidewi, "New Chaos-Based Image Encryption Scheme for RGB Components of Color Image," Comput. Sci. Eng., 2012, doi: 10.5923/j.computer.20120205.06.

[4] M. SaberiKamarposhti, I. AlBedawi, and D. Mohamad, "A new hybrid method for image encryption using DNA sequence and chaotic logistic map," Aust. J. Basic Appl. Sci., vol. 6, no. 3, pp. 371–380, 2012.

[5] J. B. . Choudhary, R., &Arun, "Secure Image Transmission and Evaluation of Image Encryption.," Int. J. Innov. Sci. Eng. Technol., 2014.

[6] "Jain, A, Pixel chaotic shuffling and Arnold map based Image Security Using Complex Wavelet Transform.," J. Netw. Commun. Emerg. Technol., vol. 6, no. 5, pp. 8–11, 2016.

[7] J. Fridrich, "Image encryption based on chaotic maps," Proc. IEEE Int. Conf. Syst. Man Cybern., vol. 2, pp. 1105–1110, 1997, doi: 10.1109/icsmc.1997.638097.

[8] M. Farajallah, S. El Assad, and O. Deforges, "Fast and Secure Chaos-Based Cryptosystem for Images," Int. J. Bifurc. Chaos, vol. 26, no. 2, 2016, doi: 10.1142/S0218127416500218.

[9] V. R. Divya, V. V., Sudha, S. K., &Resmy, "Simple and Secure Image Encryption.," Int. J. Comput. Sci. Issues., vol. 9, no. 3, pp. 286–289, 2012.

[10] H. Choi, J., Seok, S., Seo, H., & Kim, "A Fast ARX Model-based Image Encryption Scheme.," Multimed. Tools Appl., vol. 75, no. 22, pp. 14685–14706.

[11] S. Bashir, Z., Rashid, T., & Zafar, "Hyperchaotic Dynamical System based Image Encryption Scheme with Time-Varying Delays.," Pacific Sci. Rev. A Nat. Sci. Eng., 2016.

[12] S. A. (2016) Kulsoom, A., Xiao, D., & Abbas, "An Efficient and Noise Resistive Selective Image Encryption Scheme for Gray Images based on

Chaotic Maps and DNA Complementary Rules," Multimed. Tools Appl., vol. 75, no. 1, pp. 1–23, 2016.

[13] N. Nasser, M. Anan, M. F. C. Awad, H. Bin-Abbas, and L. Karim, "An expert crowd monitoring and management framework for Hajj," 2017, doi: 10.1109/WINCOM.2017.8238202.

[14] Z. Gu, G., Ling, J., Xie, G., & Li, "A Chaotic-cipher-based Packet Body Encryption Algorithm for JPEG2000 Images.," Signal Process. Image Commun., vol. 40, pp. 52–64, 2016.

[15] M. Enayatifar, R., Abdullah, A. H., Isnin, I. F., Altameem, A., & Lee, "Image Encryption using a Synchronous Permutation-Diffusion Technique," Opt. Lasers Eng., vol. 90, pp. 146–154, 2017.

[16] S. Nagaraj, G. S. V. P. Raju, and K. Koteswara Rao, "Image encryption using elliptic curve cryptograhy and matrix," in Procedia Computer Science, 2015, vol. 48, no. C, doi: 10.1016/j.procs.2015.04.182.

[17] S. Xu, Y. Wang, J. Wang, and Y. Guo, "A fast image encryption scheme based on a nonlinear chaotic map," ICSPS 2010 - Proc. 2010 2nd Int. Conf. Signal Process. Syst., vol. 2, pp. 1–16, 2010, doi: 10.1109/ICSPS.2010.5555472.

[18] H. Gao, Y. Zhang, S. Liang, and D. Li, "A new chaotic algorithm for image encryption," Chaos, Solitons and Fractals, vol. 29, no. 2, pp. 393–399, 2006, doi: 10.1016/j.chaos.2005.08.110.

[19] E. Seeram, "Digital image processing.," Radiol. Technol., vol. 75, no. 6, 2004, doi: 10.4324/9781315693125-12.

[20] M. SaberiKamarposhti, I. AlBedawi, and D. Mohamad, "A new algorithm for image encryption using DNA sequence and cycling chaos," Aust. J. Basic Appl. Sci., vol. 6, no. 3, pp. 381–392, 2012.

[21] K. Singh and K. Kaur, "Image Encryption using Chaotic Maps and DNA Addition Operation and Noise Effects on it," Int. J. Comput. Appl., vol. 23, no. 6, pp. 17–24, 2011, doi: 10.5120/2892-3779.

[22] S. Lian, J. Sun, and Z. Wang, "Security analysis of a chaos-based image encryption algorithm," Phys. A Stat. Mech. its Appl., vol. 351, no. 2–4, 2005, doi: 10.1016/j.physa.2005.01.001.

[23] X. Liu and A. M. Eskicioglu, "Selective encryption of multimedia content in distribution networks: Challenges and new directions," Proc. Second IASTED Int. Conf. Commun. Internet, Inf. Technol., pp. 527–533, 2003.

[24] L. Kocarev, "Chaos-based cryptography: A brief overview," IEEE Circuits Syst. Mag., vol. 1, no. 3, pp. 6–21, 2001, doi: 10.1109/7384.963463.

Integrated Model to Develop Grammar Checker for Afaan Oromo using Morphological Analysis: A Rule-based Approach

Jemal Abate[1]
Department of Information Science
Haramaya University, Haramaya, Ethiopia

Vijayshri Khedkar[2]*, Sonali Kothari Tidke[3]
Symbiosis Institute of Technology, Symbiosis International
(Deemed University), Pune, Maharashtra, India

***Abstract*—This study has implemented a rule-based approach on grammar checkers by integrating a spell-checker with a morphological analyzer to improve the Afaan Oromo grammar checker. A corpus containing about 300,000 words has been prepared to be used for spell-checker. About 300 grammar rules are constructed to detect the grammar error within the Afaan Oromo text and to suggest the possible grammar correction. The developed frameworks have experimented on the document having pairs of 100 correct and incorrect sentences. The experimental result for checking the spelling errors has scored 73% of recall, 76% precision, and 75% of F-measure. The score for suggesting the correct spelling is 78% of recall, 62% precision, and 70% precision F-measure while the evaluation result for detecting the grammar errors has 47% recall, 90% precision and 68% f-measure score. For suggesting the possible correct grammar on the detected error, the system has scored 61% recall, 71% precision and 66% f-measure. The overall performance of the developed system has a good performance. However, there is still a need to conduct further research to improve the Afaan Oromo grammar checker.**

***Keywords*—Grammar checker; spell checker; part-of-speech tag; error detection; syntactic analysis; semantic analysis; morphological analyzer; NLP**

I. Introduction

For communication, a natural language is a language used by humans. Natural language processing (NLP) is a field of study mainly concerned with the communications between human languages and the computer. During a written communication, people may make an error which can be a spelling error and/or a grammar error. A spelling error is an error if the given the word does not exist in the language's vocabulary, whereas a grammatical error occurs when the written sentence is not as per the grammar rule of a given language [1].

With globalization, many individuals work on National and International languages for everyday communications. But it is difficult for many to write various contents using correct spelling and grammar, whether professional communication or a regular discussion, which makes grammar checking one important tool in the word processing software. Grammar checking has significance for having a good flow of ideas, exactness and quality of the expressed written content [1]. Grammar checking is one of the activities in the natural language's written communication and NLP application. Grammar checking checks the grammatical errors in text and suggests possible corrections in many cases. This is one of the most frequently used tools in language engineering.

One of the main challenges in the application of grammar checking for natural language is it is language-specific. Due to this, the grammar checker that works for one language may not work for another language. Therefore, it's necessary to design a grammar checker for each language as per the grammatical rule of a particular language.

Afaan Oromo is one of the most widely spoken languages in east Africa, which accounts for about 40 million speakers in Ethiopia; it is also spoken in neighboring Ethiopian countries like Kenya, Somalia, Djibouti and Egypt [2].

Today, Afaan Oromo is an official working language of Oromia national, regional, and educational languages for primary school (1-8) students in the Oromia region. As a result, many works for official and/or personal purposes use Afaan Oromo for interpersonal communication. Most of the time, while preparing various reports, peoples are vulnerable to make a grammatically wrong statement unintentionally or unknowingly. Some languages, such as English, French, etc., have a tool to catch and make corrections on these problems. However, Afaan Oromo is one of the under-resourced languages. Due to this, there are no tools that can assist the users in catching and making corrections on the grammatically wrong statements. Therefore, it is required to develop and implement the Afaan Oromo grammar checker to eliminate these problems. Developing a grammar checker for Afaan Oromo is significant for non-writers and nontechnical peoples to have accurate and quality written content. This study is intended to design and develop an integrated model to develop a grammar checker for Afaan Oromo using morphological analysis – A Rule-Based Approach.

II. Grammar Checker Approaches

There are three widely used approaches by various researchers like Syntax-based approach, statistical-based approach and rule-based checking [3].

- Syntax-based approach: In this approach, text parsing is used by assigning a tree structure to the sentence. If the text parsing mismatches, the statement is labeled as a grammatically wrong sentence; otherwise, it is considered correct.

*Corresponding Author

- Statistical-based approach: In this approach, a corpus annotated with its part of the speech tag is prepared and used to build a sequence of parts of the speech tag list. If the frequency of the tagged text is less than with the already trained model, this text is considered incorrect otherwise it's it is considered correct.

- A rule-based approach: In this approach, a set of grammar rules has been manually prepared to match against a text. If the text matches the crafted rules, it is considered a correct otherwise wrong sentence.

III. Related Works

Various studies have been conducted on natural language processing for the Afaan Oromo language. For instance, [4] have designed a spell checker for non-word Afaan Oromo spell checker. This work is focused on checking misspelled words and use the proposed algorithm to observe if the right word is generated under suggestion. But authors are not considering the context of the statement in their work.

The author in [5] attempted to design a grammar checker for Afaan Oromo using a rule-based approach. The author constructed about 123 rules for the detection of the Afaan Oromo grammar error. The rules are constructed depending on the language rule of affixation. However, false alarms in the developed framework lead the system to become challenged to catch the grammar error correctly. The author in [5] has mentioned some of the reasons for the problem that occurred. These are some words root and affix are identified wrongly by the stammer, a wrongly assigned part-of-speech tag and incompleteness of the rules. However, there are many reasons for the problems with [5] framework in addition to the reasons mentioned. Rules ranging from 81-86 were constructed to catch the error of past perfect tense. But this system may falsely alarm the correct grammar as wrong due to incorrect rules. Let's have a look at the following example, which tests his rule. One of the rules which are crafted by [5] says, "If the subject of the sentence is third-person singular masculine, so the verb must end with the suffix -ee, and the sentence must end with ture.”

Example: Callisee Bira dabruu yaalee ture. (He tried to pass in silence).

But the above rule may flag other correct statements as a grammatically wrong sentence. Because, in Afaan Oromo, if the subject of the sentence is third-person singular masculine, but the verb may end with the suffix other than -ee and also it's not necessary for the sentence to end with ‘ture’.”

Example: Callisee Bira dabruu yaalus hin milkoofne. (He tried to pass in silence, but he did not).

The author in [6] has developed the Afaan Oromo grammar checker by using a statistical approach. The author has implemented two statistical approaches: the token n-gram frequency whose probability of sequence is calculated and the tag n-gram frequency whose POS tag is assigned to all tokens, and the probability of the tag sequence is calculated for training and testing the checker. However, [6] didn't incorporate the checking of spelling errors in his work which is fundamental to the work of grammar checkers. The corpus used for training contains only 10000 words due to this POS tagger, and the Morphological analyzer has become inaccurate. Additionally, the article didn't highlight constructing a rule that will be used for training the model and improving the model performance. As a result of the factors mentioned, the work proposed in the paper has to be improved by incorporating spell checking along with a good morphological analyzer.

Even if few works have been done on grammar checkers for the Afaan Oromo language, there is still a gap in developing a good grammar corrector that can help to fix common spelling errors, instances of the passive voice, clunky and hard to understand language.

IV. Material and Methods

A. Material Used

Various development tools have been used to design the Afaan Oromo grammar checker. Python is one of the tools that were used to develop the prototype. Python is used because it's easy to work with, learn and adaptable scripting language, making it attractive for development [7]. HORN MORPHO 2.5 is used for POS tagging activity. It is a program that analyzes Amharic, AO, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the word's grammatical structure [8]. The proposed research has also used the Hunspell dictionary to store the required grammar correction rules and integrate the developed framework into Apache OpenOffice.

B. Design Approach

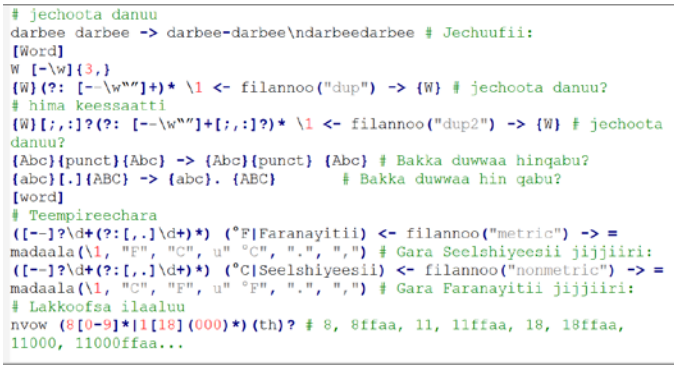
There are quite a lot of ways in which diverse approaches to grammar checking can be illustrated. This study intended to develop a grammar checker for Afaan Oromo using a rule-based approach by integrating with spell-checker and morphological analyzer. The dictionary containing about 300,000 unique words is constructed that can help to detect spelling errors. For grammar checking, a set of grammar rules has been constructed, which accounts for about 250 rules that check the errors in the grammar.

To perform the grammar checking, the developed system will find a word from the sentence. Then the tag of the word is checked against the developed grammar rules to detect various errors such as style problems, agreement in gender, number, word order, etc. If the sentence contains an error, the suggestion to the error will be displayed. The designed framework has two components, namely the spell-checker component and grammar checker component.

1) Spell-checker components: The spell-checker components have a set of words and affix rules of Afaan Oromo languages. This component helps to catch the spelling errors from the texts and suggests the possible correction to the wrong words depending on the morphological rules of the Afaan Oromo languages.

2) Grammar checker components: This module consists of a set of grammatical rules of the Afaan Oromo languages.

This component works by checking the arrangements of words in a given sentence. The crafted rules check the syntactic agreement features in the statement, i.e., the agreement in the gender (Masculine or Feminine), number (Singular or Plural), a person (1st/2nd/3rd person singular, 1st/2nd/3rd person plural) and proofreading (punctuation). In Fig. 1, some of the sample rules are constructed.



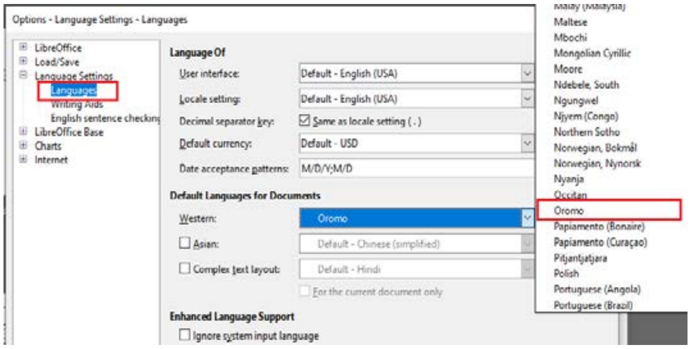
```
# jechoota danuu
darbee darbee -> darbee-darbee\ndarbeedarbee # Jechuufii:
[Word]
W [-\w]{3,}
{W}(?: [--\w"”]+)* \1 <- filannoo("dup") -> {W} # jechoota danuu?
# hima keessaatti
{W}[;,:]?(?: [--\w"”]+[;,:]?)* \1 <- filannoo("dup2") -> {W} # jechoota
danuu?
{Abc}{punct}{Abc} -> {Abc}{punct} {Abc} # Bakka duwwaa hinqabu?
{abc}[.]{ABC} -> {abc}. {ABC}        # Bakka duwwaa hin qabu?
[word]
# Teempireechara
([--]?\d+(?:[,.]\d+)*) (°F|Faranayitii) <- filannoo("metric") -> =
madaala(\1, "F", "C", u" °C", ".", ",") # Gara Seelshiyeesii jijjiiri:
([--]?\d+(?:[,.]\d+)*) (°C|Seelshiyeesii) <- filannoo("nonmetric") -> =
madaala(\1, "C", "F", u" °F", ".", ",") # Gara Faranayitii jijjiiri:
# Lakkoofsa ilaaluu
nvow (8[0-9]*|1[18](000)*)(th)? # 8, 8ffaa, 11, 11ffaa, 18, 18ffaa,
11000, 11000ffaa...
```

Fig. 1. Sample Rules.

V. Implementation and Evaluation

1) Implementation: The developed framework has been implemented using a Libreoffice text/document writer. LibreOffice is one of the document writer software supporting the integration of the Afaan Oromo spelling and grammar checker. Fig. 2 shows a list of extensions already in the Libreoffice software before installing the Afaan Oromo spelling and grammar checker.

As shown in Fig. 2, installing the developed framework as an extension of writing aids, there is no indicator of the Afaan Oromo language. However, after installing the Afaan Oromo spelling and grammar checker, the writer started to indicate an icon showing the language extension for grammar and spell-checking is installed. As shown in Fig. 3, the layout for the language extension is highlighted with a red color.



Language Of

User interface: Default - English (USA)

Locale setting: Default - English (USA)

Decimal separator key: ☑ Same as locale setting (.)

Default currency: Default - USD

Date acceptance patterns: M/D/Y;M/D

Default Languages for Documents

Western: Oromo

☐ Asian: Default - Chinese (simplified)

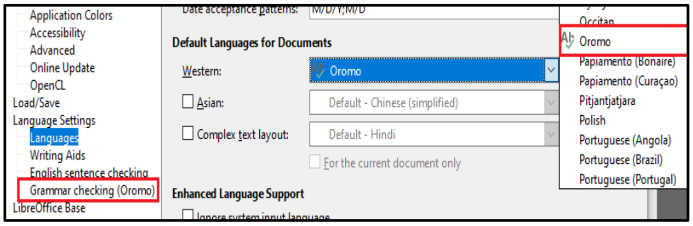
☐ Complex text layout: Default - Hindi

☐ For the current document only

Enhanced Language Support

☐ Ignore system input language

Fig. 2. Layout before Installing the Afaan Oromo Spelling and Grammar Checker.



Date acceptance patterns: M/D/Y;M/D

Default Languages for Documents

Western: Oromo

☐ Asian: Default - Chinese (simplified)

☐ Complex text layout: Default - Hindi

☐ For the current document only

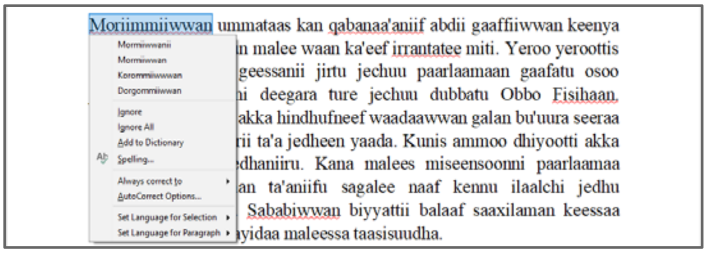
Enhanced Language Support

☐ Ignore system input language

Fig. 3. Layout after Installing the Afaan Oromo Spelling and Grammar Checker.

The developed framework has been implemented to perform spelling checking, correction, grammar error detection and suggestion.

a) Spell-Checker Implementation: At this phase, the system is implemented to detect the wrongly spelled words and suggest possible corrections. Fig. 4 shows the implementation of the spell-checker.



Moriimmiiwwan ummataas kan qabanaa'aniif abdii gaaffiiwwan keenya
n malee waan ka'eef irrantatee miti. Yeroo yeroottis
geessanii jirtu jechuu paarlaamaan gaafatu osoo
hi deegara ture jechuu dubbatu Obbo Fisihaan.
akka hindhufneef waadaawwan galan bu'uura seeraa
rii ta'a jedheen yaada. Kunis ammoo dhiyootti akka
dhaniiru. Kana malees miseensoonni paarlaamaa
an ta'aniifu sagalee naaf kennu ilaalchi jedhu
Sababiwwan biyyattii balaaf saaxilaman keessaa
yidaa maleessa taasisuudha.

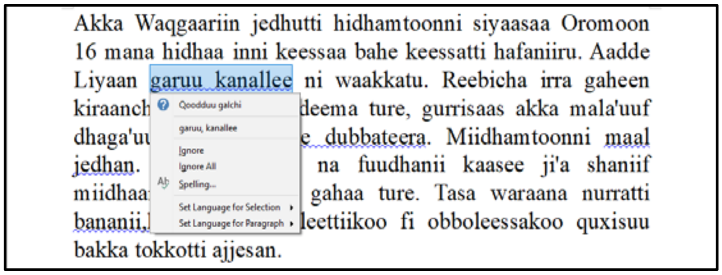
Fig. 4. Spell-Checker Implementation Sample Screenshot.

To indicate the wrongly spelled word in a sentence, all wrongly spelled words are underlined with a red flag. But if the word has no spelling error, no sign is displayed. As shown in Fig. 4, the wrongly written words are flagged with a red color underscore. When these wrongly written words are right-clicked, the possible suggestion is shown.

b) Grammar-Checker: This implementation was done to determine the wrongly written statements and extract the grammatically wrong statements within the text by the system. The prepared corpus (doc1 and doc2) is used alternatively for error detection and false alarm identification in this implementation. Fig. 5 is the screenshot of the implementation of detection and correction of the grammar of the proposed framework.

Fig. 5 shows how the system detects the grammar error and suggests the possible correction.

The statement that has a grammar error will be flagged with a blue underscore to make the user alerted and take proper action. The grammar suggestion will be shown when right-clicked on the flagged text. However, those statements written with a correct grammar rule will not be flagged.



Akka Waqgaariin jedhutti hidhamtoonni siyaasaa Oromoon
16 mana hidhaa inni keessaa bahe keessatti hafaniiru. Aadde
Liyaan garuu kanallee ni waakkatu. Reebicha irra gaheen
kiraanch deema ture, gurrisaas akka mala'uuf
dhaga'uu e dubbateera. Miidhamtoonni maal
jedhan. na fuudhanii kaasee ji'a shaniif
miidhaa gahaa ture. Tasa waraana nurratti
bananii, leettiikoo fi obboleessakoo quxisuu
bakka tokkotti ajjesan.

Fig. 5. Detection and Correction of Grammar Errors.

2) Evaluation: The developed system is evaluated to check whether the objectives of the research are achieved or not. To evaluate the performance of the systems, two documents named doc1 and doc2 are prepared. These documents are used for testing and evaluation purposes, where each document contains about 500 sentences. All of the words and sentences within the first document (doc1) are grammatically correct, whereas the second document (doc2) contains the possible wrong spelling and grammar of doc1. The evaluation task is done in two phases, and the first phase evaluation is applied to test the detection and suggestion capability of the system for the spelling error within a sentence. The second phase is evaluated to check the system's performance for grammar error detection and correction suggestions made to the identified error.

a) Evaluation of Phase 1: The performance of the system for spelling checking has been evaluated at this phase. The document with 50 words has been prepared, and of these, 25 words were wrongly spelled while the rest 25 correctly spelled. The system is expected to suggest at least three correctly spelled words related to the detected spelling error.

b) Evaluation of Phase 2

- Grammar Error Detection: The framework is tested against doc1 to identify the grammatically wrong statements within the text, whereas doc2 is used to handle false alarms.

- Correction Suggestions: The correction suggestions evaluation has been done to evaluate the system's performance for suggesting the possible correct grammar of the detected errors. The system is tested by using the document containing the wrong grammar statements (doc1). Depending on the error detected, the number of correction suggestions may vary. However, the system is evaluated to suggest a minimum of two correct suggestions per error.

3) Confusion matrix: To demonstrate the performance of the proposed system, a confusion matrix has been used. The confusion matrix is a commonly used technique to describe the classifier's performance based on test data [9]. Table I shown below, describes the confusion matrix result for the system performance.

The performance measure shows the detailed description for the activities of the systems such as; checking the spelling error, suggesting the correction for the spelling error, grammar checking and correction are shown in the Fig. 6, 7, 8 and 9.

Evaluation matrix performance can also be explained using f-measure, precision and recall values. Following are the formulas used for these calculations.

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

Table II is focused on the results of the evaluations.

TABLE I. PERFORMANCE MEASURE DESCRIPTION

S. No.	Evaluation Parameters	TP	FP	FN	TN
1	Have a Spelling Error	19	5	7	6
2	Spell-Checker Correction Suggestion	56	15	16	35
3	Have a Grammar Error	86	4	96	10
4	Predicted Correction Suggestion	122	28	78	50

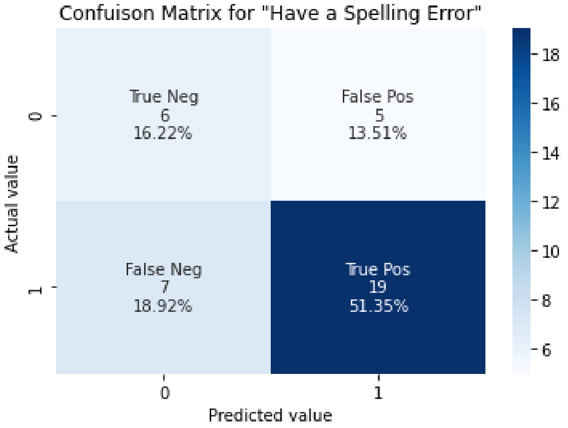


Fig. 6. Have a Spelling Error.

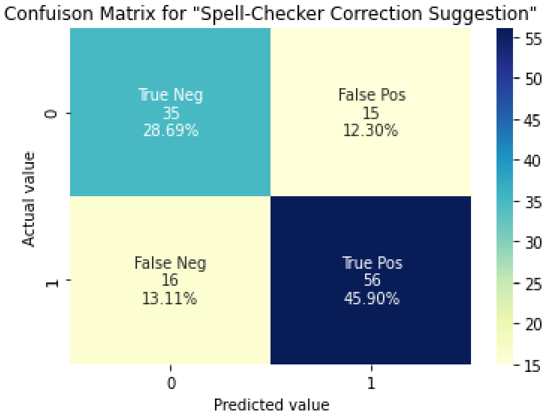


Fig. 7. Spell-Checker Correction Suggestion.

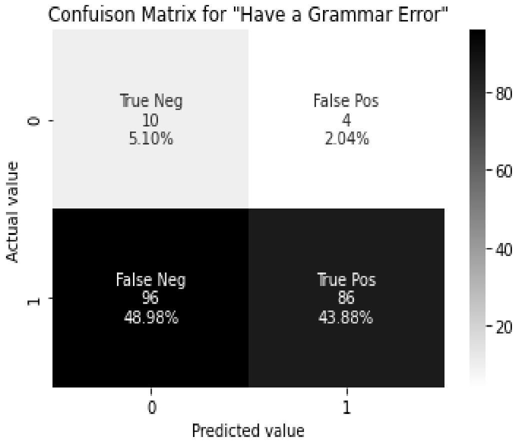


Fig. 8. Have a Grammar Error.

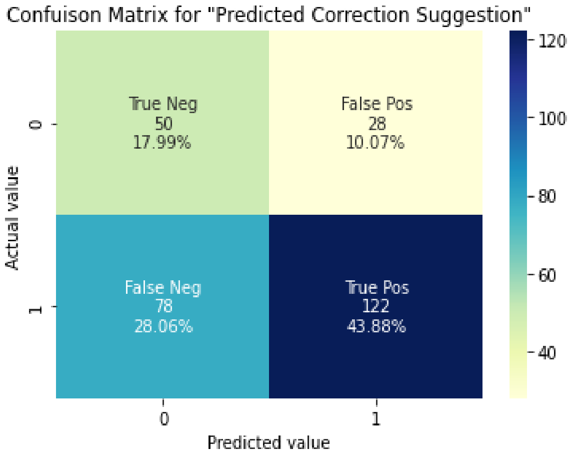


Fig. 9. Predicted Correction Suggestion.

TABLE II. PERFORMANCE EVALUATION MATRIX

S. No.	Parameters Used	Evaluation Result		
		Recall	Precision	F-Measure
1	Checking Spelling Errors	0.73	0.76	0.75
2	Suggesting Correction	0.78	0.62	0.70
3	Checking Grammar Errors	0.47	0.90	0.68
4	Suggesting Correct Grammar	0.61	0.71	0.66

VI. DISCUSSION

The developed system has been evaluated on four basic tasks: checking spelling errors within the text, suggesting the correct spelling for the detected error, checking grammar errors, and suggesting the correct grammar for the identified grammar errors. Accordingly, the confusion matrix result for checking the spelling error has scored 51.35% TP, 13.51% FP, 18.92% FN, and 16.22% TN, whereas spell-checker correction suggestion 45.90% TP, 12.30% FP, 13.11% FN and 28.69% TN score. But, for checking the grammar error and suggesting the possible correct grammar, the system has scored 43.88% TP, 2.04% FP, 48.98% FN and 5.10% TN result of confusion matrix measurement.

The performance evaluation result of the system, which is evaluated by the confusion matrix measurement, is described by using the f-measure, precision and recall values. So, for checking the spelling errors, the score is 73% recall, 76% precision and 75% f-measure. To suggest the correct spelling for the detected error, it has scored 78% recall, 62% precision and 70% f-measure. For checking grammar errors, the system has scored recall 47%, precision 90% and f-measure 68%, whereas, for the task of suggesting the correct grammar for identified error, it has 61% recall, 71% precision and 66% f-measure.

According to the experimental result presented, the framework has performed well for handling false alarm issues than error detection. The main reason that false alarm handling tasks scored higher than error detection tasks is that the written rules are primarily focused only on the common errors that mislead the meaning or make the statement meaningless. The evaluation performance for the task of grammar error detection within a statement scored lower performance. The problem with the task of error detection happened as a result of uncovered cases within the rule. Within a given statement, if the grammar problem is a semantic error, it will not be detected and/or corrected by checking only the sentence structures. To find and/or correct the semantic errors of the statement, it is must to analyze the semantic structure of the sentence. Therefore, it's possible to solve a challenge to detecting grammar errors due to the dynamic occurrences of the error.

VII. CONCLUSION

This study has developed an integrated model to develop a grammar checker for Afaan Oromo using a rule-based. Approach integrated a spell-checker along with a morphological analysis on Afaan Oromo grammar checker to improve its performance. Data has been collected from various sources such as Oromia Broadcasting Network, BBC Afaan Oromo and other sources used for designing and implementation purposes.

The experimentation was done in two phases: spell-checking & correction phase and grammar-error detection & correction phase. The system is evaluated on detecting wrong spelling and grammar from the text and the possible suggestion made to the error. As a result, the evaluation result for detecting grammar errors in a statement has a low score than other activities. One of the main reasons is due to the uncovered cases within the rule for detecting the error.

VIII. RECOMMENDATION

The designed prototype for improving the Afaan Oromo grammar checker has a promising result. Even if the model performance showed a good result, there is still work that needs to be done to improve the performance. The following are recommended as a future research direction.

- The grammar error can be syntactic, related to the structure of a sentence and/or semantic, which is related to the meaning of the sentences. But this work attempted to solve the grammars having only syntactic errors. Therefore, further research work has to be done on the Afaan Oromo Grammar checker that can check the semantic analysis of the sentences.
- Since it is difficult to catch all of the grammar errors using only a set of rules, it must implement another method. Therefore, machine learning methods combined with a rule-based approach may improve the performance of the Afaan Oromo grammar checker.

REFERENCES

[1] Anonymous., "the-major-importance-of-grammar-check-website," 01 February 2016. [Online]. Available: https://www.nounplus.net/blog/the-major-importance-of-grammar-check-website/. [Accessed 22 March 2020].

[2] Kualo, "Oromo language, alphabet and pronunciation," omniglot, 2010. [Online]. Available: http://www.omniglot.com/writing/oromo.htm. [Accessed 14 October 2018].

[3] S. D. Baviskar and S. S. Bahekar, "Comparative Study of Rule-Based Approach for Grammar Checker," *International Journal of Management, Technology And Engineering,* vol. IX, no. I, p. 1315, 2019.

[4] G. O. Ganfure and D. Dr. Midekso, "Design And Implementation Of Morphology Based Spell Checker," *International Journal of Scientific & Technology Research*, vol. 3, no. 12, 2014.

[5] D. Tesfaye, "A rule-based Afan Oromo Grammar Checker," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 8, 2011.

[6] Mideksa, "Statistical Afaan Oromo Grammar Checker," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[7] Mainsheet, "What is Python?," javatpoint, 07 April 2011. [Online]. Available: http://www.javatpoint.com/what-is-python. [Accessed 18 October 2018].

[8] M. Gasser, "HORN MORPHO2.5 User's Guide," Research group of human language technology and the democratization of information, 2012.

[9] K. Markham, "Simple guide to confusion matrix terminology," Data School, 25 March 2014. [Online]. Available: https://www.dataschool.io /simple-guide-to-confusion-matrix-terminology/. [Accessed 08 March 2021].

IoT Soil Monitoring based on LoRa Module for Oil Palm Plantation

Ahmad Alfian Ruslan[1], Shafina Mohamed Salleh[2]
Sharifah Fatmadiana Wan Muhamad Hatta[3], Aznida Abu Bakar Sajak[4]*
Computer Engineering, MIIT, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[1, 4]
Computer and Network Engineering, Gerik Vocational College, Kuala Lumpur, Malaysia[2]
Electrical and Electronics Engineering, University of Malaya, Kuala Lumpur, Malaysia[3]

***Abstract*—Internet of Things (IoT) Soil Monitoring based on Low Range (LoRa) Module for Palm Oil Plantation is a prototype that sends data from the sender to the receiver by using LoRa technology. This realises the implementation of Industrial Revolution 4.0 in the agriculture sector. Also, this prototype uses the TTGO development board for Arduino with built-in ESP32 and LoRa, pH sensor and moisture level sensor as main components. The prototype utilises the LoRa communication between the sender and the receiver. The sensors will detect soil pH along with the moisture level. The data then will be sent to the receiver, where it will be displayed in the Organic Light-Emitting Diodes (OLED) display. At the same time, the data will be uploaded to the database named ThingSpeak by using wireless communication. Users can monitor the data collected by accessing ThingSpeak's website using smartphones or laptops. The prototype is easy to set up and use to help users monitor the pH level and moisture level percentage. For future enhancement, the project can be enhanced by combining temperature and tilt sensors to get comprehensive data about the soil's condition.**

Keywords—Internet of Things (IoT); Low Range (LoRa); Organic Light-Emitting Diodes (OLED); ThingSpeak; Arduino

I. Introduction

This project concerns an Internet of Things (IoT) soil monitoring system using Arduino based on LoRa (Low Range) technology. The project is created for future technology "Industry 4.0" to boost efficiency in palm oil plantation by adopting the IoT in the agriculture sector [1]. Most of the oil palm plantations are in a rural and secluded area where the internet is scarce. Hence, LoRa introduces a new method to realise IoT in the agriculture sector. This is the contribution of this paper. Up to now, palm oil workers manually check and monitor every crop, covering a vast area measuring thousands of hectares [2]. This can be less efficient to monitor each crop for the soil moisture level because oil palm is challenging to grow.

Moreover, after they harvest the crops, they will chop off the old crops to plant new crops. This practice leads to many problems, such as boycotts from climate activists. With the existence of this project, it can reduce the frequency of replanting when monitoring alerts you exactly as soon as it is due for replanting. Oil palm has already been a staple crop, with thousands of hectares being planted and manually monitored by the workers, but it is not the most efficient way to implement it. The workers need to manually check and maintain the appropriate soil moisture level throughout the year. The farmers mostly rely on the little knowledge they gain from experience or by observing crops with naked eyes. This leads to misjudgment, causing devastation to the crops. If a sufficient level of moisture is not maintained, it can be damaging to the crops. For example, maintaining the water supply is crucial to the crops [3-7]. With LoRa and a few sensors, the moisture readings can be viewed easily while also saves time. The design of the monitoring system is utilising the sensor module comprising of temperature sensor and moisture level sensor positioned in the soil that will capture the data taken from the soil. The data then will be transferred to the Arduino development board using the LoRa module.

LoRa is an important part of this project because the plantation lacks internet access and cellular coverage. LoRa was first introduced in the previous work in [8], where the LoRaWAN platform had shown the potential to replace the use of IoT infrastructure on the existing 802.11 Wireless Standard. In [9], an animal movement tracker prototype that used LoRaWAN as a medium to transfer the GPS location of the farmed animals had been built to support smart agriculture. Hence, using LoRa will help to cover the wide radius of the plantation. Unlike the prototype in [10], where the IoT soil monitoring prototype used wireless connectivity as the medium, the LoRa module will act as a network connection to connect to Arduino by using radiofrequency. After that, the data will be kept in the database to be monitored by the plantation workers using mobile devices or laptops.

The scope of this project is to develop an IoT project with the implementation of the LoRa module as the main network connectivity. Using LoRa, the system will have the benefit of low power consumption. Besides, it supports a longer range of coverage area. For measuring data, a pH sensor is needed to measure acidity or alkalinity with a range from 0 to 14. Another sensor is a moisture sensor, which is needed to detect moisture level in the soil and manage irrigation systems (the system will turn on when moisture level falls below a certain predefined threshold value) if available. A mobile application will be developed as well to present the data for the workers.

In summary, this project is created to provide a solution to overcome the manual monitoring system used by the workers for a very long time. The IoT monitoring system will ensure to increase productivity and efficiency of the planter at the tip of their finger.

*Corresponding Author

This paper will discuss the methodology of the project, the hardware and software development, the results and discussion and the conclusion.

II. Methodology

The research methodology section provides a step by step guideline to complete the project. A suitable methodology for this monitoring system can ensure the success of the project. In this section, the explanation of the chosen methodology as shown in Fig. 1, along with the description of how it is applied in developing a prototype, will be discussed. The result will also be discussed to gather as much information as possible about soil monitoring using the LoRa module. Fig. 2 is the flowchart of the project.

A. Feasibility Study

A feasibility study is an analysis or evaluation for the project proposed, either it is feasible to be implemented. Thus, the feasibility study consists of the problems statement, objectives, and phases to complete the project. The developer must have various strategies or mitigate planning if something wrong happens in terms of resources required, such as the cost of development and the development time.

B. Requirement Analysis

In this phase, all of the possible requirements and information regarding the project was collected and documented. The important thing to do in this phase is to identify the project's objectives, requirements, and specifications to proceed to the next phase.

C. System Design

In the design phase, the prototype was designed. The design must meet the requirements stated in Table I. It will have TTGO ESP32 with a built-in LoRa module, pH sensor, and moisture sensor on the transmitter side. For the receiver part, it will be designed to have TTGO ESP32 with a built-in LoRa module. ThinkSpeak will act as a database to store the data.

D. Coding

In the coding phase, the transmitter sent data; the receiver must receive the data. The transmitter (client) hardware consists of a pH sensor and a moisture sensor; both were fixed to the soil, while both the sensor and LoRa module will be installed to the TTGO ESP32. On the other hand, the receiver side consists of Arduino UNO and installed with the LoRa module to receive data from the transmitter. A database was configured with ThingSpeak to store the data. A mobile application was installed in the receiver to view the data. The TTGO ESP32 was configured by using the Arduino IDE from a laptop. The connectivity enables the data to be sent and received by using the LoRa module.

E. Testing

The project was tested to identify any defects in the system. Each of the connected sensors was tested whether it is working properly. Any defects found in the system were fixed and tested again to ensure that the defects were fixed. The test was conducted by using natural material (soil) to ensure the correct data was collected. This ensures the objectives are achieved. LoRa module was tested to ensure the connectivity between the transmitter and the receiver. ThinkSpeak was able to store the correct pH and moisture level readings.

F. Maintenance

In this phase, a maintenance phase was conducted to correct any errors discovered during the product development phases. This ensures developers always improve implementation and enhance the functionalities of the system.

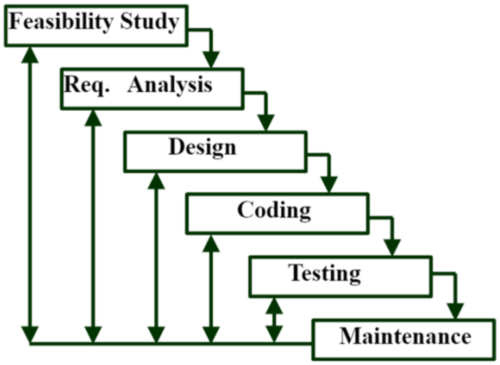


Fig. 1. Iterative Waterfall Model.

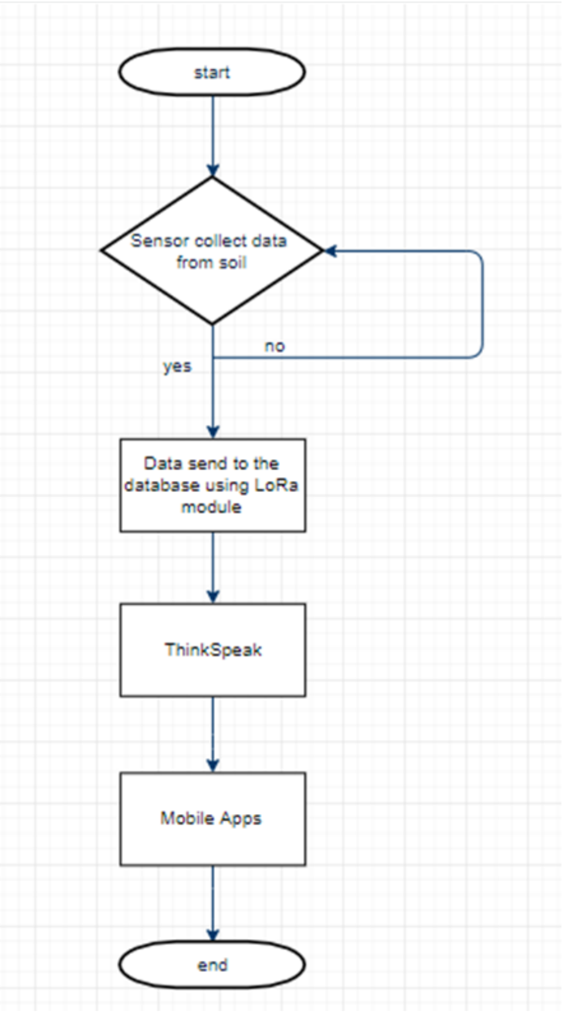

start
Sensor collect data from soil
no
yes
Data send to the database using LoRa module
ThinkSpeak
Mobile Apps
end

Fig. 2. Project Flowchart.

III. Hardware and Software

In this project, the hardware, as shown in Table I, is necessary for development. This step explains the project's purposes, requirements, and specification, including tooling and software. This phase is important for gathering project information to progress to the next phase.

A. *TTGO LoRa SX1278 with ESP32*

TTGO LoRa SX1278, shown in Fig. 3, is the main component for both the transmitter and the receiver. This board is built-in together with LoRa SX1278 and ESP32. Thus, the board is working on LoRa and WiFi communication. The device was used in both the transmitter and the receiver to act as a medium to transmit and receive the data over LoRa communication.

B. *Analogue pH Sensor*

Based on Fig. 4, pH sensors are connected to the board TTGO LoRa. There were three pins involved to connect the pH sensor to the TTGO LoRa board.

- VCC = +5.5V pin
- GND = GND pin
- GPIO = 35

This sensor can be used to detect the pH value from soil and water. An example can be seen in Fig. 4. The test was made to test the pH value from the distilled water.

TABLE I. Hardware and Software

	Item	Unit
Hardware	pH sensor	1
	Moisture sensor	1
	TTGO LoRa SX1278 ESP32 (OLED)	1
	TTGO LoRa SX1278 ESP32	1
	Jumper cable	1
Software	Arduino IDE	-
	Laptop	-
	ThinkSpeak	-

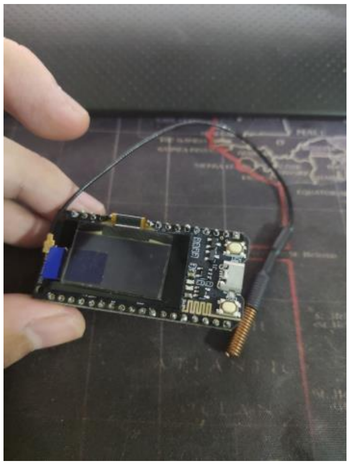


Fig. 3. TTGO ESP32 Build-in LoRa Module.

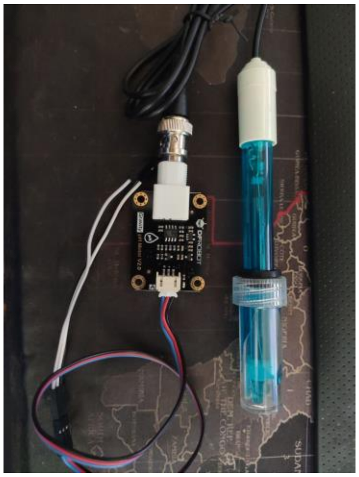


Fig. 4. pH Sensor by DFRobot.

C. *Moisture Sensor*

Based on Fig. 5, this sensor has three pins: VCC, GND, and Analog Output, to connect the sensor to the TTGO board.

- VCC = +5.5V pin
- GND = GND pin
- GPIO = 32

This sensor can be used to detect the moisture of water. The sensor will be read in the form of voltage. From the voltage, the higher reading can be interpreted as a high level of moisture. To facilitate reading, the data will be converted into a percentage.

D. *ThingSpeak*

A database is needed to ensure the data received can be collected and processed, and ThingSpeak, as shown in Fig. 6, is the best choice for this. The user accessed the ThingSpeak website to register and create a database channel.

Simple coding, as shown in Fig. 7, was needed at the receiver to ensure the data received was properly sent to the ThingSpeak database. The “apiKey” can be found from the channel. “MY_SSID” & “MY_PWD” were set up based on the network that connected to the internet.

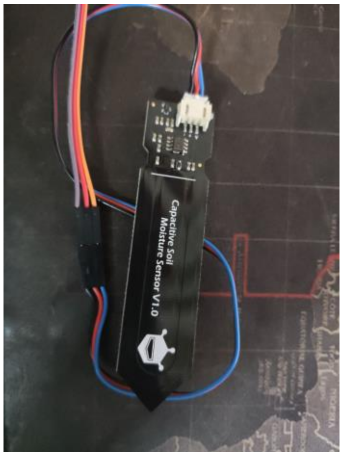
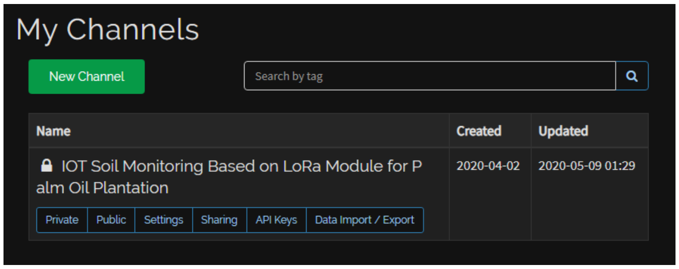


Fig. 5. Moisture Sensor.

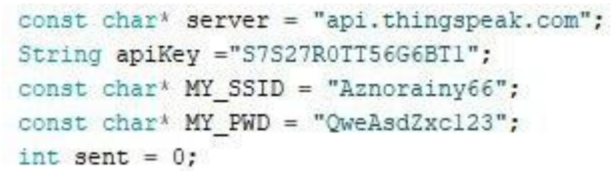


My Channels

New Channel | Search by tag

Name	Created	Updated
IOT Soil Monitoring Based on LoRa Module for Palm Oil Plantation Private | Public | Settings | Sharing | API Keys | Data Import / Export	2020-04-02	2020-05-09 01:29

Fig. 6. Project Channel at ThingSpeak.



```
const char* server = "api.thingspeak.com";
String apiKey ="S7S27R0TT56G6BT1";
const char* MY_SSID = "Aznorainy66";
const char* MY_PWD = "QweAsdZxc123";
int sent = 0;
```

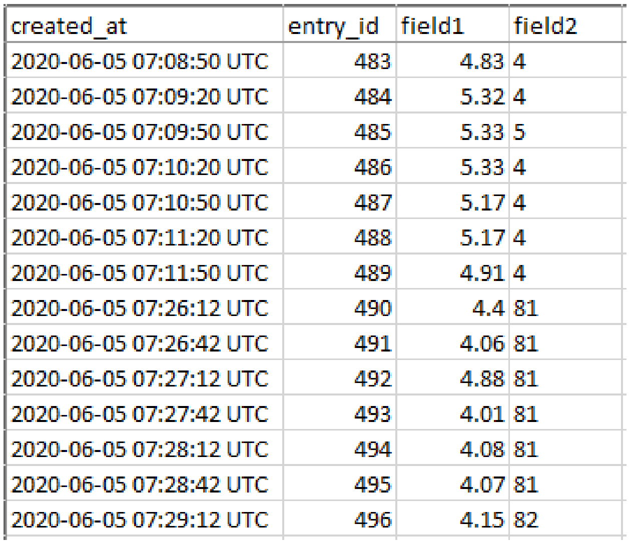
Fig. 7. Code for ThingSpeak.

IV. Result and Discussion

Fig. 8 shows the results retrieved from the sensor, stored in the channel's database. In Fig. 8, there were four columns. The first column is to show the exact time when the data was collected. The second was the "entry_id", which was the sequence of data collected. The third was the "field1", the data retrieved from the pH sensor. The last column was "field2", which was the data from the moisture sensor. All of the data was retrieved together in a packet.

A. Results

Fig. 9 shows the data from the pH sensor, which was plotted to a line graph along with the numeric value. For the line graph, the x-axis is the date as per data being collected, while the y-axis is a range of numbers of the pH readings, which ranged from zero to 14. The line graph will show us the flow of data collected through the duration setup. The graph was set up to show the latest ten collected data. The numeric value will show the latest data retrieved to make it easier to read.



created_at	entry_id	field1	field2
2020-06-05 07:08:50 UTC	483	4.83	4
2020-06-05 07:09:20 UTC	484	5.32	4
2020-06-05 07:09:50 UTC	485	5.33	5
2020-06-05 07:10:20 UTC	486	5.33	4
2020-06-05 07:10:50 UTC	487	5.17	4
2020-06-05 07:11:20 UTC	488	5.17	4
2020-06-05 07:11:50 UTC	489	4.91	4
2020-06-05 07:26:12 UTC	490	4.4	81
2020-06-05 07:26:42 UTC	491	4.06	81
2020-06-05 07:27:12 UTC	492	4.88	81
2020-06-05 07:27:42 UTC	493	4.01	81
2020-06-05 07:28:12 UTC	494	4.08	81
2020-06-05 07:28:42 UTC	495	4.07	81
2020-06-05 07:29:12 UTC	496	4.15	82

Fig. 8. Database Table.

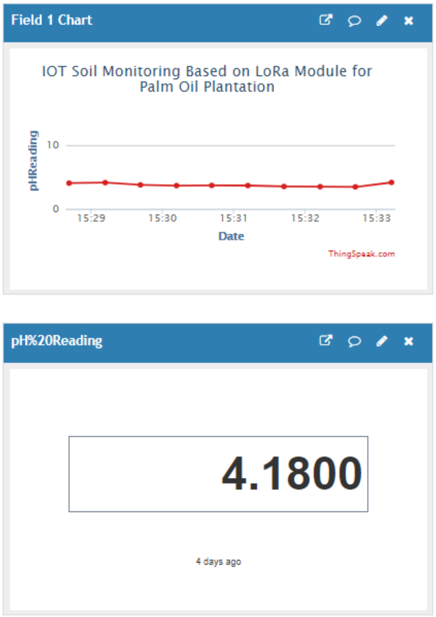


Fig. 9. pH Graph.

In Fig. 10, the data from the moisture sensor was plotted to a line graph, and a numeric value was also displayed. Referring to the line graph, the x-axis is the date as per data being collected, while the y-axis was the range of numbers from zero to 100, which corresponded to percentages of the moisture level. The graph was set up to show the latest ten collected readings. The numeric value showed the latest data retrieved for ease of reading.

Fig. 11 showed the database could be accessed via smartphones. It has a similar interface as the PC browser interface, which consisted of a line graph and a digital reading of pH and moisture level. Thus, the users can understand and view the data from their smartphone.

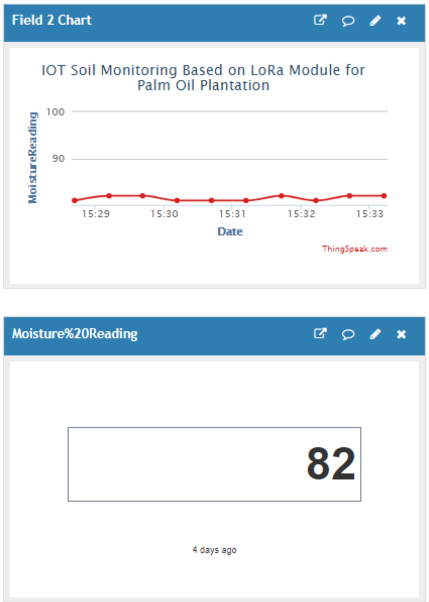


Fig. 10. Moisture Graph.

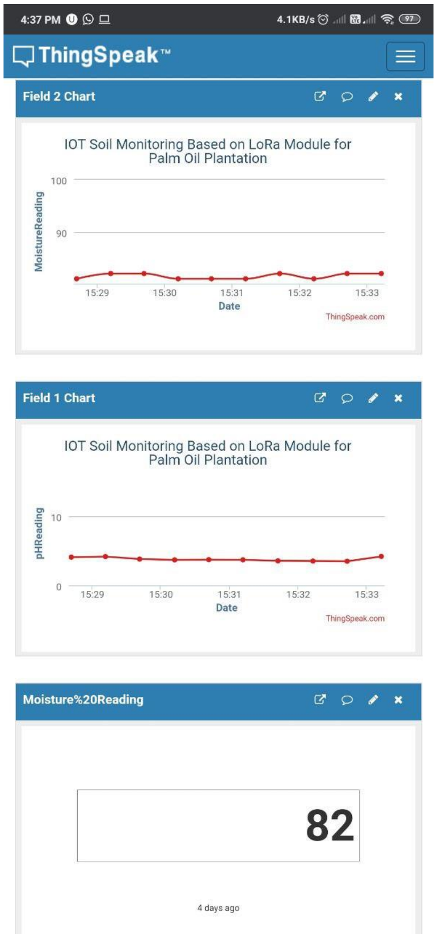


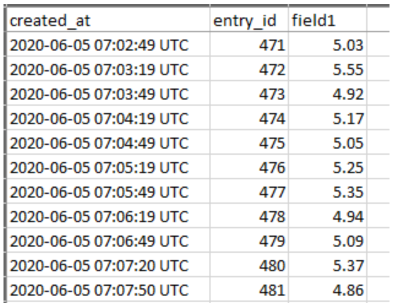
Fig. 11. ThingSpeak via Smartphone view.

B. Analysis

1) Recommended soil pH: Fig. 12 shows the soil pH readings from the ThinkSpeak database in tabular form. The ideal recommended soil pH for oil palm is between 4.3 to 6.5 [6], which is more acidic. Other readings, aside from the recommended values, will be harmful to the crops. From the table, the entry_id readings between 471 and 481 are ideal. When the readings are between this range, the worker doesn't have to do anything. The worker only takes action if the reading from the sensor is outside the recommended value.

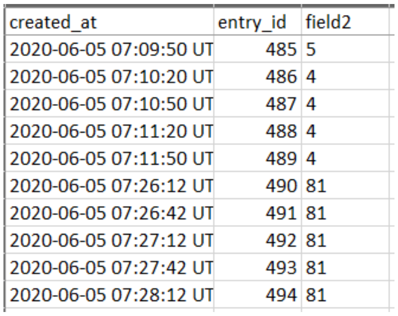
2) Recommended soil moisture: Fig. 13 shows the soil moisture reading from the ThinkSpeak database in tabular form. 80% [7] soil moisture level is the recommended reading for oil palm. Taking the data from the table as an example, the readings of entry_id between 490 and 494 were sufficient, while the readings between 485 and 489 were insufficient, as this meant that the soil was so dry and can cause the crops to die due to lack of water. Thus, the worker needs to water the plant to moist the soil to the recommended moisture level.

3) LoRa signal analysis: Fig. 14 shows the map of the checkpoints for the testing sites. For this test, there were six different locations to test out the LoRa connectivity. The test sites were located around the residential area and high-rise buildings.



created_at	entry_id	field1
2020-06-05 07:02:49 UTC	471	5.03
2020-06-05 07:03:19 UTC	472	5.55
2020-06-05 07:03:49 UTC	473	4.92
2020-06-05 07:04:19 UTC	474	5.17
2020-06-05 07:04:49 UTC	475	5.05
2020-06-05 07:05:19 UTC	476	5.25
2020-06-05 07:05:49 UTC	477	5.35
2020-06-05 07:06:19 UTC	478	4.94
2020-06-05 07:06:49 UTC	479	5.09
2020-06-05 07:07:20 UTC	480	5.37
2020-06-05 07:07:50 UTC	481	4.86

Fig. 12. Soil pH Database.



created_at	entry_id	field2
2020-06-05 07:09:50 UT	485	5
2020-06-05 07:10:20 UT	486	4
2020-06-05 07:10:50 UT	487	4
2020-06-05 07:11:20 UT	488	4
2020-06-05 07:11:50 UT	489	4
2020-06-05 07:26:12 UT	490	81
2020-06-05 07:26:42 UT	491	81
2020-06-05 07:27:12 UT	492	81
2020-06-05 07:27:42 UT	493	81
2020-06-05 07:28:12 UT	494	81

Fig. 13. Soil Moisture Database.

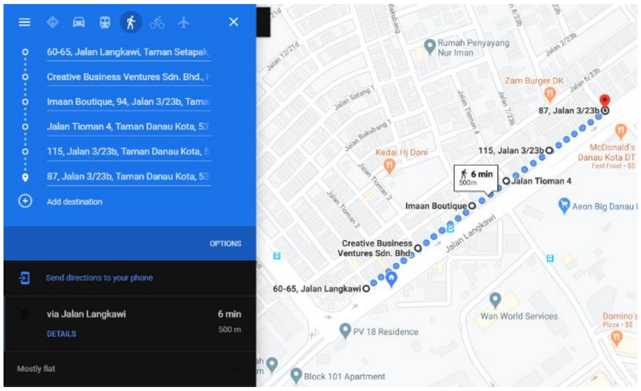


Fig. 14. Checkpoint Test Location.

Table II shows that the distance for each checkpoint, A through G. It showed that maximum LoRa connection was available up to 500 meters, and the connection was lost at checkpoint G. This might be due to interferences coming from cars, motorcycles, trees or might be due to the loss of Line of Sight (LOS). Next, for the received signal strength indicator (RSSI) measurements, the signal was much better as the signal value neared 0 dBm. At checkpoint A, the RSSI reading is -79 dBm, while at another checkpoint, RSSI's was greater than -100 dBm. When the signal is -100 dBm, and below, the signal strength is not good as the better signal strength readings should be closer to 0 dBm.

TABLE II. LORA CONNECTIVITY RESULT

	A	B	C	D	E	F	G
Distance (m)	5	100	200	300	400	500	700
Connectivity	✓	✓	✓	✓	✓	✓	x
RSSI	-79	-114	-117	-116	-110	-116	x

V. Conclusion

Palm oil production is vital for the Malaysian economy, which is the world's second-largest producer of the commodity, after Indonesia. Thus, the implementation of IoT in this agriculture sector will greatly increase the profit to the country. As internet connectivity is scarce in rural area, where most agriculture sectors are located, Lora had been identified as the solution for data transmission. Using this project, we can monitor every aspect of oil palm management and plantation. With the increasing application of IoT in the sector, the efficiency and productivity will be increased and the profit margin of the palm oil industry. Workers can view and monitor real-time soil condition along with various other agriculture variables via apps or websites. This is because of the sensors installed in the plantation; with the telemetry, all of the data can be displayed on the screen ready to be analysed, and solution can be searched for should there be any problem with the data obtained. Also, unnecessary and troublesome travels across the plantation to monitor the condition of the crop can be minimised. Adding more sensors such as tilt and temperature sensors such as in [11] will result in a more comprehensive prototype that can test the soil's condition more accurately.

References

[1] “Ruler: Industry 4.0 tech lagging in palm oil sector | The Star.” [Online]. Available:https://www.thestar.com.my/business/business-news/2019/07/16/ruler-industry-40-tech-lagging-in-palm-oil-sector. [Accessed: 09-Jun-2020].

[2] "Industrial-scale pyrolysis the easiest to use: The Role of Biochar In Era of Precision Farming and IoT in Palm Oil Plantations." [Online]. Available: http://jfe-project.blogspot.com/2017/07/the-role-of-biochar-in-era-of-precision.html. [Accessed: 09-Jun-2020].

[3] R. Dagar, S. Som, and S. K. Khatri, “Smart Farming - IoT in Agriculture,” Proc. Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2018, no. Icirca, pp. 1052–1056, 2018.

[4] S. Verma, R. Gala, S. Madhavan, S. Burkule, S. Chauhan, and C. Prakash, “An Internet of Things (IoT) Architecture for Smart Agriculture,” Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–4, 2018.

[5] S. Heble, A. Kumar, K. V. V. D. Prasad, S. Samirana, P. Rajalakshmi, and U. B. Desai, “A low power IoT network for smart agriculture,” IEEE World Forum Internet Things, WF-IoT 2018 - Proc., vol. 2018-Janua, pp. 609–614, 2018.

[6] R. Rozieta, A. R. Sahibin, and I. Wan Mohd Razi, “Physico-chemical properties of soil at oil palm plantation area, Labu, Negeri Sembilan,” in AIP Conference Proceedings, 2015, vol. 1678.

[7] “Plantations International Palm Oil.” [Online]. Available: https://www.plantationsinternational.com/palm-oil/. [Accessed: 09-Jun-2020].

[8] J. Jaffar, A. F. Abdul Malik, M. Farez Azril Zuhairi, A. A. Bakar Sajak and M. Taha Ismail, "Development of the LoRaWAN-based Movement Tracking System," 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), Taichung, Taiwan, 2020, pp. 1-6, doi: 10.1109/IMCOM48794.2020.9001689.

[9] S. Alimin Mahama Chedaod, A. Abu Bakar Sajak, J.Jaffar, and M. Sallehin Mohd Kassim, “LoRaWAN based Movement Tracker for Smart Agriculture”, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol.9, No. 1.5, July 2020, p. 253 – 258.

[10] M. Shakhil Iman Hasnan, R. Rawi and A. Abu Bakar Sajak , “Palm Oil Soil Monitoring System for Smart Agriculture”, The International Journal of Integrated Engineering, Vol. 12, No. 6, September 2020, p. 189 – 199.

[11] M.Syarafuddin Md Saleh, M. Ariff Majmi Zaaba, R. Mohamad and A. Abu Bakar Sajak “ IoT Real-Time Soil Monitoring based on LoRa for Palm Oil Plantation ”, unpublished.

Workload Partitioning of a Bio-inspired Simultaneous Localization and Mapping Algorithm on an Embedded Architecture

Amraoui Mounir[1], Latif Rachid[2]
LISTI, ENSA
Ibn Zohr University
Agadir,80000, Morocco

Abdelhafid El Ouardi[3]
SATIE, Digiteo Labs
Paris-Saclay University
Orsay, France

Abdelouahed Tajer[4]
LISA, ENSA
Cadi Ayyad University
Marrakech, 40140, Morocco

Abstract—**Many algorithms were developed to perform visual localization and mapping (SLAM) for robotic applications. These algorithms used monocular or stereovision systems to solve constraints related to the navigation in unknown or dynamic environment. The requirement of SLAM systems in terms of processing time and precision is a factor that limits their use in many embedded applications like UAVs or autonomous vehicles. Meanwhile, trends towards low-cost and low-power processing require massive parallelism on hardware architectures. The emergence of recent heterogeneous embedded architectures should help design embedded systems dedicated to Visual SLAM applications. It was demonstrated in a previous work that bio-inspired algorithms are competitive compared to classical methods based on image processing and environment perception. This paper is a study of a bio-inspired SLAM algorithm with the aim of making it suitable for an implementation on a heterogeneous architecture dedicated for embedded applications. An algorithm-architecture adequation approach is used to achieve a workload partitioning on CPU-GPU architecture and hence speeding up processing tasks.**

Keywords—Simultaneous localization and mapping (SLAM); Bio-inspired algorithms; CPU-GPU workload partitioning; embedded systems; visual acuity (VA); hardware/software codesign

I. INTRODUCTION

The robot navigation is not always possible in some special circumstances, due to the unavailability of a map or because the environment keeps changing. Hence, for its localization, the robot needs to know its pose accurately which is not possible without a map, which brings us back to the initial issue. The dilemma of what should come first the map, or the pose is complex to solve, because it needs to calculate the pose and construct the map at the same time. A solution to this navigation problem is a Visual Simultaneous Localization and Mapping system also called V-SLAM [1]. The implemented algorithms use data inputs from different sensors mounted on the mobile robot, like cameras since they provide much more information about the environment.

To run a Visual SLAM algorithm, the minimum hardware requirement is a CPU based architecture and a monocular camera. Usually, as depth cannot be observed with one camera, most of systems use multi-view techniques to allow map reconstruction. Camera data is used to perform the scene photogrammetry and rebuild the trajectory map. This sensor presents several advantages (price, availability of a high amount of information) compared to other sensors it also has constraints such as the need for calibration, and sensitivity to light changes and intensity.

Bio-inspired approaches are based on learning concepts from nature and applying them to design an enhanced real time SLAM system. Hence most of these algorithms are aiming to simulate the biological retina and brain-based methods for features detection and description, which make the model complex and also its parallelization a real challenge.

Eyes represented by cameras are used to provide inputs data for front-end operations, but images are processed in a different way compared to classic methods. It can be categorized into simple eyes with one concave photoreceptor lens like for humans, and compound eyes like for some insects with multiple lenses [2]. This study is limited to simple eyes because it is the most feasible to be simulated in a mobile robot. Binocular and stereo vision systems can be then considered as a reference.

Simple eyes are grouped into two known categories based on their photoreceptor's cellular construction. Therefore, human and rodent were selected because they are both on the top of each group, since they are representing the best visual acuity (VA) [3].

In previous studies [4 - 5], bio-inspired approaches were proved to be very competitive methods, in term of accuracy and execution time compared to classic ones. This paper is sharping these results by studding two different models: the retina model HOOFR-SLAM [6] and hippocampal model for rodent RAT-SLAM [7].

The contribution is:

- an efficient partitioning model of a bio-inspired algorithm on a heterogeneous CPU-GPU architecture to improve the processing time performance.
- the evaluation of this model on a dedicated architecture based embedded application.

The aim is to be able to determine if bio-inspired V-SLAM methods can be used on a real-time application, despite their algorithmic complexity.

The reported resulting values are given based on the mean of 20 run results for at least 1000 timestamps. Datasets used are from the well-known KITTI benchmark [8] and Oxford New college opensource dataset [9]. Calibration and other algorithm parameters were adapted accordingly to each dataset to always have the maximum performance.

II. Background

One of the advantages of Rat-SLAM navigation system is its ability to run in dynamic environment using cheap camera sensor, due to the nature of Rat retina that has a low Visual acuity (VA) compared to the human one [3]. From another hand the use of odometrical data in this algorithm is mandatory in order to combine visual scene matching and a semi-metric topological map representation. The used image model is a simulation of a biological rat retina perception. HOOFR-SLAM doesn't need any odometrical data and can rely only on inputs provided by a well calibrated camera.

Human vision is trichromatic because it has three types of color cones: long-wavelength “red”, middle-wavelength “green”, and short-wavelength “blue”. Rat's vision is dichromatic because, they have only two wavelengths: a shorter blue wavelength than human but shifted toward and middle "green". Therefore, rats can see into the ultraviolet, this doesn't mean that they are color blind, they just have different color perception compared to human.

It is very important to understand how images are seen by each bio-inspired system. So, based on above, RAT-SLAM have to spend less time to convert images to gray scale than HOOFR-SLAM, with better adaptation to lower light environment, but accuracy is also lower, since human Visual Acuity (VA) is higher, this is why the use of an additional sensor (odometer) is needed.

Since Rats have eyes on both sides of the head, the vision is binocular, the field is large and panoramic. So, a representation in a simple flat screen will distort the image, since computer screens are flat and adapted to human forward-facing eyes which have a smaller visual field and more binocular vision. Due to these facts, running the same dataset on HOOFR-SLAM and RAT-SLAM will be a real challenge due to the difference in image perception and cameras position, and must be converted and adapted to each bio-inspired system before use, also the dataset will never have the same length, making a direct comparison simply not possible.

Now, after standing up on all the differences related to visual performance and image perception, the next sections are describing the Rat-SLAM and HOOFR-SLAM algorithm concept and the details of each model front end, back-end, kernels and evaluate their algorithmic complexity.

Finally, a parallelization and an implementation on a single CPU, multi-cores CPU and CPU-GPU on laptop, then a Jetson TX1 system on ship (SoC) architecture.

III. The RAT-SLAM

A. Algorithm Inputs

The Rat-SLAM system as defined by [7] use a self-motion sensor data to create a representation known as experience map, also used to facilitate the exploration. This technic is used in robotic when the human intervention is not or low needed. This navigation system has two goals, short term goals using a local obstacle map, and long-term goals relaying on experience map, with the aim to reach the desired destination. These two steps can be resumed to Matching [10] and mapping processes, if compared to HOOFR-SLAM as per Fig. 1.

The RAT-SLAM algorithm uses two different sensors, as a main data input source; a camera and an odometer. If a robot operating system (ROS) is used, both inputs will be then received as ROS messages. Also, for simulations reasons, both inputs camera and odometrical messages derived from the mobile robot wheels encoders are provided together in a dataset file (*.bag file), therefore in this case a visual odometry functional block will not be used, this data is included in the opensource dataset benchmark from [9].

Cameras are placed on both sides of the robot to simulate rat eyes position, the vison is binocular, so the depth is calculated using information on head direction, calibration, functional regions of the hippocampus and surrounding areas. The camera motion is limited to the translational and rotational speed and it is estimated via an image comparison process, more details on how it is done are provided by [11].

The rotational velocity estimation is done by minimizing horizontal offset of two consecutive scanline profiles that are generated by summing the images in the vertical direction represented by *f* as follows:

$$f(S,I^j,I^k) = \frac{1}{w-|s|}\left(\sum_{n=1}^{w-|s|}\left|I^j_{n+\max(s,0)} - I^K_{n+\min(s,0)}\right|\right) \quad (1)$$

where S is the profile shift, I^j and I^k are the scanline intensity profiles to be compared, and w is the image width. The pixel shift S_m in consecutive images I^j and I^k is the value of s that minimizes *f* for those two profiles.

$$S_m = \underset{s\,\epsilon|\rho-w,w-\rho|}{\operatorname{argmin}}\ f(S,I^j,I^k) \quad (2)$$

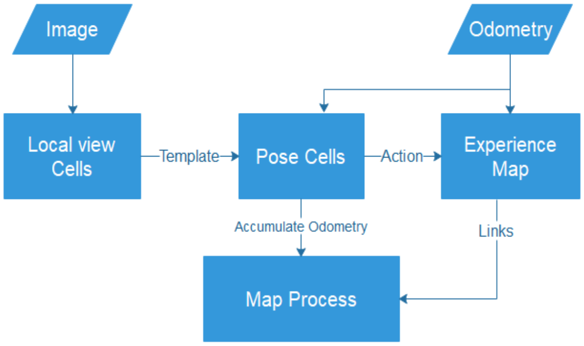


Fig. 1. Rat-SLAM Vision System Diagram with Inputs and Outputs Related to each Functional Block (FB), Local view Cells and Pose Cells are the Font-end and can be Considered as the Tracking Process, using Camera Image Data Plus Odometry as Inputs to Extract Features (Neural Approach) and Provide Actions as Output to the Experience Map. The Mapping Process uses Both Pose Cells Cumulated Odometry Outputs and Links to Avoid Infinity and to Close the Loop.

where ρ is the offset that ensures that there is enough overlap between the profiles. The translational velocity estimation is limited to a maximum value to prevent large changes in illumination and done by multiplying the minimum difference by a scaling factor. Two ROS message are then provided to the system: sensor_msgs::CompressedImage and nav_msgs:Odometry.

B. Main Functional Blocks

RAT-SLAM is a robust bio-inspired navigation algorithm, used when dependence on human assistance is not or low needed, based on sensor inputs message data described above. The algorithm has three main functional blocks and sub-functional blocks as per Fig. 2. They are identified as follows:

Local view cells: This node uses image data from camera, to make a matching process, by using a comparison technique in order to determine if the visualized scene is familiar or novel. The block algorithm sequence is the following:

Local view cells algorithm:

```
Image Conversion, Cropping, sizing and Matching (FB1)
{
 Convert image to gray scale ()

/* For Rat-SLAM the input image is already converted Remove visually bland features like roads */

 Cropping image ()
 Convert_view_to_view_template ()

/* cropping to specified region of the original camera image */

 Do patch normalization ()
 Create_template ()

/* Calculate the sum of absolute difference between current view template that represents the current camera image, and all previously learnt templates*/

 Compare ()
}
```

Pose Cells: as per Fig. 1, it uses the odometric data, visual template from local view cell, to provide actions to make decisions used by experience map node. This node has two inputs and one output that are processed as follows:

- Template: is the output of the local view cell image comparison process, if the templates already exist it is directly injected into associated location in the pose cells, if not a new a local ID is associated to the pose cell network. Also, in case of a successive detection of the same template for a long period of time, it could be understood as the robot is not moving.
- Odometric: inputs are processed in Pose Cells; the functional block algorithm sequence is described as follow:

Pose Cells Algorithm:

```
Check templates and update template database (FB2)
{
 if (New Template)
 create_view_template ()
 else (Template exist)
 give a local template ID ()
 }
Local and global excitation, path integration, action and energy management (FB3)
{
 Local excitation ()
/*Add energy around each active pose cell */
 Local inhibition ()
 /*Remove energy from around each inactive pose cell */
 Global inhibition ()
 /*Remove energy from all active pose cells above zero */
 Network energy normalization ()
 Path integration using odometrical information ()
 Get_current_exp_id ()

/* Identify the centroid of the dominant activity packet in the network */
 Get_action ()
 Update_scene ()
 }
```

Experience Map: Manage based on both local view and pose Cells outputs, the graph building, relaxation and path planning. This block has three main outputs:

1) A path message with information on the trajectory to the goal.

2) A list of nodes and edges that makes a representation of the experience map.

3) The robot poses in the experience map.

4) Get the goal way point

The experience map algorithm is the following:

Experience Map algorithm:

```
Update experience map get new target (FB4)
{
 Update Experience map ()
/* update the current position of the experience map since the last experience*/

 Calculate_path_to_goal ()
 Get_goal_waypoint ()
 }

 Map Converging, add new goal (FB5)
{
 CREATE_NODE
 CREATE_EDGE
 SET_NODE
 iterate ()
 /* iterate the experience map. Perform a graph relaxing algorithm to allow the map to partially converge */
Set_goal_pose_callback ()
Add_goal ()
/* Setting and handling goals */
 }
```

According above description, RAT-SLAM software architecture allows a multithread processing; therefore it can be parallelized in an heterogenous System on chip (Soc). However, the proposal is to proceed with the code profiling first, to stand on every function block use rate, calculate accurately the number of times each function is called and get an overview on the memory usage for each block.

C. Code Profiling and Workload Identification

In order to have better understanding about how the algorithm is behaving once run on a CPU, the source code has been profiled using a profiling tool to measure the processing time of the functional blocks, the parameters dependencies and hence the functional blocks that could be parallelized.

Profilers are designed tools for analyzing and improving performance of code execution. They allow analyzing the algorithm, to measure, while the code is running, how long a routine takes to execute, how often it is called, where it is called from and how much of total time at some spot is spent executing that routine.

By using the opensource Callgrind profiling tool [12], which is considered as a binary instrumentation profiler, the call history could be recorded among functions for every main functional block. Data collected and represented in Table I, correspond to the CPU and memory workloads using an Intel® core i7 @1.8 GHZ CPU with 8 GB RAM. Also, the cache memory usage for reading and writing operations is given. This information is important because it reveals about congestion in different memory buses.

A fast information is then given about the algorithm behavior, to target where the code parallelization can be more efficient, of course an adequation architecture algorithm analysis is mandatory to have a deeper understanding and make the maximum optimizations that can be done.

TABLE I. RAT-SLAM C++ WORKLOAD ON A CPU

FB	CPU workload %	Cache memory reads %	% Cache memory writes %
Pose cells	28.98	28.30	29.99
Local View Cells	41.61	42.77	39.70
Experience Map	29.41	28.93	30.31

D. CPU-GPU Parallelization

Computing architectures have known the generalization of the concept of parallelism thanks to their significant technological evolution in recent years. Parallelization has spread to the level of processor architectures, notably with multicore processors and superscalar computers. In this context, founders and manufacturers of graphics processors have developed their architectures to be able to use them in applications with generic processing by designing GPGPU (General Purpose Graphic Processing Unit). Development languages like CUDA, OpenCL or OpenGL have been designed to use the potential of these processors for massively parallel computing purposes.

To our knowledge, in the state of the art, there is no study on the performance evaluation of a bio-inspired SLAM algorithm on GPU-based architectures.

The interest here is to explore hybrid architectures based on CPU-GPU for the implementation and parallelization of a bio-inspired SLAM algorithm. A first challenge lays in the complex data structure which results in the nature of the studied algorithm. Another interest lies in the specificity of the parallelization and programming of CPU-GPU architectures, which differs from conventional methods of parallel programming requiring mapping optimization and processing synchronizing with different coprocessors and shared memories.

Hence, the RAT-SLAM algorithm was divided into blocks with well-defined inputs and outputs (in number, type and size of data). An algorithm consists of a processing description, so it is the same for the blocks composing it.

The modeling in this work is based on a graph representation that illustrates the dependencies between the different blocks of the algorithm.

In the case of a heterogeneous architecture, the execution time of each functional block and therefore of the algorithm strongly depends on the mapping, otherwise on the way in which the functional blocks are distributed between the different processing units. The goal of the mapping optimization is to find one or more implementations allowing to reach defined constraints of real-time.

Given the results about the workload of each functional block, knowing the cache memory number of read/write accesses, measuring the execution time of each block and sub-functions, and taking into consideration the algorithmic complexity analysis, the nature of interaction between functional blocks can be understood. So, functions that needs to be sent to GPGPU for a parallel processing are separated from the ones that need to remain on the host, because processing them on CPU will be adequate due to the sequential processing and to make benefit from the higher frequency of CPU.

So, based on the above, this study propose an optimized parallelization targeting a CPU-GPGPU heterogenous architecture, where all time-consuming functional blocks related to image processing, visual template generation, intermediate map creation, pose cell experience and matching, have been transferred to the GPGPU device as explained in Fig. 2. The CPU is then managing the sequential processes and acting as a host.

As per Fig. 2, the selected functions for a parallel computing were converted to CUDA. Also, for evaluation purposes, the algorithm is executed first on a laptop (see Table II for hardware specifications), then on a Jetson TX1 which is often used in automotive applications due its better performance compared to its predecessor TK1 (see Table III for Jetson Tegra X1 hardware specifications).

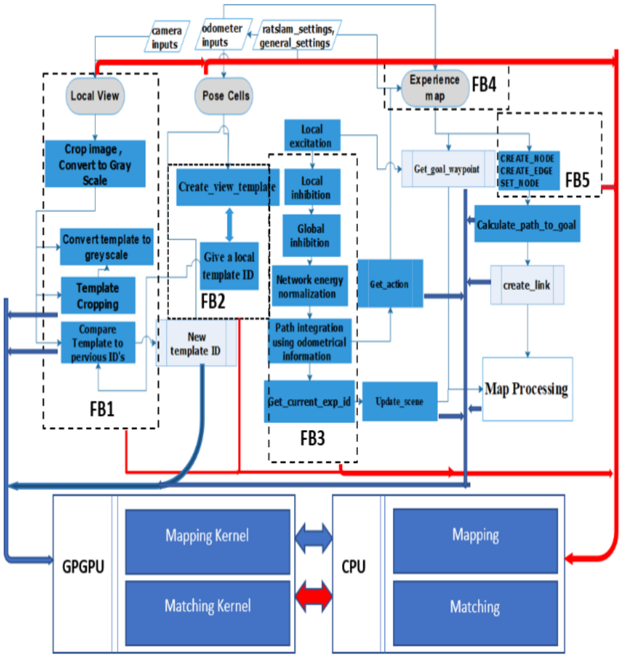


Fig. 2. Rat SLAM Algorithm flow Diagram: Interactions between main Functional Blocks and Sub-functions. Odometric Data and Image Inputs are Consecutively given by an IMU and a Camera from [9]. Buses in Red Represents the Bottleneck in Terms of Data Transfer. Bidirectional Arrows Represent Data Transferred between CPU and GPU.

TABLE II. HARDWARE SPECIFICATIONS OF THE CPU-GPU LAPTOP

CPU: Intel® core i7 @1.8 GHZ
GPU: NVIDIA GeForce MX110
RAM: 8035640 KB

TABLE III. HARDWARE SPECIFICATIONS OF JETSONTX1

CPU: Quad-Core ARM® Cortex®-A57 MP Core
GPU: 256-core NVIDIA Maxwell™ GPU
Memory: 4GB 64-bit LPDDR4 Memory

IV. HOOFR-SLAM

HOOFR is a bio-inspired binary detector, inspired from human retina. As introduced by Nguyan et al. [6], it is a novel model inspired from Hessian ORB - Overlapped FREAK (HOOFR) and based on the combination of the ORB detector [13-14] and the FREAK bio-inspired descriptor [15]. HOOFR use a similar method as ORB except that for filtering features it uses, instead of Harris filter, a Hessian Matrix represented by equation 3, together with Gaussian represented by equation 4 for smoothing.

$$H = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix} \qquad (3)$$

Where I is the pixel intensity and $\partial x, \partial y$ are the derivative in x and y direction, respectively.

$$G(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \qquad (4)$$

Where G(x,y) is a Gaussian 2D distribution and σ is the standard deviation of the distribution.

Because the human retina has a better visual acuity, the rotation estimation is done based on the sum of gradients over a selected pair, by using long pairs to compute the global orientation, whereas select pairs are mainly with symmetric receptive fields with respect to the center, as per Fig. 3.

The set of selected pairs is done according to the following equation 5:

$$O = \frac{1}{M} \sum_{P_0 \in G} \left(I\left(P_0^{r_1}\right) - I\left(P_0^{r_2}\right)\right) \frac{P_0^{r_1} - P_0^{r_2}}{\left\| P_0^{r_1} - P_0^{r_2} \right\|} \qquad (5)$$

where M is the number of pairs in G and P is the 2D vector of the spatial coordinates of the center of receptive field. However, compared to classical binary descriptor (like BRISK), HOOFR retinal pattern has more error in the orientation due to a larger receptive field allowing more estimation error, but this issue is solved by discretizing the space of orientations in much bigger steps, leading to more than 5 times smaller memory load (about 7 MB against 40 MB), and therefore a smaller execution time.

The main difference between Rat-SLAM and HOOFR-SLAM matching method, is that the first algorithm is comparing the template image against the source image by sliding it, then a sum of square is calculated to give the movement direction, so this method is operating directly on the pixel value, while HOOFR, from the other hand, use a feature matching method. In general, for this kind of techniques, the accuracy and efficiency are strongly dependent on the used detector and descriptor. An evaluation has been done in [4-5] show the impact of execution time and accuracy by using different feature detectors/descriptors combination. Fig. 4 is schematizing the flow of each algorithm front-end process.

Unlike RAT-SLAM, HOOFR-SLAM has been parallelized in a previous work [16] and implemented in different heterogenous architectures, so it is re-evaluated on a laptop, in order to determine the acceleration rate in a the same hardware architectures, and to have a fear comparison with results obtained for Rat-SLAM.

Since both bio-inspired algorithms has different images perception and since the camera position in HOOFR is frontal but lateral for RAT, any comparison using the same dataset will not be equitable, even after a modification, because the number of evaluated templates will not be the same. Therefore, a corresponding dataset is used for each algorithm in order to get time spent per iteration, images examples are given by Fig. 5.

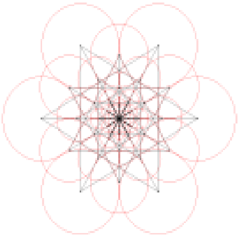


Fig. 3. Illustration of the Pairs Selected to Compute the Orientation.

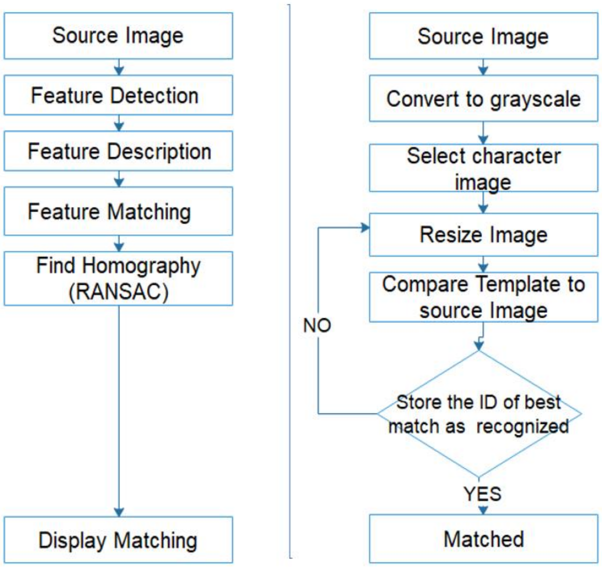


Fig. 4. Front End Process flow Diagram Showing differences between HOOFR-SLAM Feature Matching (Right) and RAT-SALM Template based Matching Method (Left).

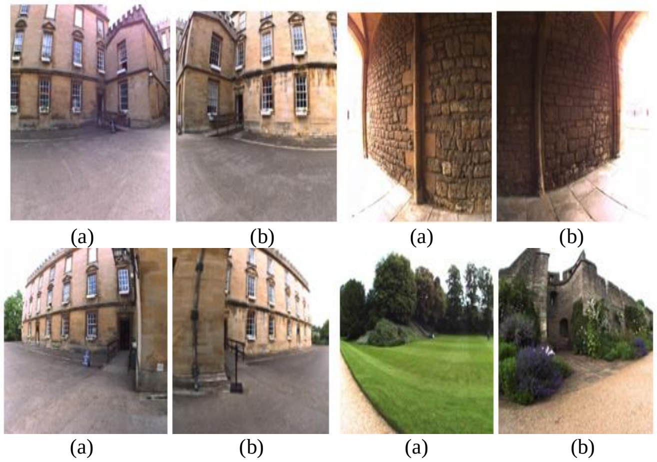


Fig. 5. Left and Right Images taken at different Lighting Conditions, by a RealSense Stereo Camera System, Simulating Rat Eyes and Placed in Lateral Position to have Similar Visual Perception as for a Rat. Right (a) and Left (b) Images used for RAT-SLAM, Images are Panoramic and May Appear Distorted to Human Eyes.

The evaluation was based on outdoor sequences where different lighting conditions are applied for better performances check.

V. Experimental Results

A first evaluation for both selected bio-inspired algorithms is done based on a Multi-core Intel CPU and an NVIDIA GPGPU with a high-end 64-bit implementation. This configuration shown previously in Table II, allows the evaluation of different run modes on a CPU, multi-cores CPU and CPU-GPGPU combined. All presented results are the mean values of 20 runs.

Tables IV and V gives the processing times per iteration when the functional blocks are transferred to GPGPU as shown in Fig. 6.

TABLE IV. Average Processing Time (MS) Evaluation for RAT-SLAM with Parallel Implementation on a Multi-core CPU, and a CPU-GPGPU Laptop

		Intel® core i7	NVIDIA GeForce MX110	Acceleration %
		Multi Core CPU	CPU-GPGPU	
Local View	FB1	183.17	31.93	82.57
Pose cells	FB2	0.75	0.22	70.67
	FB3	38.49	22.32	42.01
Experience Map	FB4	133.93	84.81	36.68
	FB5	0.16	0.08	50.00

TABLE V. Average Processing time (MS) Evaluation for RAT-SLAM with Parallel Implementation on an ARM CPU, and CPU--GPGPU using ARM-Nvidia TX1

		ARM®Cortex -A57	NVIDIA Maxwell™	Acceleration%
		Multi Core CPU	CPU-GPGPU	
Local View	FB1	111.76	32.24	71.15
Pose cells	FB2	0.75	0.29	61.33
	FB3	26.59	20.94	21.25
Experience Map	FB4	113.82	86.86	23.68
	FB5	0.13	0.09	30.73

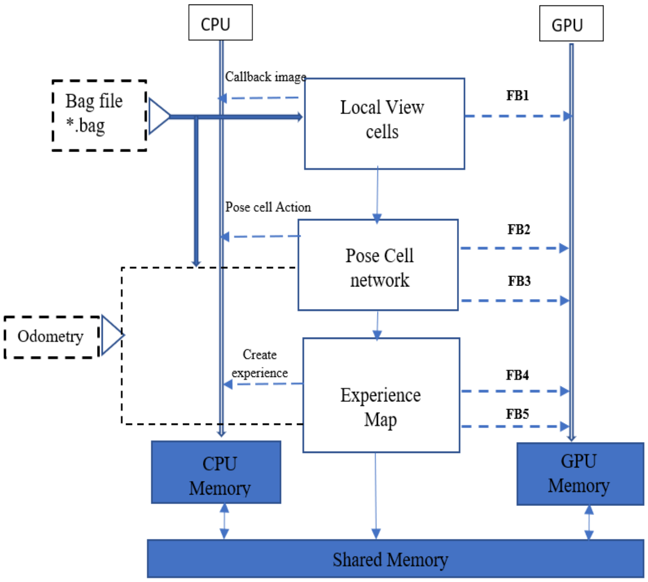
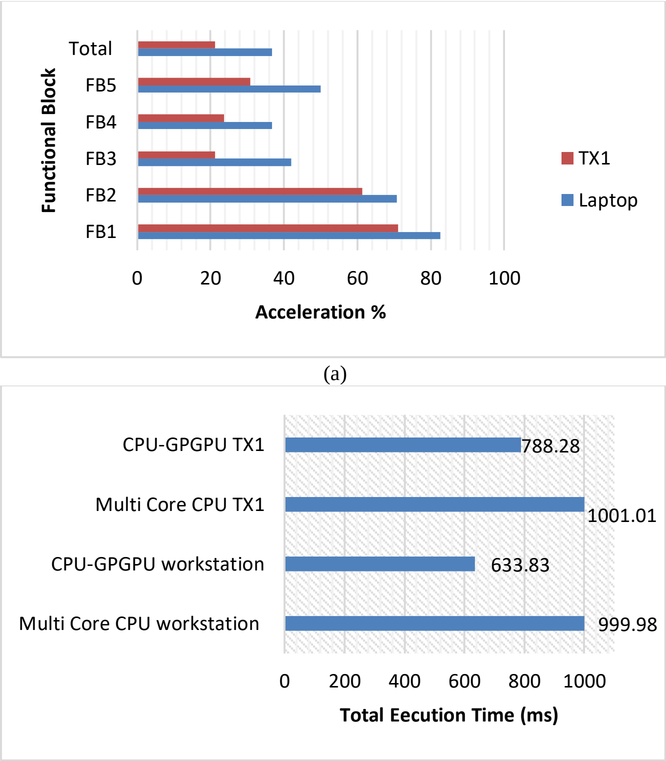


Fig. 6. Simplified Schematic Representation of Functional Blocks Sent to GPU and the One Remaining on CPU for Rat-SLAM.

The evaluation was achieved using the well-known open-source dataset Oxford New College, 2008 [9], settings and camera calibration parameters are shared by the same source. Images are panoramic and were converted to gray scale to match the most the rodent eyes perception to comply with the template matching process requirement.

Based on above results, the processing time of the most time-consuming functional blocks is drastically reduced using a parallel implementation by 36.68% on the laptop and 23.68% for TX1, this is due to the higher GPU performance and number of cores on the laptop compared to TX1, Fig. 7 is a graphical representation of the total execution time speed up for the algorithm on GPGPU.



(a)

(b)

Fig. 7. Comparison of Acceleration (A) and Total Execution time (B) for Oxford New College [9] on Laptop and on TX1 Hardware.

Below, graphs in Fig. 8, 9 and 10 represent the average execution timing measurement related to NVIDIA GeForce MX110 GPGPU and on ARM Cortex-A57 on TX1, also the power consumption in Watts for both Laptop CPU and GPU.

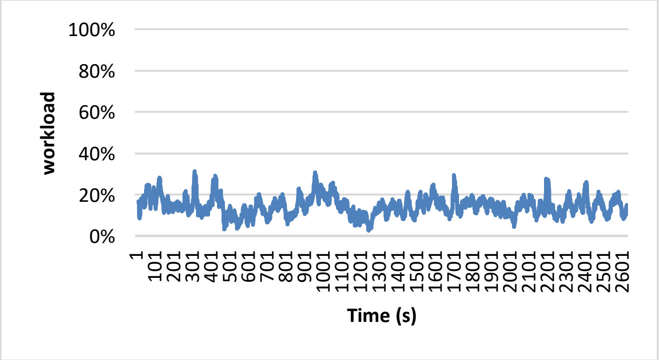


Fig. 8. Average Workload Evolution for RatSLAM on Laptop NVIDIA GeForce MX110 GPU, the Workload Everage is around 20% during the Total Execution Time of the Selected Dataset [9].

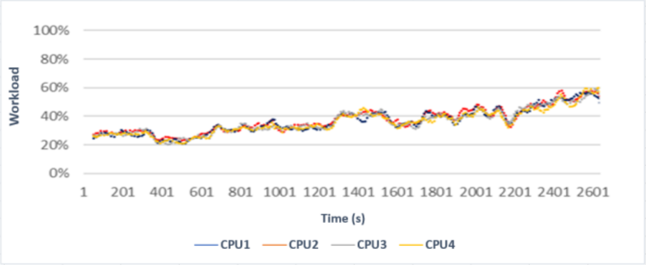


Fig. 9. Average Workload Evolution for RatSLAM on Laptop Intel Core i7 CPU, the Workload Everage is Variating between 20% and 60% during Execution of the Selected Dataset [9].

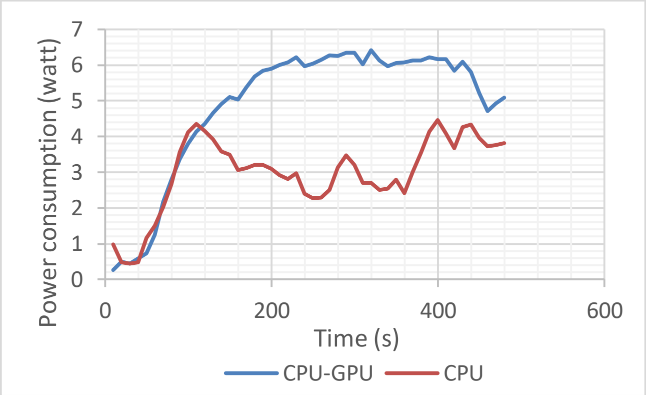


Fig. 10. The CPU-GPGPU Power Comsumption is Calcuted only for RATSLAM Algorithm when the Power Resulting form Operating System is not Considered.

As per above graphs , the CPU workload has been reduced by an average of 20% when running the RatSLAM algorithm, by running parallel functional blocks on GPGPU, but due to the nature of the algorithm some blocks cannot run on GPU because they are sequential and therefore are kept on CPU.

As expected the temperature will also increase on the CPU in the same way as per below Fig. 11, which can be considered as a week point for implementation of Rat SLAM in an embedded architecture where the use of a cooling system is not always possible.

An evaluation of HOOFR-SALM using KITTI-07 open source dataset, gives the results shown in Tables VI and VII.

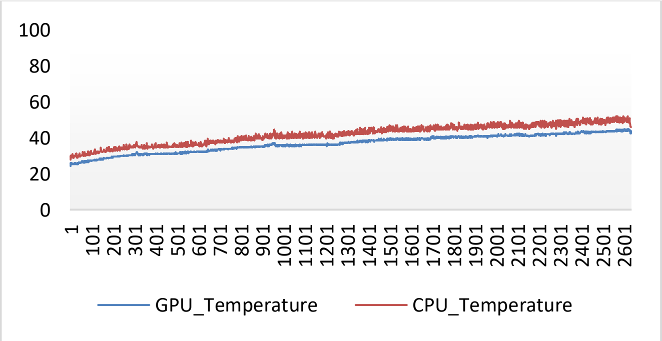


Fig. 11. Temperature Profile for Laptop CPU and GPU, it Increases Over Time Due to the Amount of Input Data.

TABLE VI. AVERAGE PROCESSING TIME (MS) EVALUATION FOR HOOFR-SLAM WITH PARALLEL IMPLEMENTATION ON MULTI-CORE CPU, CPU-GPGPU ON INTEL® CORE I7 / NVIDIA GEFORCE MX110

	Intel® core i7 / NVIDIA GeForce MX110				
	CPU (ms)	Multi Core CPU (ms)	Acceleration %	CPU-GPU (ms)	Acceleration %
Extraction	8.6	8.6	0.00	8.6	0.01
Mapping	52.79	52.76	0.06	27.88	47.15
Loop detection	15.38	15.34	0.26	8.05	47.52
Map Processing	0.48	0.45	6.25	0.19	58.85

TABLE VII. AVERAGE PROCESSING TIME (MS) EVALUATION FOR HOOFR-SLAM FOR PARALLEL IMPLEMENTATION ON MULTI-CORE CPU, CPU-GPGPU ON TX1 ARM®CORTEX-A57 NVIDIA MAXWELL™

	ARM®Cortex-A57 NVIDIA Maxwell™		
	Multi Core CPU	CPU-GPU	Acceleration %
Extraction	16.78	16.73	0.31
Mapping	21.92	16.47	25
Loop detection	21.92	16.47	25
Map Processing	0.58	0.4	31

VI. CONCLUSION

This paper presented an algorithmic complexity study for two bio-inspired algorithms. It proposed an optimized parallel implementation on a CPU-GPU by studying, in a practical way, optimization possibilities for workload partitioning. It also presented understanding of bio-inspired algorithms with necessary techniques to accelerate there processing times for real time SLAM applications.

From one hand, based on above temporal evaluation results, a first conclusion is that the use of multiple CPU's cores cannot accelerate much the algorithm compared to one CPU core. This is due to the congestion of data in buses at the on-chip memory level since it is a shared resource for all CPU's cores. The memory is considered as a bottleneck in this case and using a higher CPU frequency or more memory will not help much.

From another hand, considering a real time sequence where the frequency is higher than 30fps, and based on the experimental results when executing both algorithms, it is clearly seen that despite the considerable acceleration, Rat SLAM still cannot fulfill the real time expectation as it should be executed in less than 33ms per frame, due to the matching sequence that is dependent on the number of images perceived by the camera sensor. Furthermore, the algorithm needs to keep previously seen templates in the memory for localization and to fine tune the map (loop closing), which has an impact on the final execution time.

In the case of HOOFR-SLAM, which is a feature-based approach that doesn't depend on the dataset size, the matching process time is not increased by increasing the number of input images. This is very important because the processing time will remain practically the same for all iterations and therefore the parallelization is more efficient.

These two studies covering the exploration of hybrid architectures based on GPU-CPU for the implementation and parallelization of bio-inspired SLAM applications, allowed to draw conclusions about the challenges to be met related to the complexity, the structure of data and the nature of the algorithms studied. The results presented in this paper confirm that future heterogeneous architectures will represent potential candidates to embed complex algorithms such as those of bio-inspired SLAM applications.

Future work will focus on the implementation of selected functional blocks on FPGA architectures in order to bring defined processing closer to the sensor and hence allow image processing on the fly and reserve the GPU for massively parallel processing.

AKNOWLEDGMENT

This research is supported by Information Systems and Technologies Engineering Laboratory, LISTI. ENSA Ibn Zohr University, Agadir, Morocco and SATIE Paris-Saclay University, Orsay, France.

REFERENCES

[1] T. Savaria and R. Balasubramanian, "V-SLAM: Vision-based simultaneous localization and map building for an autonomous mobile robot," 2010 IEEE Conference on Multisensor Fusion and Integration, Salt Lake City, UT, USA, 2010, pp. 1-6.

[2] Land, M.F.; Fernald, R.D. (1992). "The evolution of eyes". Annual Review of Neuroscience. 15: 1–29. doi:10.1146/annurev.ne.15.030192.000245. PMID 1575438

[3] Caves EM, Brandley NC, Johnsen S. Visual Acuity and the Evolution of Signals. Trends Ecol Evol. 2018;33(5):358-372. doi: 10.1016/j.tree.2018.03.001.

[4] M. Amraoui, R. Latif, A. Elouardi and A. Tajer, "Features Extractors Evaluation Based V-SLAM Applications," 2019 4th World Conference on Complex Systems (WCCS), Ouarzazate.

[5] M. Amraoui, R. Latif, A.E. Ouardi, A. Tajer "Feature Extractors Evaluation Based V-SLAM for Autonomous Vehicles", Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 5, pp. 1137-1146 (2020).

[6] D. Nguyen, A. El Ouardi, E. Aldea and S. Bouaziz, "HOOFR: An enhanced bio-inspired feature extractor," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, 2016, pp. 2977-2982.

[7] Milford, M.J., Wyeth, G.F., Prasser, D.: Ratslam: a hippocampal model for simultaneous localization and mapping. In: 2004 IEEE International Conference on Robotics and Automation, Proceedings, ICRA 2004, vol. 1, pp. 403–408. IEEE (2004).

[8] Andreas Geiger , Philip Lenz , Christoph Stiller , Raquel Urtasun; "Vision meets Robotics: The KITTI Dataset"; International Journal of Robotics Research, IJRR,2013.

[9] M. Smith, I. Baldwin, Churchill, R. Paul, Newman, "The new college vision and laser data set", The International Journal of Robotics Research, vol.28, issn.0278-3649, May.2009.

[10] U.Muhammad, M.Tanvir, K.Khurshid, "Feature Based Correspondence: A Comparative Study on Image Matching Algorithms" International Journal of Advanced Computer Science and Applications(IJACSA), 7(3), 2016.

[11] D.Ball, S.Heath, , J.Wiles, et al. OpenRatSLAM: an open source brain-based SLAM system. Auton Robot 34, 149–176 (2013).

[12] Antonio J. Peña, Pavan Balaji, "A data-oriented profiler to assist in data partitioning and distribution for heterogeneous memory

in HPC Parallel Computing”, Volume 51,2016, Pages 46-55, SSN 0167-8191.

[13] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262, 2017.

[14] E.Adel, M.Elmogy,H.Elbakry, “Image Stitching System Based on ORB Feature-Based Technique and Compensation Blending” International Journal of Advanced Computer Science and Applications(IJACSA),6(9),2015.

[15] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast Retina Keypoint”, In Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 510-517.

[16] Nguyen, Dai-Duong. “A vision system based real-time SLAM applications. (Un système de vision pour la localisation et cartographie temps-réel).” (2018).

An Approach based on Machine Learning Algorithms for the Recommendation of Scientific Cultural Heritage Objects

Fouad Nafis[1]*, Khalid AL FARARNI[2]
Ali YAHYAOUY[3]
LISAC Laboratory, Department of Informatics, FSDM
Sidi Mohamed Ben Abdellah University
Fez, Morocco

Badraddine AGHOUTANE[4]
IA Laboratory, Science Faculty
My Ismail University
Meknes, Morocco

***Abstract*—The Scientific Cultural Heritage (SCH) of the Drâa-Tafilalet region in south-eastern Morocco is a rich source of data testifying to the ingenuity of an older generation that has shaped the past of the region. These data must be preserved for future generations, particularly with new technologies and the semantic web. Recommendation systems (RS) are intended to assist prospective users in recommending the most suitable services based on their profile and expectations. The collaborative filtering (CF), content filtering (CB) or hybrid filtering (CF) RS has shown promising results in order to explore the problems experienced especially in CH. However, there are some limitations to be resolved, mostly due to the ability of these methods to build a stable and complete framework, which can provide a complete image of the user profile and suggest the most appropriate offers. This paper presents a hybrid recommender system for SCH data; a field little explored despite its historical importance and the value it generates. The results presented in this paper belong to the data collected from the region of Drâa-Tafilalet in southern Morocco.**

***Keywords*—Cultural heritage; CIDOC-CRM; ontologies; OWL; recommender system; semantic web; RDF**

I. Introduction

Given the many features and applications developed in the Big Data age, a Recommender System (RS) is an essential and effective user support tool. The user loses a considerable amount of time because the access to relevant information is difficult and the value of the services offered is challenged. An RS is a data filtering method that defines a collection of resources that are important to a particular user. There are three different forms of RS to think about. Content-driven SRs monitor user behavior and make suggestions for new items based on the user's interests. The guiding forces behind this work are the need to protect the Drâa-Tafilalet heritage. A region with a rare richness and diversity in its tangible and intangible cultural heritage, the use of which was restricted to a few a number of projects, without noticeable effects, by some public institutions. Using emerging technology will help make this heritage known, preserve it and make it a valuable added vector for a country with practically non-existent economic activities. Technologies for semantic web, Linked Data, for example, will go a long way toward resolving some of the issues that have arisen in several applications of this type. In particular, the lack and dispersion of data sources and the redundancy of information available on the web in data sources held by the public sector and private organizations. Using link tools, we can expect to have a single complete Moroccan Heritage data source utilizing multiple data sources. As a result, the stored data is automatically processed to extract useful information for the end-user, adding a significant dimension to the proposed system's architecture [1]. This document is structured as follows: Section II provides a synopsis of how SRS works for CH. Section III deals with some terminologies on Drâa-Tafilalet's unique and rich SCH. Section IV describes the architecture of the proposed system

and presents some results obtained, while Section V presents a conclusion and some future directions.

II. Related Works

The literature includes a limited number of works discussed by machine learning (ML) [13] algorithms for the issue of SCH conservation. This study focuses on the creation and testing of a SCH recommendation framework based on ML algorithms and semantic data through the use of a CIDOC CRM-reference model. The use of SRS in CH has been widely debated by researchers. In view of the great demand displayed, especially for cultural tourism, many players, especially in the last 10 years, continue to express their interest. In the semantic integration of the Draa-Tafilalet area CH data from a generic Big Data architecture a complete analysis was carried out and applied [2]. It is the primary motivation for this work as well as a rich source of data for the future works. The authors in [4] propose a novel methodology for implementing a route planner within cultural sites such as museums by combining recommendation facilities with agent-based planning techniques. No clear idea is provided on the data collected and its nature. The PCS data is not illustrated in this case. The authors present INTHELEX, an application of a first-order logic incremental learning system that allows learning the automatic identification rules of a wide range of essential document classes and their related components in [5]. When the set of documents is constantly expanded, incrementally plays a critical role. The Folksonomy-based Item Recommender System (FIRST) [24] is a CHAT-developed content-based recommender system. The goal is to propose a

*Corresponding Author

method for adaptive exploitation of digital libraries based on a flexible architecture that allows the recommendation of artworks located in the Vatican Picture Gallery (Pinacoteca Vaticana), providing users with a personalized tour of the museum based on their preferences.

CHAT-Bot [9] is a chatbot that recommends adaptive tourist itineraries with points of interest and associated services based on the tourist's profile and contextual factors. The authors of [10] present a Big Data architecture supporting popular applications for CH (query, content analysis, navigation). The objective is to propose a new user focused approach to suggest different cultural artifacts. SMARTMUSEUM [17] is a mobile application that offers customized, contextual knowledge and content suggestions to museum visitors using geo-location, Near Field Communication (NFC), and Radio Frequency Identification (RFID) techniques. In most of the works reviewed, CH is dealt with in a general manner or, alternatively, by taking into account a case study concerning a particular object that may be part of the tangible or intangible heritage. Processing of SCH objects such as zawyas, khizanas and oulamas is almost absent or in the majority of the cases treated from its historical and cultural side. This is devalorizing given its importance in the history and civilization of a generation that has tried so hard to make this heritage known. The use of new technologies of the semantic web such as RDF graphs and ontologies will surely allow attributing to the SCH data an interesting semantic side and therefore benefiting from all the advantages and advances that these technologies offer.

In this work, we focus on the SCH data. The number of objects belonging to this type of data and the dimension of each object are two important factors. This justifies the choice to focus on the objects: Medersas, Zawyas and Oulamas. Each Zawya may contain a Khizana or library containing documents of several types (books, manuscripts, documents, maps...)

The SCH objects processed can be classified according to the structure given in Fig. 1.

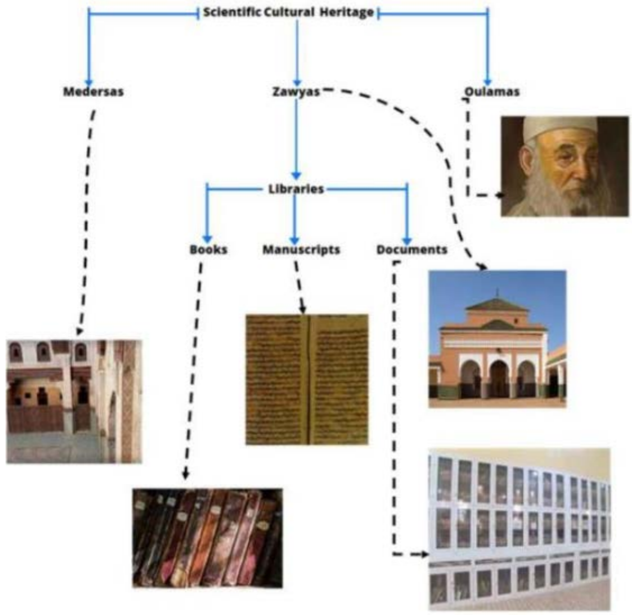


Fig. 1. The Different Forms of the Moroccan SCH Considered.

III. Terminology and Basic Functionalities

A. Scientific Cultural Heritage

Drâa-Tafilalet region in Morocco has a diverse and rich SCH. This heritage includes objects, buildings, and monuments from various periods and architectural styles, including religious architecture, funerary architecture, military architecture, and domestic architecture.

1) Zawya: Zawya is a Moroccan term that refers to the location where Sufi disciples congregate in the presence of Sufi sheikhs to "purify the soul and refine behavior." The Zawya is a scientific, social and religious entity that has received special attention because of a historical relationship between Morocco's rulers and the Zawya's sheikhs, who played an important role in the country's political stability and spiritual security. In general, every Zawya has a library with its name. The library contains books, old manuscripts and documents that reflect the country's history and culture. There are dozens of Zawyas in the region, many of which are still active and have been for centuries. Some of them still keep their extensive libraries.

2) Oulamas: The Oulamas are Muslim scientists who have worked in a variety of fields, including science, literature, and law. They have left an indelible mark on Moroccan history, as well as the history of all Islamic countries throughout history. Ultimately, the goal of this research is to identify the Oulamas associated with the Tafilalet region who lived within its walls, refine their talents and abilities, and improve their knowledge and achievements in Shari'ah and related sciences. It is therefore another type of CH of the diversity of the Drâa- Tafilalet region.

3) Medersa: A "medersa", "Madrassa" or Koranic school is a Muslim theological school operated by a religious foundation known as "waqf". The Merinid medersa of Morocco, located in Fez, is the most notable among them, with impressive architecture. In addition, those from the Drâa-Tafilalet region, who is main function is to teach the Koran and the Hadith. Semantic web references such as RDF/RDFS and Owl are used to represent the region's collected scientific heritage data. Several knowledge bases are included to enrich the information presented to the end user and ensure semantic consideration of user queries, regardless of formalism or language used. Table I illustrates a semantic representation of some SCH concepts.

These semantic equivalences are exploited to provide rich and optimized content since some descriptions of some commonly used cultural objects will derive from reference knowledge bases such as DBpedia. In Fig. 2, an overview of the representation of the Zawya NASSIRIYA library in Tamegroute is illustrated.

B. Basic Functionalities

In general, a SRS integrates the following main functionalities:

TABLE I. EXAMPLES OF SOME CONCEPTS AND THEIR SYNONYMS USED IN THE SCH

Concept	*related terms*	*DBpedia source*
Khizana	bibliotheque Maktaba	https://dbpedia.org/page/Library
Medersa	School Koranic school Madrassa	https://dbpedia.org/page/Madrasa
Book	Livre Kitab	https://dbpedia.org/page/Book
Manuscript	Manuscrit Makhtout Makhtouta	https://dbpedia.org/page/Manuscript

1) Acquisition/learning: Acquisition and learning are the first step of every SRS. It is a delicate and necessary step since any system's output depends on the quality and volume of input data, and it consumes the majority of the time spent on a project. The geographical dimension and constraints associated with the region of study, as well as the nature of the data collected, are also important considerations.

2) Creation of the user profile: The development of a user profile takes place according to the previous stage and the interactions an actor has in the system, particularly if the SRS is of hybrid type as in the study. One of SRS's most difficult challenges is the cold start problem, where the machine does not have details about the actor who is supposed to use it. This issue is addressed by assigning a default profile to each new user based on the semantic average of what is visited and what is most desired by other users, owing to the semantically close relationship between the objects manipulated by the system.

3) Extracting objects of interest: The next step is to extract the most interesting things for the user. Several techniques are used in SRS to ensure the semantic side of extracting the items of interest, which is dependent on the capacity of the system to interpret and understand user interactions (ontologies, RDF graphs, SPARQL queries, etc.). CIDOC-CRM standard was used for the integration of CH data of the study area. CIDOC CRM is the most widely used ontology and offers several advantages, which make it a reference, justifying its use in our case. The use of ontology aims to identify objects that are semantically linked to the user's interest items in order to ensure as rich as content as possible. The actor may be interested in the historical aspect of a KHIZANA from the area. However, he may be more interested in this one's content, which is rich in historical and cultural data.

4) Classification: The classification helps for the organization of the extracted objects based on their continuously modified profile and their importance to the end user. A number of algorithms were used in the literature to accomplish this mission, based on expected features and performance. The advantages of ML and DL technologies have been demonstrated in a wide range of fields, especially in CH (CNN, Naive-Bays, SVM, CNN, etc.) (Fig. 3).

5) Recommendation: A list of items that can interest the end user is produced in the recommendation process. This list is typically presented in the order of relevance based on the user's created profile. Based on the user's interactions with the system, this profile is constantly updated. A semantic link depth is defined to prevent the end-user from becoming lost in the massive amount of data presented to him. The best depth is currently defined by what is observed practically through user interactions, but it will be the subject of a much more in-depth study based on what is observed.

In this paper, we represent each object O_1 by a set of n features (f_1; f_2. . .; f_n). We note ***d*** the distance function used to compare objects. if D is a Database (relational DB and Knowledge base) of objects, the k-nearset neighbor of an object noted q can be described by :

$$NN(k,q,D) = \begin{cases} s \in D \backslash \forall\, r \in D\ d(q,r) \geq d(q,s)\ if\ k = 1 \\ NN(k-1, q, D \backslash NN(k-1,q,D)\ if\ k \geq 2 \end{cases}$$

The user through his interactions with the system identifies the inputs. The new connections are automatically assigned a set of concepts extracted from his profile or a generic profile. This algorithm is of the KNN type, and the number of k closest neighbors shown in the output adapts to the display device used by the end user. The depth is limited to three levels of semantic links extracted from the SCH concept semantic representation, as shown in Fig. 2. This restriction was imposed following a series of experiments and tests. The semantic distance is calculated for each extracted concept, and if it is less than a certain threshold, the item is added to the output list.

Algorithm 1 : Best item recommendation

```
Inputs     Target User : U
           Set of Concepts : C
           Number of items to recommend $n \in \mathbb{N}$
           $p \longleftarrow 3$ (link depth)
Outputs
foreach    $x \in C$ do
    $pr \longleftarrow depth\ (x, U)$
    If $pr \leq 3$ then
        $d \longleftarrow measure(x, U)$
        If $d \leq \epsilon$ then
            R.add(x)
        end
    end
end
```

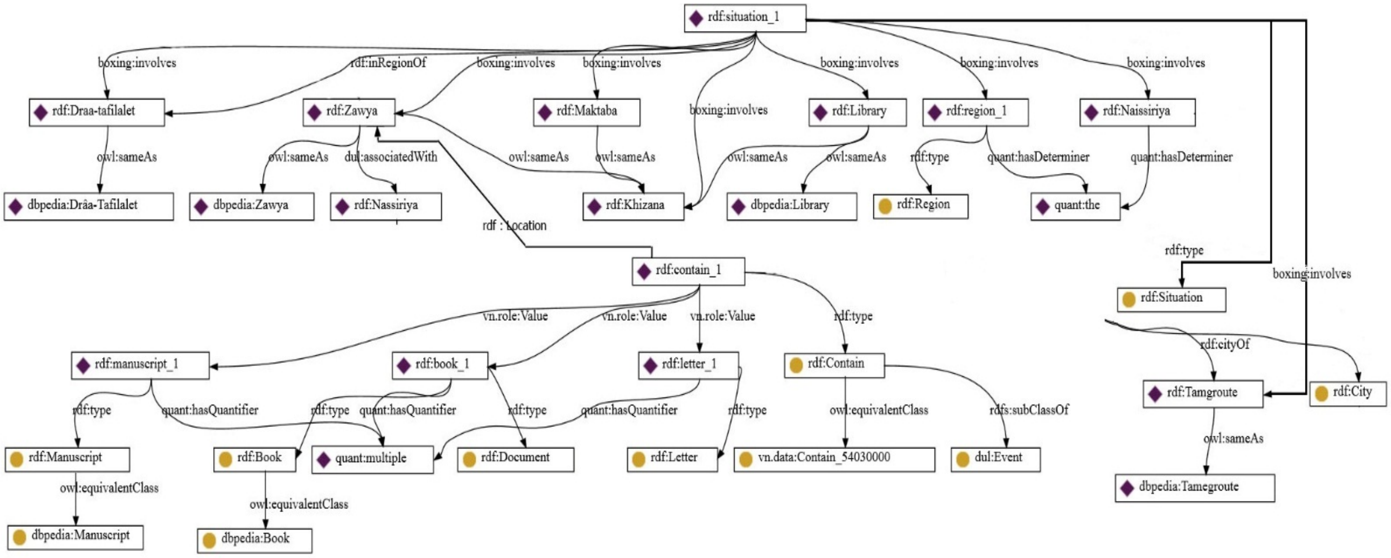


Fig. 2. Semantic Graph Representing the Khizana of Zawya NASSIRIYA in Tamegroute.

C. Algorithms of ML for CH

A preliminary study was carried out on the various ML algorithms used to process cultural heritage data (rule-based algorithms, genetic algorithms, SVM, KNN, etc.). The goal of this part is to find the most efficient algorithms to use for the CH of the study region and to increase the performance of the proposed system. This paper describes nine ML algorithms, which are as follows:

- SVMs (Support Vector Machines) [21]: This is a set of supervised learning algorithms that are frequently used for regression or classification. SVM is an algorithm based on statistical learning. These algorithms search for all data points similar to those in the other classes. These data points, known as support vectors, are used for the classification task, the others data points are ignored. The best dividing line, known as the decision boundary, is then defined.

- Naive Bayes (NB): This is a statistical approach founded on the theorem of the probabilities of Bayes. NB employs statistical functions to determine the likelihood that input is relevant for a specific predefined class. This algorithm returns the most likely class.

- Decision Tree (DT) [15, 19]: As a logic-based algorithm, DT models data sets in hierarchical structures using a series of if/else statement comparisons. Each node in the tree is made up of either decision nodes that contain terms (or objects that are more complex) or leaves that contain class label predictions. The weight of each word or object is labeled on the branches.

- KNN (K-Nearest Neighbors)[6, 19, 21, 22, 30]: this is a statistical method for predicting new input by calculating the similarity between the test data and the new instance by locating the closest data points (or data objects) in the training dataset based on certain distance functions. K denotes the number of closest data points (i.e., neighbors). The value of K is frequently determined using experimental test data.

- Rule-Based (RB) [5] classifications are algorithms in which collections of rules represent the data set. In contrast to DTs, which use a strictly hierarchical approach, RB classifiers allow for overlaps in the decision space. These rules are divided into two sections: the left side is made up of conditions, and the right side is made up of classes. The dataset is used to generate these rules.

- CNN (Convolutional Neural Networks) [20, 21, 23, 25, 30, 31] are widely used for image or video processing, including image classification. The term “convolutional network” refers to a mathematical concept known as the convolution product. In simple terms, we apply a filter to the input image, and the parameters of the filter are learned as we go. Following that, a learned filter will be used to recognize and classify a more complex image or object. To classify architectural heritage images, the authors propose and implement a pre-trained CNN such as GoogLeNet, resnet18, and resnet50 in [31]. The goal is to improve image database management and make it easier to search for a specific element, thereby facilitating the study and analysis of the relevant heritage object.

- Genetic-Based (GB) [6, 7]: GB algorithms are a subset of evolutionary algorithms. The goal is to use an optimization mechanism to approximate the solution to an NP-complete problem. The GB Algorithm begins with a population of candidate solutions known as individuals, which evolves from generation to generation until the first one contains the best solutions. Each individual has unique characteristics that can be influenced by genetic mutations (mutation, crossing, etc.). Each individual is evaluated, and its fitness value is used as a criterion for survival from generation to generation.

- Conditional Random Fields (CRF) [19] are statistical modeling techniques used in machine learning. A classifier predicts a single sample label without taking into account the context, whereas a CRF can. The prediction is represented as a graphical model, the type of which depends on the application, and it incorporates dependencies between predictions.
- Gaussian Mixture Models (GMM) [18,19] are probabilistic models that distribute points into different groups using the flexible clustering approach. GMMs assume a large number of Gaussian distributions, each of which represents a cluster. As a result, a Gaussian mixture model tends to cluster data points that belong to the same distribution.

In comparison to generic quadratic programming (QP) algorithms, Sequential Minimum Optimization (SMO) [19, 21] is a simple and efficient algorithm used to solve the learning problem in SVMs vectors. At each step, SMO decomposes the global QP problem into QP sub-problems and solves the minor possible optimization problem. Fig. 3 describes a summary of the outcomes obtained by distinguishing between tangible and intangible heritage objects. SRS is a type of semantic information filtering that has benefited greatly from the significant progress made in the semantic web world (ontologies, RDF/RDFS, OWL, etc.). A study was carried out in order to gain a clear picture of the most recent advances and technologies used in SRS for CH and its algorithms. Table II summarizes the findings.

The following are the most relevant comparison criteria that have been identified in terms of the basic functionality that any SRS for CH must provide:

- The objects of CH used: Cultural places, Tripes, Museum, painting, events, etc.
- The form of CH: Tangible (TCH) or intangible (ICH)
- Methods and Algorithms used: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Semantic Web (SW), etc.
- Type of Recommendation: Context-Based, Collaborative-Based, Hybrid, etc.
- Proposed GUI: Web, mobile, etc.
- Date: Date of start or publication of the first works and prototypes.

This will provide an idea of the historical side of the system.

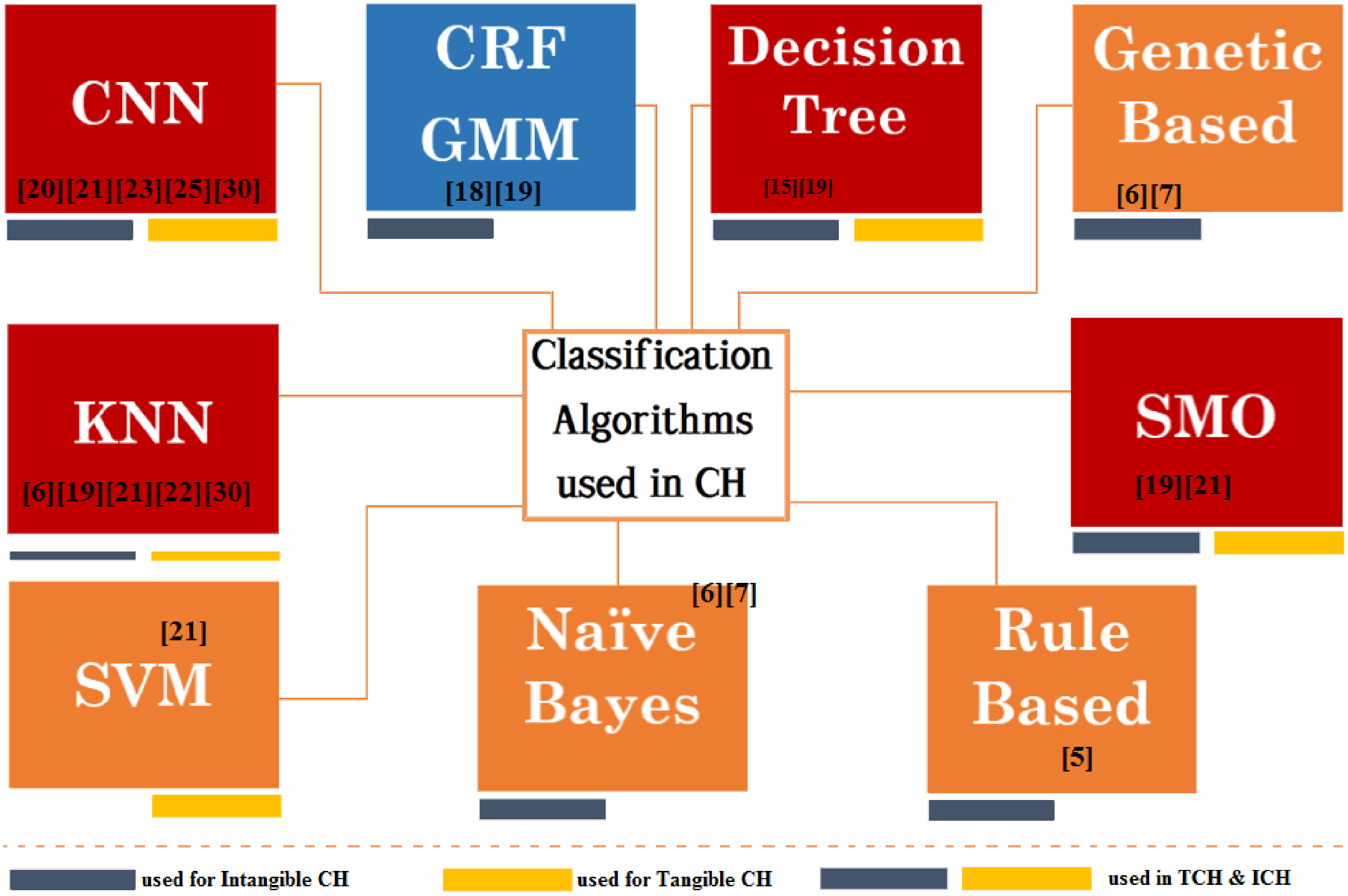


Fig. 3. Most used Classification Algorithms in the Field of Cultural Heritage.

IV. Results and Discussion

The developed system offers two options for support: a mobile application for Android systems and a web application. The end-user can access the rich content via a simple graphical interface, which provides relevant information on the KHIZANAS of the study region (Libraries). Each KHIZANA contains a number of documents (manuscripts, books, maps, geometric drawings, etc.). The user can also operate a query on the available data. Following that, three functions are carried out: Acquisition to extract keywords from a query and then detect important concepts implemented to feed the user profile or, if necessary, create a new profile. Several data sources are queried based on the entries of the users (database, knowledge base, .etc.). Based on the CIDOC CRM repository, an ontology was created at the start of this project to represent the SCH of the study region. The goals of this ontology are as follows:

- Integrate the semantic aspect to the system to guarantee complete handling of the user request.
- Have a controlled vocabulary for all the concepts used and their semantic links.
- Extract concepts semantically linked to those extracted from the user query (according to a predefined link depth) to provide more prosperous and more relevant content.
- Once this task is completed, the system provides adaptive content as recommendations that may interest the final user. The activation of one of these contents will help to enrich the profile already created.

The proposed system's architecture (Fig. 4) allows a user to interact by formulating queries and collecting the responses directly as SCH articles. It also makes recommendations based on the user's previous activities and, if necessary, the history of visits by other users if no profile has yet been created, reducing the cold start problem that any SRS must deal with. This is because SCH data is already semantically close, and the recommended objects have a high likelihood of being accepted by the current user. A particular work is done in the background to ensure the consistency of the recommendations provided. The back-end is primarily made up of algorithms that allow for the extraction of relevant concepts, their classification, and recommendation algorithms by implementing knowledge bases that allow for a semantic representation of the items related to the SCH and the various data collected, as well as the various records stored in various formats. The system administrator creates profile prototypes that correspond to the developed system preferences and functionalities and have a close relationship with the various SCH's forms that have already been introduced. The main problem that recommender systems face is the cold start problem. This is solved by creating a default profile (SCH) for each new user. This profile is increasingly customized to provide more targeted and relevant recommendations and services based on its interactions (and the history of other users) with the system (requests, consultation of proposed assets, comments, etc.).

TABLE II. Comparison of some Semantic RS

	System	Date	Objects used	Type of CH	Algorithms & techniques used	Type of RS	Proposed GUI
[27], [28]	-		Artifact, movies, books, music (songs & artists)	TCH	Linked Open Data techniques	Hybrid	
[5]	INTHELEX	2004	CH Documents	TCH	Graphs, Object identity, ML	CF	Web app
[4]	-		Museum objects	TCH	multi-agent systems	Hybrid	-
[17]	SMARTMUSEUM		Museum	TCH	Geopositioning and NFC/RFID techniques	CF	
[11]	EventAware	2018	Events	ICH	Tags extraction, profile learning, Tag-based filtering	Hybrid	Mobile app
[9]	CHATBot	2020	Cultural sites	TCH	Text analysis	Hybrid	Mobile app
[16]	-		Painting and Sculptures	TCH	Social affinity graph	CF	
[14]	Ifar		Castles (The graz castle)	TCH	Serious games, AR, VR	CF	Mobile app
[24]	FIRSt (CHAT project)		Museum objects, movies, music, books,…etc.	TCH (Digital libraries)	Semantic analysis(Tags), ML	Profiling and CF	
[8]	-	2018	Cultural places	TCH	ML algorithms		Mobile app
[26]	Apollo		Museum	TCH	Minimax approach, ML, AI	Hybrid	
[12]	Père		Places & services	General	Semantic web technologies (RDF/OWL, SPARQL)		Web app
[10]	CHIS	2018	Cultural places	General	Big Data algorithms	CB	Mobile + web portal
[29]	-		Trips, attraction places		AI algorithms	Hybrid	Mobile app

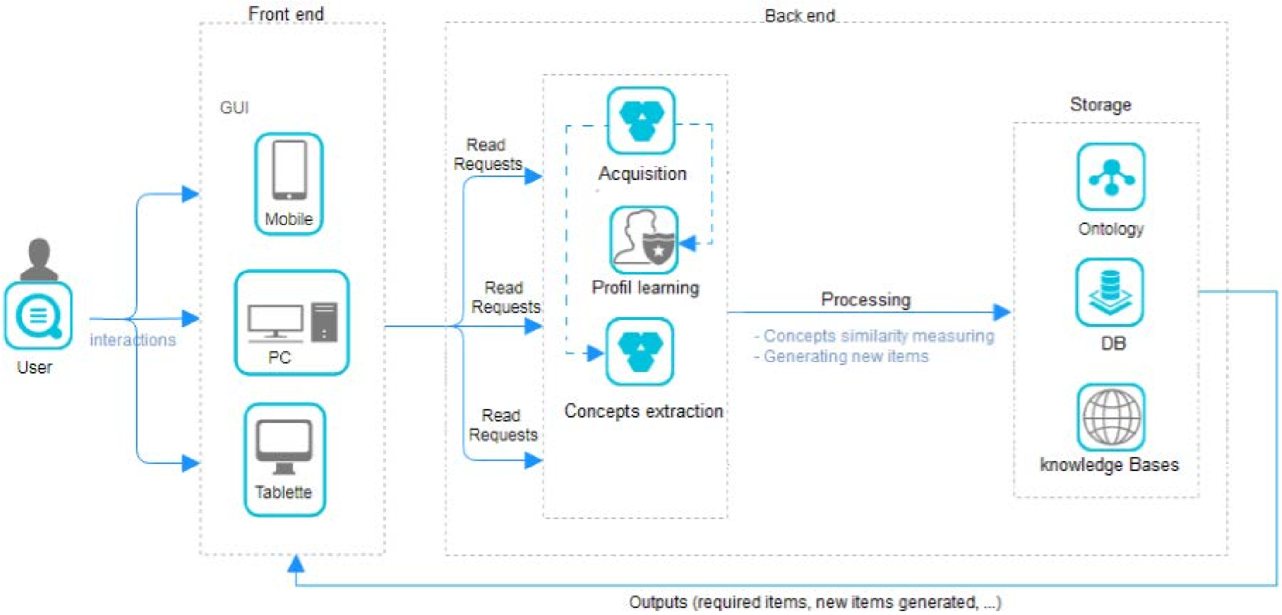
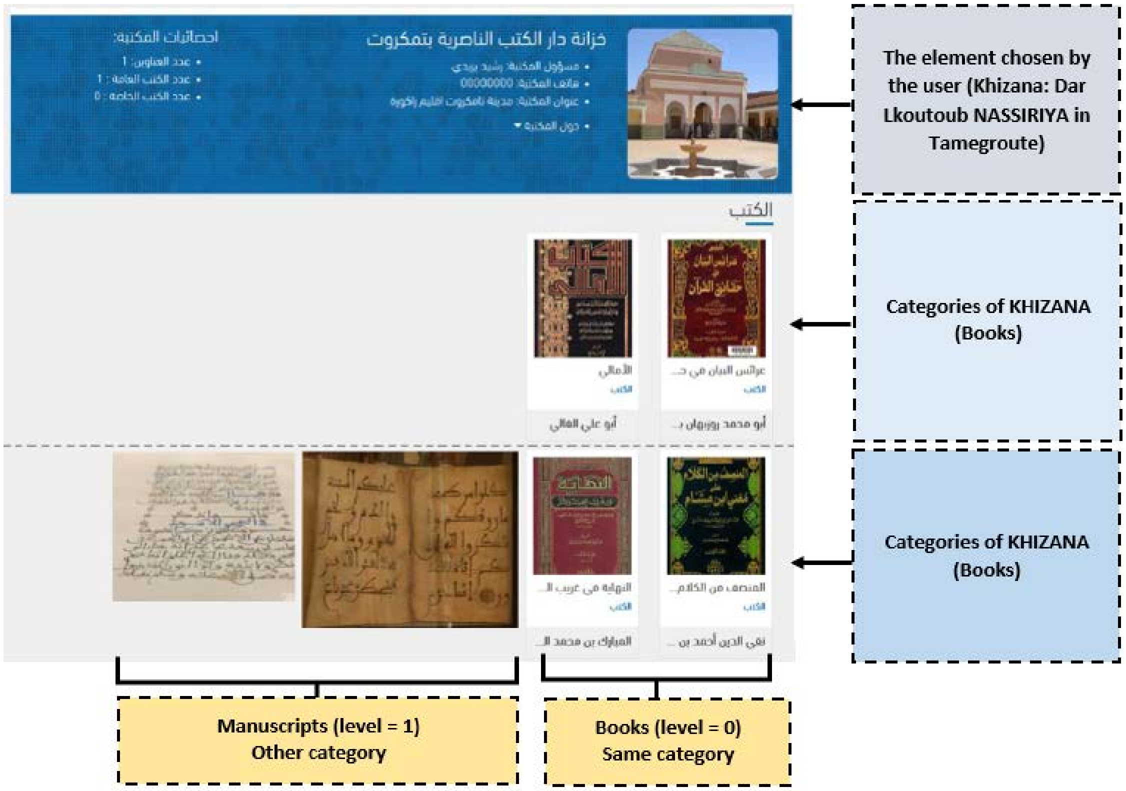


Fig. 4. The Architecture of the Proposed System.

Fig. 5 shows the results of integrating several semantic levels in the case of users interested in Books, resulting in a recommendation of objects with the same level of semantic depth (Books in this case) or higher levels (Manuscripts in this case) based on the semantic knowledge base already developed. Fig. 6 shows an example of a recommendation based on a user's interactions with objects of type: Khizana (Libraries), in this case, the Khizana Nassiriya, one of the world's oldest Khizana and houses to incomparable historical documents. In this case, the user will see recommended Khizana from the same category; it is the case where the semantic depth leads to objects from the same category.


خزانة دار الكتب الناصرية بتمكروت

- مسؤول المكتبة: رشيد بريدي
- هاتف المكتبة: 00000000
- عنوان المكتبة: مدينة تامكروت اقليم زاكورة
- دول المكتبة

احصائيات المكتبة:

- عدد العناوين: 1
- عدد الكتب العامة: 1
- عدد الكتب الخاصة: 0

الكتب

عرائس البيان في حـ...
الكتب
أبو محمد روزبهان بـ...

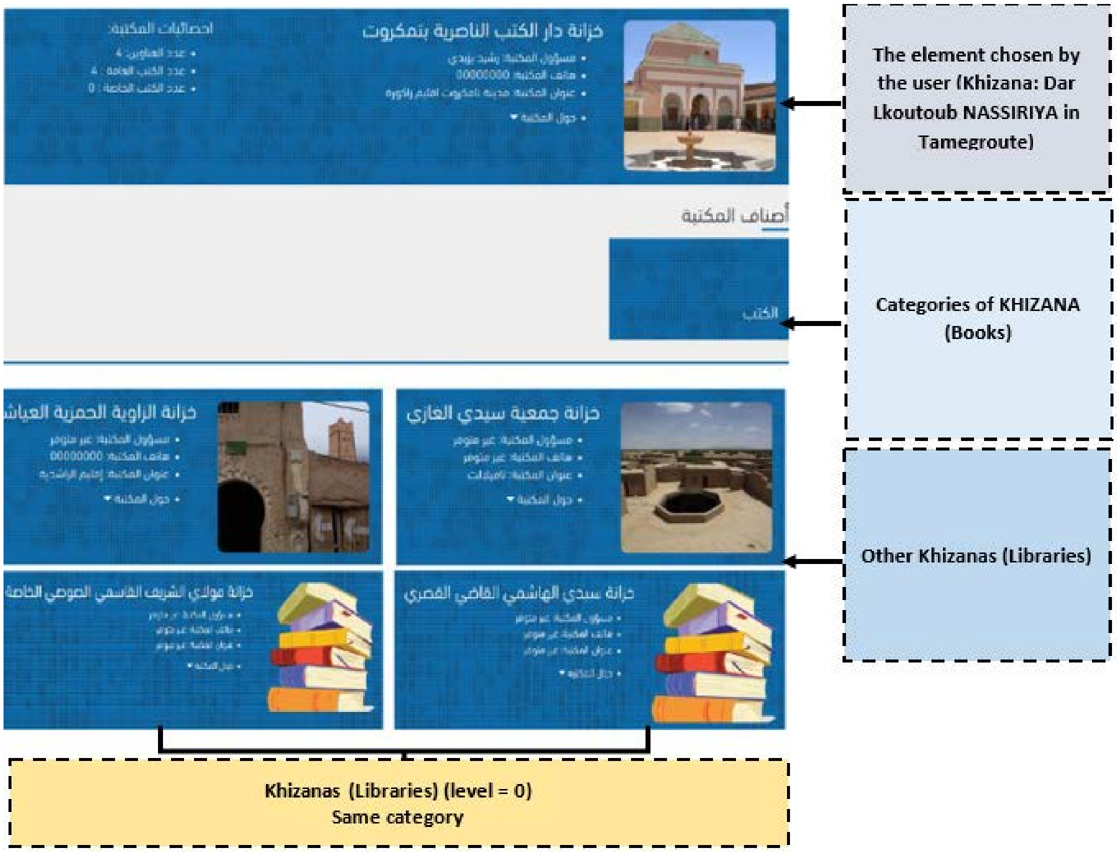
الأمالي
الكتب
أبو علي القالي

المنصف من الكلام...
الكتب
تقي الدين أحمد بن ...

النهاية في غريب الـ...
الكتب
المبارك بن محمد الـ...

The element chosen by the user (Khizana: Dar Lkoutoub NASSIRIYA in Tamegroute)
Categories of KHIZANA (Books)
Categories of KHIZANA (Books)
Manuscripts (level = 1)
Other category
Books (level = 0)
Same category

Fig. 5. Example of Results of the Approach Adopted for a Book Recommendation.



خزانة دار الكتب الناصرية بتمكروت

- مسؤول المكتبة: رشيد بويدي
- هاتف المكتبة: 00000000
- عنوان المكتبة: مدينة تامكروت اقليم زاكورة
- حول المكتبة ▾

احصائيات المكتبة:

- عدد العناوين: 4
- عدد الكتب العامة : 4
- عدد الكتب الخاصة : 0

أصناف المكتبة

الكتب

خزانة الزاوية الحمزية العياش

- مسؤول المكتبة: غير متوفر
- هاتف المكتبة: 00000000
- عنوان المكتبة: إقليم الراشدية
- حول المكتبة ▾

خزانة جمعية سيدي الغازي

- مسؤول المكتبة: غير متوفر
- هاتف المكتبة: غير متوفر
- عنوان المكتبة: تافيلالت
- حول المكتبة ▾

خزانة مولاي الشريف القاسمي الصوصي الخاصة

- مسؤول المكتبة: غير متوفر
- هاتف المكتبة: غير متوفر
- عنوان المكتبة: غير متوفر
- حول المكتبة ▾

خزانة سيدي الهاشمي القاضي القصري

- مسؤول المكتبة: غير متوفر
- هاتف المكتبة: غير متوفر
- عنوان المكتبة: غير متوفر
- حول المكتبة ▾

The element chosen by the user (Khizana: Dar Lkoutoub NASSIRIYA in Tamegroute)
Categories of KHIZANA (Books)
Other Khizanas (Libraries)
Khizanas (Libraries) (level = 0)
Same category

Fig. 6. Example of Results of the Approach Adopted for Khizana Recommendation.

V. Conclusion

Until now, very few works in the literature have focused on the valorization of the SCH of a region very rich in heritage data that has remained ignored. The added value of this work consists largely in the exploitation of some ML algorithms for the realization of an intelligent system allowing the recommendation of SCH objects. This will allow the preservation of this heritage and make it recognized by a maximum of actors. In this perspective, a Semantic RS for SCH has been implemented and tested for Drâa-Tafilalet region. The process of creating and assigning profiles considers end-user preferences. The process enables semi-supervised (content based) profiling, which is then refined using other available profiles (based on user collaboration (history of other users)). The system automatically adjusts the content to the user's terminal (mobile or web). The proposed system will enable the valorization of a little-known CH despite its significant potential in several fields of the economy of a region whose primary source of income is tourism. The majority of the functions described in this paper are already operational, and test data was used to validate the final prototype. The next step is to incorporate this component into the overall system in order to achieve the primary goal of this study: to have a comprehensive system for preserving and presenting the CH of the study region. Another point to raise, which will serve as the foundation for future work, is the incorporation of an interface that will allow cultural heritage specialists in the region to validate available data before it is explored via the platform.

References

[1] F. Nafis and D. Chiadmi: "Methods and Systems for the Linked Data". In Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015 (pp. 587-592). (2016) Springer, Cham.

[2] F. Nafis, A. Yahyaouy, and B. Aghoutane: "Semantic integration of Moroccan Cultural Heritage using CIDOC CRM: case of Drâa-Tafilalet zone", IJCC- ICBDSDE'19, 2021 "in Press".

[3] K. Al Fararni., F. Nafis, B. Aghoutane, A. Yahyaouy, J. Riffi and A. Sabri : "Hybrid recommender system for tourism based on big data and AI: A conceptual framework". Big Data Min. Anal. 4, 47–55, 2021.

[4] F. Amato, F. Moscato, V. Moscato, F. Pascale, and A. Picariello, « An agent-based approach for recommending cultural tours », Pattern Recognition Letters, vol. 131, p. 341 347, mars 2020, doi: 10.1016/j.patrec.2020.01.005.

[5] T.M.A. Basile, S. Ferilli, N. Di Mauro, F. Esposito: "Incremental induction of classification rules for cultural heritage documents". In: Proceedings of the 17th International Conference on Innovations in Applied Artificial Intelligence. pp. 915-923. IEA/AIE'2004, Springer Springer Verlag Inc (2004). doi:10.1007/b97304.

[6] D. Brodi´c and A. Amelio: "Discrimination of different Serbian pronunciations from shtokavian dialect". Procedia Computer Science 112, 1935 - 1944 (2017). doi : 10.1016/j.procs.2017.08.047, knowledge Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

[7] D. Brodi´c and A. Amelio: "Recognizing the orthography changes for identifying the temporal origin on the example of the balkan historical documents". Neural Computing and Applications (Nov 2017). doi:10.1007/s00521-017-3292-1.

[8] P.J.S Cardoso, P. Guerreiro, J. Monteiro, and J.M.F. Rodrigues: ‘Applying an Implicit Recommender System in the Preparation of Visits to Cultural Heritage Places’. In Universal Access in Human-Computer Interaction. Virtual, Augmented, and Intelligent Environments, M. Antona, and C. Stephanidis, eds. (Cham: Springer International Publishing), pp. 421–436, 2018.

[9] M. Casillo, F. Clarizia, G. D’Aniello, M. De Santo, M. Lombardi and D. Santaniello : “CHAT-Bot: A cultural heritage aware teller-bot for supporting touristic experiences”. Pattern Recognition Letters 131, 234–243, 2020.

[10] A. Castiglione, F. Colace, V. Moscato and F. Palmieri: “CHIS: A big data infrastructure to manage digital cultural items”. Future Generation Computer Systems 86, 1134–1145, 2018.

[11] D. Horowitz, D. Contreras, and M. Salamó : ‘EventAware: A mobile recommender system for events’. Pattern Recognition Letters 105, 121–134, 2018.

[12] G.-L. Chetreanu, A. E. Mihaila, , I. Vascu, G.-A. Vlad, S. C. Buraga and L. Alboaie: “PeRe: A location-based personal semantic web recommender”, in Proceedings of 2012 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, Romania, mai 2012, p. 486 490, doi: 10.1109/AQTR.2012.623776.

[13] F. Esposito, D. Malerba, G. Semeraro, S. Ferilli, O. Altamura, T.M.A Basile, M. Berardi, M. Ceci and N.D. Mauro: “Machine learning methods for automatically processing historical documents: from paper acquisition to xml transformation”. In: First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings. pp. 328-335 (Jan 2004). doi:10.1109/DIAL.2004.1263262.

[14] S. Gerhard, H. Agnesa and J. FH “iFAR: mobileAR for Cultural Heritage », présenté à XChange Reality”, St. Pölten, Austria, avr. 2020.

[15] E. Grilli, D. Dininno, G. Petrucci and F. Remondino: “From 2d to 3d supervised segmentation and classification for cultural heritage applications”. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2, 399-406 (2018). Doi : 10.5194/ isprs-archives-XLII-2-399-2018.

[16] M. Hong, J. Jung, F. Piccialli and A. Chianese: “Social recommendation service for cultural heritage”, Pers Ubiquit Comput, vol. 21, no 2, p. 191 201, avr. 2017, doi: 10.1007/s00779-016-0985-x.

[17] A. Kuusik, S. Roche and F. Weis: “SMARTMUSEUM: Cultural Content Recommendation System for Mobile Users”, in 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, Korea, 2009, p. 477- 482, doi: 10.1109/ICCIT.2009.257.

[18] J. Li, J. Ding, X. Yang: “The regional style classification of chinese folk songs based on gmmcrf model”. In: Proceedings of the 9th International Conference on Computer and Automation Engineering. pp. 66{72. ICCAE ’17, ACM, New York, NY, USA (2017). doi:10.1145/3057039.3057069.

[19] Y. Liu, Q. Xiang, Y. Wang, L. Cai: “Cultural style based music classification of audio signals”. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 57-60 (April 2009). doi:10.1109/ICASSP.2009.4959519.

[20] Llamas, J., M. Lerones, P., Medina, R., Zalama, E., Gmez-Garca-Bermejo, J.: Classification of architectural heritage images using deep learning techniques. Applied Sciences 7(10) (2017). doi:10.3390/app7100992.

[21] M. Makridis and P. Daras,: “Automatic classification of archaeological pottery sherds”. J. Comput. Cult. Herit. 5(4), 15:21 (Jan 2013). doi:10.1145/2399180.2399183.

[22] T. Mensink and van J. Gemert: “The rijksmuseum challenge: Museum-centered visual recognition”. In: Proceedings of International Conference on Multimedia Retrieval. pp. 451-454. ICMR ’14, ACM, New York, NY, USA (2014). doi:10.1145/2578726.2578791.

[23] E. Michon, M.Q. Pham, J. Crego, J. Senellart: “Neural network architectures for arabic dialect identification”. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018).

[24] C. Musto, F. Narducci, P. Lops, M. de Gemmis, and G. Semeraro: “Integrating a Content Based Recommender System into Digital Libraries for Cultural Heritage”, in Digital Libraries, vol. 91, M. Agosti, F. Esposito, et C. Thanos, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, p. 27 38.

[25] A.M. Obeso, M.S.G. Vazquez, A.A.R. Acosta, J. Benois-Pineau: “Connoisseur: Classification of styles of mexican architectural heritage with deep learning and visual attention prediction”. In: Proceedings of the 15th International Workshop on Content Based Multimedia Indexing. pp. 16:1 {16:7. CBMI ’17, ACM, New York, NY, USA (2017). Doi : 10.1145 / 3095713.3095730.

[26] G. Pavlidis : “Apollo - A Hybrid Recommender for Museums and Cultural Tourism”, in International Conference on Intelligent Systems (IS), Funchal - Madeira, Portugal, sept. 2018, p. 94 101, doi: 10.1109/IS.2018.8710494.

[27] G. Sansonetti, F. Gasparetti, A. Micarelli : “Cross-Domain Recommendation for Enhancing Cultural Heritage Experience”, in Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP’19 Adjunct, Larnaca, Cyprus, 2019, p. 413- 415, doi: 10.1145/3314183.3323869.

[28] G. Sansonetti, F. Gasparetti, A. Micarelli, F. Cena and C. Gena : “Enhancing cultural recommendations through social and linked open data”, User Model User-Adap Inter, vol. 29, no 1, p. 121 159, mars 2019, doi: 10.1007/s11257-019-09225-8.

[29] A. Smirnov, A. Kashevnik, N. Shilov, S. Mikhailov, O. Gusikhin, and H. Martinez : “Intelligent Content Management System for Tourism Smart Mobility: Approach and Cloud-based Android Application”, (SCITEPRESS - Science and Technology Publications), pp. 426–433, 2019.

[30] A.M Yasser, K. Clawson, C. Bowerman: “Saving cultural heritage with digital make-believe: Machine learning and digital techniques to the rescue”. In: Proceedings of the 31st British Computer Society Human Computer Interaction Conference. pp. 97:1{97:5. HCI ’17, BCS Learning & Development Ltd., Swindon, UK (2017). doi: 10.14236/ewic/HCI2017.9.

[31] M. H. Abed, M. Al-Asfoor, et Z. M. Hussain, ”Architectural Heritage Images Classification Using Deep Learning With CNN”, in Proceedings of the 2nd International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, Bari, Italy, janv. 2020, p. 12.

Comprehensive Survey and Research Directions on Blockchain IoT Access Control

Hafiz Adnan Hussain[1], Zulkefli Mansor[2], Zarina Shukur[3]
Faculty of Science and Information Technology
The National University of Malaysia
Bandar Baru Bangi, 43000
Selangor, Malaysia

***Abstract*—The Internet of Things (IoT) is a widely used technology in the last decade in different applications. The Internet of things is wirelessly or wired to communicate, store, compute and track various real-time scenarios. This survey mainly discussed the core problems of Internet of things security and access control to unauthorized users and security requirements for IoT. The Internet of things is a heterogeneous device and has low memory, less processing power because of the small sizes. Nowadays, IoT systems are not sure and powerless to protect themselves against cyber attacks. It is mainly due to inadequate space in IoT gadgets, immature standards, and the lack of protected hardware and software design, development, and deployment. To meet IoT requirements, the authors discussed the limitations of traditional access control. Then the authors examined the potential to spread access control by implementing the safe architecture accommodated by the Blockchain. The authors also addressed how to use the Blockchain to work with and resolve some of the standards relevant to IoT security issues. In the end, an analysis of this survey shows future, open-ended problems, and challenges. It offers how the Blockchain potentially ensures reliable, scalable, and more efficient security solutions for IoT and further research work.**

***Keywords*—Blockchain; Internet of Things; IoT; access control; access control management**

I. Introduction

Internet of Things (IoT) is an infrastructure of smart things that exchange information over the Internet. In a different world, the Internet of things is used to obtain data, and certain events are triggered. The linked IoT devices are expected to be 50 billion in late 2020, according to CISCO. As the entire world population grows, the advancement of IoT devices is rapidly evolving. There seem to be enormous data produced by IoT devices. The infinite interlinking between physical and virtual objects notably stimulates processing, storage, and data exchange on the IoT. It is a primary base for rendering services in critical areas such as hospitals, cars, bridges, schools, retail outlets, public constructions, communities, and even within human bodies in the form of wearable devices. The foremost issues are how can efficiently handle services and data collected by IoT devices because they can contain personal information or even save people's lives. There are many difficulties in implementing IoT. In numerous IoT implementations, uniformity, interoperability, data management, processing, user authentication, identity, confidentiality, completeness, affordability, protection, and privacy are among the several open challenges. One of the famous approaches for data protection and privacy is making sure only validated and accredited users can see that data [1]. Data is coming massively from these devices, storing data to make sure it is secured and protected from hackers, and authenticating the user who wants to access that data is the correct user on the server. The authenticated user must have limited access to limited data.

Access control is a collection of rules and policies that enable a nominated user to be enabled or not accessed by users, services, procedures, or other approved mechanisms to access or limit access to an information system's resources. It decides whether or not the requirement for access to the related order is allowed [2]. Various concepts, like access control policy, layout, and method, have established a basis for developing an access control system. The access control policies form the basis for access arrangements implemented through the mechanisms for access control. The essential parts of the access control system are the object, subject, and owner of the thing. The current element is a subject that requests access to objects. The object is a receptive element in a system that is petitioned by the subject. The connection to an object determines access to documents, areas, directories, programs, and network nodes to its information. The network is also recognized as objects-related devices, including smart devices, routers, and mechanical elements. The owner of each entity defines access policies and supplies the necessary authorizations.

In IoT access control systems, certain restrictions are imposed because they have a complex and clustered infrastructure that does not meet the wide variety of IoT gadgets and the versatility situation in which nodes can connect and drop the network. IoT devices often need a lightweight low-latency access control system for CPUs, memory, and battery life [3]. Numerous literature attempts to create a distributed IoT control monitor utilizing edge paradigms such as Adhoc Web Cloud and Fog Computing by reducing the mentioned problems. However, due to the absence of trusted entities working to provide arbitrary services, the security problem persists. For example, a group of vehicles can accommodate road security assistance to other transports under traffic networks.

Blockchain is a Peer-to-Peer (P2P) distributed and decentralized public ledger over the network [4, 5]. It is used to store transactions, events, and smart contracts. The smart contract includes a programable code that any customer can create and publish in the Blockchain as a transaction [6]. It has a specific number assigned by Blockchain. The contract is executed when the Blockchain user or smart contract calls it, and the smart contracts can communicate, among others. The first Blockchain scenario is a shared P2P digital currency. It is a service for high-security distributed applications with the introduction of a smart contract framework. Therefore, to make a distributed IoT access control, this paper can exploit the latest use of Blockchain as a distributed and decentralized infrastructure.

The rest of the paper is organized as in Section 2; related research work based on Blockchain in IoT access control like problems and limitations of traditional access models are identified. The security requirements for IoT are described in Section 3. Analysis and discussion of the entire survey are summarized in Section 4. Future research directions and open challenges are described in Section 5. The paper concludes with a conclusion of the work done in Section 6.

II. Related Research Work on Blockchain-based Access Control for IoT

A. Access Control and IoT

Reference [7] provides a full detailed review of various IoT access control solutions. The study shows that the new access control method employed in IoT and claims that Internet protocols widely used cannot be extended to compelled environments. Based on its comprehensive review of the literature, [7] Identify and Lines of 3 access control and decentralized authorization solutions: the IBM Adept (Autonomous Decentralized P2P Telemetry) [8] framework, DOAuth (Decentralized Open Authentication), and FairAccess [9, 10]. However, in these references, the author explained the OAuth-based access control solutions are the heavy mechanism for IoT situations because of their low processing overhead and communications. Lastly, IBM Adept offers a collaboration and file storage system to develop IoT apps without creating a process for access control. Further, [7] passed over the work has done in the IETF called CoAP Management Interface (CoMI) [11], and the result is done by the Open Mobile Alliance (OMA) called LWM2M [12] CoMI and LWM2M are primary mechanisms in a centralized environment for IoT device management.

B. Blockchain and IoT

This paper identified two main access control patterns using Blockchain, access control depending on the transaction and access control depending on the smart contract. Their pros and cons are summarized in Table I.

1) Access control depending on the transaction: Conoscenti et al. [13] performed systematic literature on the new technology Blockchain for the IoT. The study mentioned many articles that handle the data gathered from IoT devices. For example, [14] shows a method for verifying the status of the data, and [15] reports the procedure for maintaining the data holding of IoT gadgets. None of the mentioned research articles suggest an architecture. The administrators can control the entire IoT life cycle access policies rather than their roots or location. The best of experience, the old work relevant to Blockchain access control for IoT, is [9] that shows a cryptocurrency Blockchain access control structure called FairAccess. Nevertheless, Access Control Policies describe the creation of transactions for that Smart Contract by creating various Smart Contracts for the Access Control Policy of each source request pair that are not suitable for the IoT environment.

The authors in [16] suggested a new method of Authentication of access control and a user to make IoT secure and safe for illegal users and get open access to information. The proposed system is based on the following: i) Registration Authority (RA), and ii). Home Registration Authority (HRA). The RA was created to simplify the authentication system for IoT gadgets. Each device should be registered with the RA.

The authors in [17, 18] suggest a multi-tier Blockchain base system for sharing data between communities and individuals for users and IoT gadgets. The mentioned design has three key components: data management protocol, data storage mechanism, and message services. The data management system offers a structure for the data manager, data requestor, or correspondence channel. The messaging tool used in this context increases device scalability by publishing / subscribing patterns. In the end, Blockchain uses data storage systems to store data anonymously.

In [19], The author proposed the Blockchain base design to enforce the access control mechanism based on the attribute. The policies are formulated by XACML and processed in Blockchain as compressed transactions. The smart contract codes all the elements needed for policy assessment. Authors design storage and maintenance feature by utilizing innovative contracts. It reflects self-assessment procedures pursued straightforwardly and transparently when the user is requesting access. This method merges smart contract and transaction structures to create an access control system that allows users to understand the policies that affect their access requests. It offers a centralized audit system and identifies sections that are fraudulently modifying the rights given by executable policies. Proof of the concept (PoS) implemented in the above method by using the Ethereum Blockchain to show and verify the proposal's validity.

2) Access control depending on smart contract: In the Alphand, Amoretti [20] architecture based in Blockchain, the authors propose secure management of access resources from end to end, named IoT Chain. Resources servers of resource owners hold their resources through a proxy server in an encrypted and signed format [21]. The third party who requests access to protected resources is the customer. It calls for a critical server decryption key, which checks that a blockchain is allowed to contain. The approval process works in the following way: the resource owner establishes and publishes an intelligent contract for customers in the Blockchain. If certain conditions are met, the customer calls the relevant, intelligent contract to produce an access token. The permission tokens are not sent to the customer but stored in the permanent memory of the transaction. The Client must seek the appropriate key to decrypt resources of the critical server, a node Blockchain, and hold a duplicate file of the Blockchain ledger once a token has been stored in an internal Contract database. At this point, the critical server tests that the Client has a token and transfers the key using a DTLS mode on Blockchain's smart contract system. The Client then installs and decrypts the encrypted device from the proxy server—this approach is primarily designed to replace a trusted ACE request system with a secure Blockchain permit.

In [22], the authors: described the main issues in the IoT access management system. The first is due to the core architecture, and the second is because access policies are handled dynamically. If the data is satisfied by the requester's access policies, the control contract is executed automatically, and a token of authorization is created and allocated to the requester. The input information relates to trust and reputation safety parameters to assists the resource owner in dynamically develop or change security policies. Order in verifying its validity; however, the model proposed requires evidence of description.

In [23] proposed the machine learning algorithms, and Smart contract access management checks the efficiency of multiple user access to a shared resource by maintaining sophisticated access control. The network architecture is made up of a single Judge Contract (JC), multiple Access Contracts (ACCs), and one Register Agreement (RC). Per ACC describes a subject-resource pair access control system and applies to update access control rules. A record of misconduct is maintained in the ACC smart contract for each property. It describes the actions of the subject matter of this platform, with several requests being identified in a short time and the decision of the Judge Contract (JC) penalty. The ACC is performed once an individual has been appointed to obtain access, and the ACC reports to the JC contract if the misconduct is detected. Based on an incorrect evaluation method, the contract for JC shows the appropriate penalty as a temporary blocking of subject access.

The author in [24] performed an access control and smart contract verification to tackle an IoT device security scalability problem. In reality, one or more IoT devices are operated by a customer, and each device needs its credentials. In this instance, however, the user should authenticate independently on each unit. This approach results in overhead verification and is challenging to measure. The main reason for using a smart contract is the validation of the user and IoT. The user signs up for the smart contract, verifying device identity by using the Ethereum wallet address. It is performed and examined in the blockchain environment of Ethereum. [25] introduces the Blockchain-enabled fog nodes for user authentication and authorization. The fog nodes manage and validating the authenticity of access to IoT devices on the Ethereum network interface. The system manager uses a smart contract to map all of the registered fog nodes and their associated IoT devices. Besides, it consists of the collection and permissions for entry to registered users. The arrangement includes the functionalities of registration, Authentication, and access control IoT device user link with a contract to verify the validity of the user. A token with access parameters is created if Authentication is successful. The next step is to sign the token and send it to the fog node to monitor the resource. The signature and token specifications are verified, and user access to the IoT gadgets is then granted or denied. A safe SSL connection between the user and the IoT device will be established for data exchange.

TABLE I. PROS AND CONS OF REFERRED ACCESS CONTROL SOLUTIONS

Ref.	Pros	Cons	Security Measures	Implemented
[7]	The authors highlighted how each solution produced various security specifications. They declared that centralized and distributed methods could complement each other.	Access Control Policies describe the creation of transactions for that Smart Contract by creating various Smart Contract to assign different Access Control Policy of each resource-request pair that are not suitable for the IoT environment.	VL	NA
[16]	IoT safe and secure for unauthorized users and open access This approach is secure for a man-in-the-middle attack	Critical scalability: The need for every device to have a RA and, similarly, for every user to have HRA could be a constraint for scalability	M	NO
[17]	Keep data privately Decentralized, open, and accessible data collected from data storage and architecture elements	Simple to do, but not feasible in all situations, as a large amount of computer power is needed for every node. It sends to the "Server," which decrypts an encrypted version of the info.	H	YES
[19]	Self-assessment policies continued in a straightforward and transparent way This model connects transactions and smart contract structures to create an access control system.	It offers a centralized audit system and identifies sections that are fraudulently modifying the rights granted by enforceable policies	VL	NO
[20]	Blockchain-focused IoTChain with ACE and OSCAR (IoT Object Security Architecture) authority. The suspect method to handle approval when OSCAR uses the public registry to create multidisciplinary groups for authorized customers.	Difficult to preserving the availability of the IoT	M	YES
[23]	To carry out centralized, secure access management for IoT networks, the author proposed a smart contract-based architecture composed of different access control contracts, an authority contract, and a registered agreement.	A large amount of contract requirements for a massive crowd is a daunting task.	M	YES
[24]	In a smart contract, users are authenticated, and an IoT token is issued. The contract decides whether the user will access the services and transmits tokens to the consumer and the required IoT computer.	For large IoT networks, this method suffers from the scalability problems associated with Blockchain. Ethereum smart contracts have the biggest drawback of fluctuating Ethereum rates, which is a problem for the consumers.	H	YES

III. Security Requirements for IoT

Various mechanisms and parameters must be considered, as listed below, for a secure IoT deployment.

A. Data Integrity, Privacy, and Confidentiality

When IoT data move across multiple hops throughout the network, a conventional encryption process is needed to guarantee data confidentiality [26]. Since systems, frameworks, and networks are configured differently, data held on a device are susceptible to protection and privacy infringement by disturbing live IoT network nodes. Attack vulnerable IoT devices may enable an intruder to contact data integrity by malicious data handling.

B. Accounting, Authentication, and Authorization

Authentication is needed for two parties to interact with each other to secure communication in IoT. Applications must be encrypted for exclusive access to services. The variation of IoT authentication procedure lives primarily because of different heterogeneous architectures and environments that help IoT gadgets. Such conditions pose a complexity in defining the standard global IoT authentication protocol [27]. Same as the authorization mechanisms guarantee that authorized persons have access to the systems or information. Usual authorization and authentication results are implemented in a stable environment that guarantees a protected communication environment. Moreover, accounting for the use of resources, reporting, and auditing produces a secure network security system.

C. Available for Services

Attacks on IoT devices will avoid general denial of service attacks involving utilities. Different tactics have led to IoT's consumers, including sinkhole assaults, jamming rivals, or replay attacks use IoT components on various levels to deteriorate the level of service (QoS) [28, 29].

D. Trustworthy

To maintain the end-to-end integrity of data gathered and related communications, the IoT applications require trust mechanisms that cover these scales. In addition to the capacity to evaluate these processes and interactions, the transparency of data collection systems and relevant experiences are the key to satisfying these criteria [30]. Both the requirements of clarity and auditing drive Blockchain to create trust in IoT.

E. Energy Efficiency and Cost-Effective

IoT gadgets are generally restricted to resources and have a lower capacity to store data. The author in [31] attacks on IoT systems may result in improved electricity consumption by intravenous or false service inquiries and by exhausting IoT resources.

F. Single Points of Server Failure

A significant number of single points of vulnerability That may depreciate service provisioned by IoT may be exposed by the ongoing development of heterogeneous IoT connectivity networks [32]. It includes designing a strategic framework for a broad category of IoT gadgets and implementing new methods for utilizing a network of fault tolerance.

IV. Analysis and Discussion

Based on the literature analysis and survey, Blockchain technology may be seen as a new bearing in IoT access control. The incentive to use the Blockchain is to help its stable and secure distributed nature that solves many IoT access requirements. In this survey, the authors also defined two ways of access control for Blockchain in IoT perspectives.

The first one consists of the transaction system to request, receive, assign, and revoke connections. In essence, the transaction is used to make a connection between the asset owner and the subject. Connection decisions may be made directly by the owner of the asset. If this is not the case, the access request shall be transferred to the external entity responsible for assessing the appeal, making the decision, and returning it to the asset owner, as stated in [33, 34]. In this case, Blockchain's primary aim is to securely transfer the access token by defining individuals' access rights and guaranteeing to check and trace all access transactions. The power delegate is often an essential method in the collaboration framework, which can be delegated to the new topic from the current issue in a verifiable manner depending on the transaction. It implies the freedom of a subject to pass partly or entirely the right to access another individual. The delegated receiver is then allowed to carry out the delegating customer's activities. To restore the transaction-based access regulation, unified access token management can fix the dual cost issue and guarantee the trackability of all transactions. However, The recognition of an entry is rendered by a single person who may be the property owner or another agency identified by the Access Control Model application. The model can be called hybrid and not distributed.

The second approach evaluates a user control demand using a smart contract definition. It takes an access option depending upon the rules defined by the property owner and applied in the agreement. The contract is executed until the customer requirement is met with the access agreement, and the effect is a consumer consent authority token. Ultimately, the token is sent to the permission applicant utilizing a particular operation. The principal objective of this strategy is to return a single permission server with a distributed smart contract to construct a distributed permission network [20]. All-access control functions may be executed on the authenticated and recorded contract in the registry of Blockchain. Blockchain nodes can establish a mutual copy and carry out a contract without a mediator. Distributed smart contract ownership per the delivery and implementation by Blockchain will solve a single point of server failure in a centralized access control manner.

In comparison, the corresponding studies investigate the feasibility of producing dynamic access control evaluation by smart contracts that incorporate a machine-readable algorithm to find and detect behavioral subjects [22]. However, there is no approved smart contract access control paradigm validated and examined in specific application domains such as Intelligent Transport System or Smart Cities to prove its viability. Most of the proposed approaches show that they are applied using one of the Blockchain platforms, such as Bitcoin and Ethereumare, not evaluated in real environments.

For future access control, the Blockchain would be the essential engine. It has new dynamics and technologies that solve big systems problems. This is only the start, and approaches to measure their success should be tested. The goal is to build a transparent access control platform built on Blockchain, which will enable the future generation of the distributed network.

V. Future Research Directions and Open Challenges

This segment explains the proposed issues for the efficient implementation of IoT security.

A. Limitations of Resources

IoT's resource-restricted nature had been a significant obstacle to identifying a reliable security mechanism. Cryptographic algorithms can only operate under these powers, unlike normal ones. By [35] ensure efficient implementation of IoT security and communication protocols over the network, any communications or multicasts needed to transfer the key or certificate, storage and resources must be dealt with it. It means that these protocols are optimized to be lightweight and energy-efficient, given the need for sophisticated computation and advanced energy harvesting techniques.

B. Heterogeneous Devices

A multi-layer security structure must be discussed, as with heterogeneous sensors, exceptionally compact, high-end servers, and low-power sensor systems. The structure will first adjust to new resources and selections on the collection of IoT layer security mechanisms even before services are granted to end-users [36]. Such a flexible and compact structure requires information that is dependent on the uniformity of IoT architectural tools.

C. Single Points of Server Failure

For heterogeneous systems, structures, and protocols, the IoT standard is risky to single-point-of-server-failure than any other system. There needs to be more research work required to ensure appropriate IoT elements, especially in mission-critical applications. It would need devices and guidelines to perform continuity in mind the trade-off among values and the functionality of the entire infrastructure.

D. Interoperability of Security Protocols

Protocols intended at various layers require interoperation within the requirement of translation mechanisms to regulate the global security structure for IoT [37]. The active synthesis of safety measures on each layer can then be determined in the worldwide system framework, taking into account architectural limitations.

E. Trusted Updates and Management

Scalable and reliable software management and upgrades to millions of IoT gadgets lead to open issues for future research. Besides, problems associated with the safe and trusted IoT device supply chain, ownership, and data privacy are the main research concerns that need to be tackled by the researchers to raise significant and broad IoT acceptance [38]. These IoT security solutions are also available in Blockchain technology. However, blockchain technology itself claims to face problems in scalability, accuracy, arbitration/regulation, and essential collisions.

F. Hardware/Firmware Culnerabilities

The IoT structure becomes vulnerable to hardware flaws when the low-cost and low-performance system is traditional. It is not just a physical malfunction, but it must be verified before IoT deployment to implement security algorithms in hardware, routing, and packet processing operations [39]. Some Security flaws exposed since launch have displayed challenging to identify and mitigate. The regular confirmation protocol is, consequently, a requirement for the use of IoT security.

G. Blockchain Vulnerabilities

Despite affording robust solutions to IoT security, blockchain systems are also exposed [40]. The consensus mechanism based on the hacking capacity can be violated to enable the attacker to handle the private database keys with minimal randomness that may also be used for blockchain accounts negotiation [41, 42]. There is still a need to develop efficient mechanisms to protect transactions' privacy and prevent race attacks, resulting in duplication of transaction costs.

VI. Conclusion

In this survey, the authors first study the various security problems and challenges in IoT applications and access control for authorized and unauthorized users. Secondly, the authors have deep dive into previous research to figure out their solutions and existing problems. From the survey, it was found that some of the research has been already done in IoT access control by using Blockchain, and found that IoT systems are vulnerable and powerless to defend themselves. Due to insufficient resources in IoT gadgets, immature standards, and the lack of reliable software and hardware development, design, and deployment, as well as trusted updates and managements. The authors have acknowledged the restrictions of old access control to reply to IoT demands and investigated the ability to use the secure Blockchain system to manage access control. The authors demonstrate how the Blockchain can address and resolve some of the fundamental IoT security problems. The article also explains and recognizes future issues and challenges that the researchers need to provide reliable, effective, and scalable IoT security solutions.

Acknowledgment

This work was supported by The National University of Malaysia under the Research Excellence Consortium Fund (Grant number KKP/2020/UKM-UKM/4/3). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] A Hassen, O., A Abdulhussein, A., M Darwish, S., Othman, Z. A., Tiun, S., & A Lotfy, Y. (2020). Towards a Secure Signature Scheme Based on Multimodal Biometric Technology: Application for IoT Blockchain Network. Symmetry, 12(10), 1699.

[2] Alliance, O. M. (2017). Lightweight machine to machine technical specification. Approved Version, 1(1).

[3] Almadhoun, R., Kadadha, M., Alhemeiri, M., Alshehhi, M., & Salah, K. (2018). A user authentication scheme of IoT devices using blockchain-

enabled fog nodes. Paper presented at the 2018 IEEE/ACS 15th international conference on computer systems and applications (AICCSA).

[4] Alphand, O., Amoretti, M., Claeys, T., Dall'Asta, S., Duda, A., Ferrari, G., . . . Zanichelli, F. (2018). IoTChain: A blockchain security architecture for the Internet of Things. Paper presented at the 2018 IEEE wireless communications and networking conference (WCNC).

[5] Aman, A. H. M., Yadegaridehkordi, E., Attarbashi, Z. S., Hassan, R., & Park, Y.-J. (2020). A survey on trend and classification of Internet of things reviews. Ieee Access, 8, 111763-111782.

[6] Atlam, H. F., Alenezi, A., Alassafi, M. O., & Wills, G. (2018). Blockchain with Internet of things: Benefits, challenges, and future directions. International Journal of Intelligent Systems and Applications, 10(6), 40-48.

[7] Ching, T. W., Aman, A. H. M., Azamuddin, W. M. H., Sallehuddin, H., & Attarbashi, Z. S. (2021). Performance Analysis of Internet of Things Routing Protocol for Low Power and Lossy Networks (RPL): Energy, Overhead and Packet Delivery. Paper presented at the 2021 3rd International Cyber Resilience Conference (CRC).

[8] Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the Internet of things. Ieee Access, 4, 2292-2303.

[9] Cohn, J. M., Finn, P. G., Nair, S. P., Panikkar, S. B., & Pureswaran, V. S. (2019). Autonomous decentralized peer-to-peer telemetry. In: Google Patents.

[10] Conoscenti, M., Vetro, A., & De Martin, J. C. (2016). Blockchain for the Internet of Things: A systematic literature review. Paper presented at the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA).

[11] Dai, H.-N., Zheng, Z., & Zhang, Y. (2019). Blockchain for Internet of Things: A survey. IEEE Internet of Things Journal, 6(5), 8076-8094.

[12] Dedeoglu, V., Jurdak, R., Putra, G. D., Dorri, A., & Kanhere, S. S. (2019). A trust architecture for Blockchain in IoT. Paper presented at the Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.

[13] Di Pietro, R., Salleras, X., Signorini, M., & Waisbard, E. (2018). A blockchain-based Trust System for the Internet of Things. Paper presented at the Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies.

[14] Din, Z., Jambari, D. I., Yusof, M. M., & Yahaya, J. (2019). Challenges in Managing Information Systems Security for Internet of Things-enabled Smart Cities. Paper presented at the 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS).

[15] Hashemi, S. H., Faghri, F., Rausch, P., & Campbell, R. H. (2016). World of empowered IoT users. Paper presented at the 2016 IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI).

[16] Hernan, S., Lambert, S., Ostwald, T., & Shostack, A. Uncover security design flaws using the STRIDE approach (2006). URL http://msdn. microsoft. com/en-gb/magazine/cc163519. aspx, 15.

[17] Jafar, U., & Ab Aziz, M. J. (2020). A State of the Art Survey and Research Directions on Blockchain Based Electronic Voting System. Paper presented at the International Conference on Advances in Cyber Security.

[18] Jalal, I., Shukur, Z., & Bakar, K. A. A. (2020). A Study on Public Blockchain Consensus Algorithms: A Systematic Literature Review.

[19] Kamalinejad, P., Mahapatra, C., Sheng, Z., Mirabbasi, S., Leung, V. C., & Guan, Y. L. (2015). Wireless energy harvesting for the Internet of Things. IEEE Communications Magazine, 53(6), 102-108.

[20] Khan, M. A., & Salah, K. (2018). IoT security: Review, blockchain solutions, and open challenges. Future Generation Computer Systems, 82, 395-411.

[21] Lee, B., & Lee, J.-H. (2017). Blockchain-based secure firmware update for embedded devices in an Internet of Things environment. The Journal of Supercomputing, 73(3), 1152-1167.

[22] Li, X., Jiang, P., Chen, T., Luo, X., & Wen, Q. (2020). A survey on the security of blockchain systems. Future Generation Computer Systems, 107, 841-853.

[23] Li, Z., Kang, J., Yu, R., Ye, D., Deng, Q., & Zhang, Y. (2017). Consortium blockchain for secure energy trading in industrial Internet of things. IEEE transactions on industrial informatics, 14(8), 3690-3700.

[24] Liu, B., Yu, X. L., Chen, S., Xu, X., & Zhu, L. (2017). Blockchain based data integrity service framework for IoT data. Paper presented at the 2017 IEEE International Conference on Web Services (ICWS).

[25] Maesa, D. D. F., Mori, P., & Ricci, L. (2017). Blockchain based access control. Paper presented at the IFIP international conference on distributed applications and interoperable systems.

[26] Minoli, D., & Occhiogrosso, B. (2018). Blockchain mechanisms for IoT security. Internet of Things, 1, 1-13.

[27] Mohanta, B. K., Jena, D., Satapathy, U., & Patnaik, S. (2020). Survey on IoT security: challenges and solution using machine learning, artificial intelligence and blockchain technology. Internet of Things, 100227.

[28] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system," http://bitcoin. org/bitcoin. pdf.

[29] Ouaddah, A., Abou Elkalam, A., & Ait Ouahman, A. (2016). FairAccess: a new Blockchain‐based access control framework for the Internet of Things. Security and Communication Networks, 9(18), 5943-5964.

[30] Ouaddah, A., Abou Elkalam, A., & Ouahman, A. A. (2017). Towards a novel privacy-preserving access control model based on blockchain technology in IoT. In Europe and MENA cooperation advances in information and communication technologies (pp. 523-533): Springer.

[31] Ouaddah, A., Mousannif, H., Abou Elkalam, A., & Ouahman, A. A. (2017). Access control in the Internet of Things: Big challenges and new opportunities. Computer networks, 112, 237-262.

[32] Ourad, A. Z., Belgacem, B., & Salah, K. (2018). Using Blockchain for IOT access control and authentication management. Paper presented at the International Conference on Internet of Things.

[33] Outchakoucht, A., Hamza, E., & Leroy, J. P. (2017). Dynamic access control policy based on Blockchain and machine learning for the Internet of things. Int. J. Adv. Comput. Sci. Appl, 8(7), 417-424.

[34] Putra, G. D., Dedeoglu, V., Kanhere, S. S., & Jurdak, R. (2020). Trust management in decentralized iot access control system. Paper presented at the 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC).

[35] Sahraoui, S., & Bilami, A. (2014). Compressed and distributed host identity protocol for end-to-end security in the IoT. Paper presented at the 2014 International Conference on Next Generation Networks and Services (NGNS).

[36] Tseng, L., Wong, L., Otoum, S., Aloqaily, M., & Othman, J. B. (2020). Blockchain for managing heterogeneous Internet of things: A perspective architecture. IEEE Network, 34(1), 16-23.

[37] Van der Stok, P., & Greevenbosch, B. (2014). CoAP management interfaces (draft-vanderstok-core-comi-04). IETF, available at: https://datatracker. ietf. org/doc/draft-vanderstok-core-comi.

[38] Wilson, D., & Ateniese, G. (2015). From pretty good to great: Enhancing PGP using bitcoin and the Blockchain. Paper presented at the International conference on network and system security.

[39] Xu, R., Chen, Y., Blasch, E., & Chen, G. (2018). Blendcac: A smart contract enabled decentralized capability-based access control mechanism for the iot. Computers, 7(3), 39.

[40] Zhang, Y., Kasahara, S., Shen, Y., Jiang, X., & Wan, J. (2018). Smart contract-based access control for the Internet of things. IEEE Internet of Things Journal, 6(2), 1594-1605.

[41] Zhang, Y., & Wu, X. (2016). Access control in Internet of things: A survey. arXiv preprint arXiv:1610.01065.

[42] Zyskind, G., & Nathan, O. (2015). Decentralizing privacy: Using Blockchain to protect personal data. Paper presented at the 2015 IEEE Security and Privacy Workshops.

Improving Performance of ABAC Security Policies Validation using a Novel Clustering Approach

K.Vijayalakshmi[1]
Vels Institute of Science
Technology and Advanced Studies
Chennai, India

Dr.V.Jayalakshmi[2]
School of Computing Sciences
Vels Institute of Science, Technology and Advanced
Studies, Chennai, India

***Abstract*—Cloud computing offers several services, such as storage, software, networking, and other computing services. Cloud storage is a boon for big data and big data owners. Although big data owners can easily avail cloud storage without spending much on infrastructure and software to manage their data, security is a big issue, and protecting the outsourced big data is challenging and ongoing research. Cloud service providers use the attribute-based access control model to detect malicious intruders and address the security requirements of today's new computing technologies. Anomalies in security policies are removed to improve the efficiency of the access control model. This paper implements a novel clustering approach to cluster security policies. Our proposed approach uses a rule-specific cluster merging technique that compares the rule with the clusters where the probability of similarity is high. Hence this technique reduces the cost, time, and complexity of clustering. Rather than verifying all rules, detecting and removing anomalies in every cluster of rules improve the performance of the intrusion detection system. Our novel clustering approach is useful for the researchers and practitioners in the ABAC policy validation.**

Keywords—Anomalies; attribute-based access control model; big data; cloud storage; clustering; intrusion detection system; security policy

I. Introduction

Cloud storage is one of the most beneficial services to leverage and manage big data efficiently [1]. Large-scale enterprises, governments, and commercial organizations use the cloud to store and manage their big data without spending much on implementing infrastructure. Data privacy and security are essential, and securing the shared big data is a real challenge in today's emerging computing technologies [2] [3]. We have surveyed about challenges, security issues, and existing methodologies for addressing the security requirements [4]. Data breaches, data loss, account hijacking, denial of services, and malicious insiders are some attacks, and various encryption techniques are used to protect the data in the distributed cloud storage [5] [6]. Cloud service providers use various access control models to implement the Intrusion Detection System [7] [8]. The access control model is a function that identifies whether a requested operation on a shared object(resource) is legal or not [9]. In other words, the access control model is a protection technique that categorizes the authorized and unauthorized users and protects the shared resources based on the Access Control List (ACL) or security policies. The access control models use rules to determine which user can get what types of accesses for a shared resource. It manages all access-rights and access-conflicts over the shared resources [10]. The term object refers to the shared resources in a distributed environment. The access control models use the term subject to refer to the process being executed for a single user or an organization, which requests access for the object. Fig. 1 shows the working mechanism of the access control model. Many access control models are developed, and each is good in some situations and gives performance up to their level [11]. In this paper we described the major four access control models Discretionary Access Control (DAC), Mandatory Access Control (MAC), Role Based Access Control (RBAC), and Attribute Based Access Control (ABAC) models.

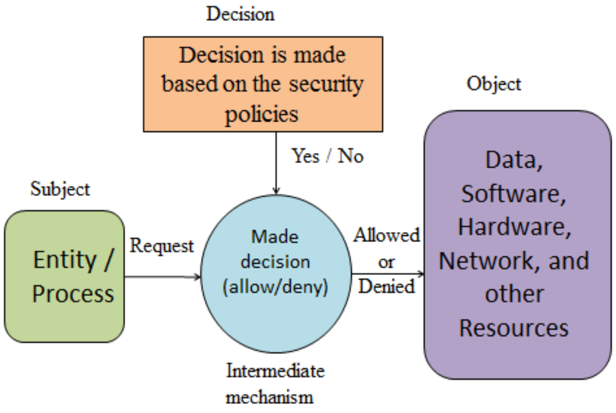

Decision
Decision is made based on the security policies
Yes / No
Object
Data, Software, Hardware, Network, and other Resources
Subject
Entity / Process
Request
Made decision (allow/deny)
Allowed or Denied
Intermediate mechanism

Fig. 1 A Working Mechanism of Access Control Model.

Discretionary Access Control (DAC) allows the owner of the object to specify the access rights of the user for their object. Thus the resource-owner decides and determines the security policies directly and explicitly unlike non-discretionary access control models. In non-discretionary access control models, rather than the resource-owner, the administrator determines the security policies [12]. DAC creates and maintains ACL for every individual resource to determine the access privileges of the user for that resource. ACL of each resource is a table of records and each record specifies the information about the user(or group of users) and the access rights [2]. The process of specifying ACL in DAC is very flexible, easily updatable, and reduces the administrator's work but the level of security is less[6]. The operating systems WINDOWS, LINUX use DAC. DAC meets the security requirements when the number of resources and users are limited but it fails to address the security requirements of today's computing technologies. The delegation feature of DAC has even more benefits like offering cooperation between users and the resource-owner [8]. But it leads to loss of confidentiality, integrity, and availability of resources if any user made a wrong decision while granting access to other users. As DAC requires a resource-owner for each object to determine the access rights, fulfilling this requirement is difficult in new computing technologies like cloud and fog computing [9][10].

Mandatory Access Control (MAC) is a central authority system where the administrator or the authorized person only can set access rights. The access rights cannot be specified or updated by the user like DAC. MAC uses a security label for each resource and subject. There are two pieces of information in a security label. The first piece of information called classification-segment determines the nature (public or private) of the resource or subject. The second segment category specifies the information about the process or application. MAC allows the requested operation on the object if the security labels of the object and subject match. The administrator only can determine the characteristics or attributes of the subject and object. Even the allowed users cannot specify or change the attributes. While DAC fails to give security on sensitive and confidential data, MAC meets these security requirements of the business and government organization. MAC is capable of managing trojan horses to protect the shared resources. Unlike DAC, the integrity and confidentiality of the resource are preserved by MAC. The feature reference monitor of MAC stores all security labels and policies in the database and monitors the access rights of every request for the object. The operating system Security-Enhanced Linux implements MAC to protect the objects. If the communication between the subject and object is high, then managing security labels and policies is a challenging issue. DAC and MAC are efficient when the size of data and the number of users are small [11]. MAC increases the complexity in managing security policies for large applications [2].

RBAC introduces a new concept of assigning roles between users and resources. RBAC (Role-Based Access Control) uses the permission-role and role-subject association to protect shared resources, whereas ABAC(Attribute-Based Access Control) uses security policies [13]. RBAC meets the security requirements of large applications. RBAC determines the security policies or access rights based on the role or job of the subject. Thus this access control model specifies the possible access rights (permissions) to the role. The subject can get access to an object based on his role or job. RBAC performs two assignments thus it first assigns all permissions required for a particular job or role for an object and second it assigns a role to the user or subjects. RBAC provides two efficient features least privilege and separation. The feature least privilege means RBAC assigns access rights or permissions required only for the role or task being processed on the object. As no additional access rights can be availed by the role, it prevents unauthorized access and preserves the integrity of the data. Another feature separation of duties distributes the process into various roles or duties and assigns the duty to the subject. As the subject can perform his/its assigned role only, RBAC prevents the entry of fraud intruders [14]. Hence no subject can perform all permissions or access rights of the shared object rather than the central authority (administrator). It increases the security level and RBAC has the capacity of ensuring whether the end-users can do their permitted transactions. Using the feature of grouping rules reduces the complexity of the administrator's work. Government sectors, military, and private organizations use RBAC to protect databases, web services, and all shared resources [15].

As RBAC performs constant or permanent associations (permission-role and role-subject), it fails to address the security needs of the computing technology that requires the dynamic associations and role-independent security policies. ABAC addresses this issue and is capable of performing dynamic relationships and generating role-independent security policies. ABAC security policies are generated with the attributes of the categories subject (who requests the operation of a resource), object (resource is shared in a distributed environment), and environment (time, the importance of the request, etc.). Both RBAC and ABAC address the security needs of large-scale applications or organizations, but ABAC fulfills the complex security requirements of today's computing technologies [9]. It gives better security even if the communication between the subjects and objects is increasing exponentially [16]. It has many features fine and coarse-grained access, dynamic mapping (subject-object), and flexibility. With the properties efficiency, flexibility, granularity, and security-level [17]. We gave a summarized analysis of the above-described access control models in Table I.

TABLE I. ANALYSIS OF ACCESS CONTROL MODELS

Access control model	Granularity (accuracy level of the model)	Flexibility (how the security policies are generated, expressed, and updated)	Efficiency (how quickly and correctly a decision is performed on the subject's request)	Security level (how the shared resources are protected)
DAC	Good at small scale applications	Good	Poor	Low
MAC	Good at small scale applications	Good	Poor	Better than DAC
RBAC	Good at large scale applications	Good	Good	Good
ABAC	Good at today's computing technologies and big data	Good	Good	Good

Most of today's computing technologies cloud, fog, edge, and IoT use ABAC to meet the security requirements. But the anomalies in the ABAC security policies dilute the reliability and efficiency of the mechanism [18]. Detection and elimination of all possible anomalies in security policies or rules improve the efficiency and accuracy of security. We can improve the performance of the detection mechanism by detecting anomalies in every cluster of similar rules, instead of detecting in every rule. This paper implemented a new approach to cluster similar ABAC security rules. The scopes of our approach are:

- Tool to generate ABAC Policies.
- Ability to measure the similarity of rules.
- Reduces the generation of more clusters.
- Avoids avoidable-redundancy (the same rule is contained in many clusters).
- Avoids the conflict clusters (not all rules in clusters are similar).
- Less time-complexity.
- Ease of implementation.

We use Section II to describe some clustering approaches already used in various research works, and Section III presents the fundamental concepts of the ABAC model and our proposed clustering approach. We use Section IV to describe the system architecture of our implementation. We analyzed output and discussion about the outcome of our approach in Section V. Finally, we concluded in Section VI.

II. Related Work

Bhatia and Vandana surveyed Nearest Neighbor techniques (NN). NN technique is simple, effective, and robust to noise. This approach generates clusters of nearest neighbors where the nearest neighbor is identified by some calculated value. The KNN(K-Nearest Neighbor) is an unsupervised clustering technique and determines the nearest neighbor based on the k-value. They categorized the NN techniques into two groups; Structure less NN techniques and structure-based NN techniques. Structure less NN determines the nearest neighbor from the training and sample data. In structure-based NN, the nearest neighbors are determined based on the structure of data such as top-down or bottom-up like a k-d tree or bell tree. Both categories are the extended or improved version KNN techniques. They conclude that researchers can improvise the NN technique based on their research [19]. Ahalya and Pandey analyzed various clustering algorithms such as the K-means algorithm, Hierarchical algorithm, self-organizing map (SoM) algorithm, and expectation maximization(EM) algorithm. They described some tools used to implement the clustering algorithms. They compared and analyzed the above algorithms based on the type and size of the data set, the number of clusters, implementation tools, and accuracy. They reported that k-means and EM give better performance than other approaches and the SoM algorithm gives more accuracy than others [20].

MaryemAit El Hadj, Mohammed Erradi, and their team proposed an approach to cluster the security policies and additionally used the information of access log to detect the fraud intruders. They used the KNN algorithm to cluster the security rules and applied the rule-sub-module-reduction technique to minimize the count of rules in each cluster [21]. MaryemAit El Hadj and his team proposed a clustering approach to cluster XACML (eXtensible Access Control Markup Language) policies. XACML is an efficient markup language to express ABAC policies. They cluster the rules based on their similarity. Every pair of rules is clustered if the similarity value of the pair or rules is greater than the threshold value of 0.8 [23]. This method ensures that each rule must be clustered once, and the same rule may be contained in more than one cluster. There must be non-empty clusters only. The above approach may produce a maximum number of clusters and clusters with a maximum number of rules [22]. In contrary to the above research, we proposed and implemented a novel approach that our approach uses the basic technique of hierarchical clustering algorithm [24]. Rather than using the distance between the new rule (data point) and the cluster center, we use the similarity value of the new rule and the existing rules in the previous cluster. We use a rule-specific-cluster-merging approach instead of using the rule-sub-module-reduction method. Thus if the similarity value of two rules Rule-1 and Rule-2 is above the threshold value, then we create a new cluster if and only if Rule-1 and Rule-2 are new rules that are not yet clustered. If Rule-1 (or Rule-2) is already clustered and Rule-2 (or Rule-1) is not clustered and matched with all rules of the existing cluster, which contains Rule-1 (or Rule-2), then Rule-2 (or Rule-1) is merged with that existing cluster. Our enhanced approach reduces the generation of more clusters, redundancy (the same rule is contained in many clusters), and avoids the conflict-clusters (not all rules in a clusters are similar). Table II summarizes our approach and the researches done in previously.

TABLE II. SUMMARIZATION OF PREVIOUS RESEARCHES TOWARDS CLUSTERING INCLUDING THIS PAPER

References	Proposed work
[19]	Surveyed Nearest Neighbor (NN) clustering techniques. The nearest neighbor is identified by the calculated value. Concluded that the NN technique is efficient, simple and robustness and researchers can use and improvise the NN techniques based on their research.
[20]	Compared and analyzed various clustering algorithms such as the K-means algorithm, Hierarchical algorithm, self-organizing map (SoM) algorithm, and expectation maximization(EM) algorithm based on the type and size of the data set, the number of clusters, implementation tools, and accuracy. Reported that k-means and EM give better performance than other approaches and the SoM algorithm gives more accuracy than others.
[21]	Proposed an approach to cluster the security policies using the KNN algorithm and additionally used the information of access log to detect the fraud intruders. Applied rule-sub-module-reduction technique to minimize the count of rules in each cluster.
[22]	Proposed a clustering approach to cluster XACML (eXtensible Access Control Markup Language) policies based on the similarity value. Pair of rules are clustered if they are similar rules. This approach generates an increased number of clusters and also the cluster contains an increased number of rules. A rule may be contained in more than one cluster (redundancy) and not all the rules in a cluster are similar (conflict-cluster).
This paper	Proposed novel approach that uses the basic technique of hierarchical clustering algorithm. We use a rule-specific-cluster-merging approach instead of using the rule-sub-module-reduction method. Our enhanced approach reduces the generation of more clusters, avoidable redundancy (the same rule is contained in many clusters), and avoids the conflict clusters (not all rules in clusters are similar).

III. PRELIMINARIES

The fundamental concept of the ABAC model is described in this section. This also section describes how the ABAC rules are expressed and how the similarity value of pair of rules is calculated.

A. ABAC Model

This section describes the basic concepts of the ABAC model and the simple authentication process. The common simple authentication process is making a decision (allow or deny) by using the information like username and password of the subject or user for the request to access the shared resources such as files, databases, software. This simple process of checking the identity of the subject is not sufficient to meet the security needs of today's emerging computing technologies. RBAC and ABAC are the models used to address the complex security requirements of new computing technologies like cloud and fog computing. While RBAC fails to generate dynamic mappings between subject and object and also it is a role-independent model, ABAC or the combined features of RBAC and ABAC can be used for the protection of shares resources [25]. In this paper, we use the ABAC model and we aim to detect and eliminate the anomalies in ABAC security policies. This paper proposes an enhanced approach for clustering the security policies to simplify the process of detection and removal of anomalies. Fig. 2 and Fig. 3 show the simple standard authentication system and the ABAC model.

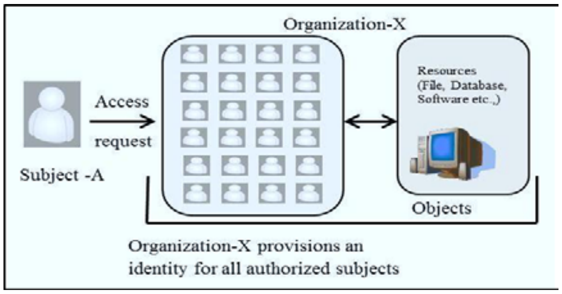

Organization-X
Access
request
Subject -A
Resources (File, Database, Software etc.,)
Objects
Organization-X provisions an identity for all authorized subjects

Fig. 2 Traditional Authentication System.

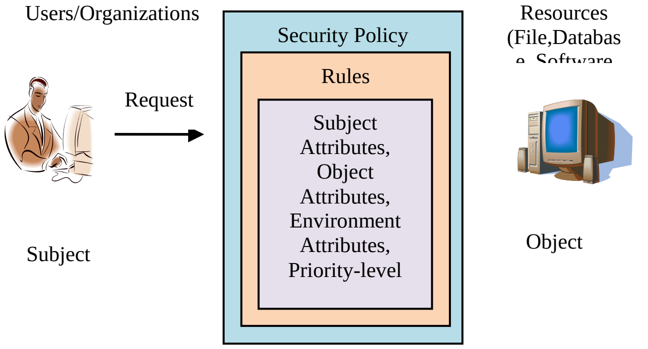

Users/Organizations
Request
Subject
Security Policy
Rules
Subject Attributes, Object Attributes, Environment Attributes, Priority-level
Resources (File,Databas e, Software
Object

Fig. 3 The Architecture of ABAC Moel.

ABAC model has a set of determined security policies where each security policy consists of security rules. In general, the security rule of the ABAC model is expressed with the decision (allow or deny), operations (read, write, print, etc.), and the attributes of subject (who requests the access), object (shared resources), and environment conditions (time, the importance of the operation, etc.). The decision either allowing or denying the subject’s request is made based on the attributes of the subject, object, and environmental conditions specified in the rules. In our research work additionally, we added one parameter Priority-level to each security rule to avoid the demand on a single object. The priority level is assigned based on the subject’s attributes and the importance of the operation on the object. The most common jargons used in the ABAC model are:

1) *Subject and its attributes:* The term subject refers to a single or group of users or organizations or the process that requests the resource. The attributes of subjects are the important credential information about the subject like name of user or process, designation, department, affiliation, etc.

2) *Object and its attributes*: Shared resources (operating system, network, file, software, and database) are named as objects. The information like name, type, owner, and date of creation are the attributes of objects.

3) Operation: The task (read, write, etc.) on the object is requested.

4) Environmental conditions: Other information such as date of request, current time, waiting period, etc.

5) Rule and decision: Constructed with the above three categories subject, object, and environmental conditions. In our proposed approach, we included the additional parameter priority-level assigned based on the attributes of the subject. The rule takes the decision (allow or deny) based on the attributes of the subject, object, and environmental conditions.

6) Policy: It is a set of rules established for the protection of objects in an organization.

B. Expression of ABAC Security Policy

The security policies of ABAC are constructed with a set of rules. Every rule contains the attributes of the subject, objects, and environment. The decision (allow or deny) is made based on the values of the attributes specified in the rule [26]. We added one additional parameter priority-level in every rule to avoid the anomaly conflict-demand. Thus more requests on a limited object are referred to as conflict-demand. The parameter priority-level is a non-negative integer. The demand for a limited object is handled based on the value of the priority level. Thus the rule or request which has the highest priority value will be allowed to get access. The pair of rules is not clustered if the priority-level of the two rules are not equal. The security policy set of the ABAC model is expressed by JavaScript Object Notation (JSON) as follows:

```
Policy-Set: [ {
Policy: <policyID>,
Rules: [{
<ruleID>,
Action: <actionName>,
Operation :<operationName>,
Subject: <subjectID>,
Resource: <resourceName>
Environmental :<environmentID>},
        }]
    }]
```

Extensible Access Control Markup Language (XACML) is the most common acceptable and widely used language to express the security policies. Each attribute is expressed with the pair key and value using a markup language. Key is the attribute name (eg. Department) specified as the tag and value (eg. Urology) is the value of the attribute name specified within the key-tag. The ABAC policy set can be expressed by XACML as follows:

```
<PolicySet>
 <Policy PolicyID=”P1”>
   <Rule RuleID=”R1” Decision=”Allow” >
     <Operation>
             <Operation-1>read</Operation-1>
             <Operation-2>write</Operation-2>
     </Operation>
     <Subject>
             <Department>urology</Department>
             <Designation>surgeon<Designation>
      </Subject>
      <Object>
                                   <ResourceName>

    PatID_005_Urine_Culture_Report
             </ ResourceName >
      </Object>
      <EnvironmentalCondition>
             <Duration>8:18</Duration>
      </EnvironmentalCondition>
    </Rule>  ………// more rules can be specified
  </Policy>   ………// more policies can be specified
</PolicySet>
```

In the above example, rule R1 states that security policy allows the surgeon in the department of urology to read and write the file ‘PatID_005_Urine_Culture_Report’ during the time 8:18 hours. The following standard is usually used to write the rules.

Rule$_1$= {Allow$_{read}$ | Designation = {duty-doctor, surgeon}, Department = { hematology}, File-Name = {pat_007_blood_report}, Time= {8:18}, PriorityLevel=2}

Rule$_2$ = {Allow$_{read}$ | Designation = {head-nurse}, Department = {hematology}, File-Name = { pat_007_blood_report }, Time= {8:18}, PriorityLevel=1}

The above rules Rule$_1$ and Rule$_2$ are security rules represented including the parameter PriorityLevel. The attributes Designation and Department are the categories of subject, File-Name is the attribute of category object, and the attribute Time is the category of environment. The rules state that the file ‘pat_007_blood_report’ can be read by the users ‘duty doctor, ‘surgeon’ and ‘head nurse’ belong to the department ‘hematology’. When all these three subjects are trying to get access to the file, ‘duty doctor, ‘surgeon’ will get access first due to the highest priority. The relationship diagram of security policy and working principle of ABAC model is shown in Fig. 4. This security system provides two stages of authorization. The first stage performs the common traditional authorization process where the second stage completes the ABAC mechanism to increase the level of protection.

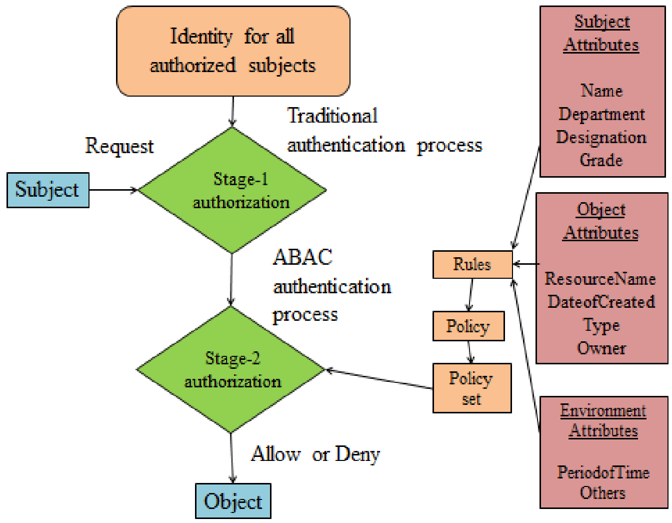


Fig. 4 The Security System Provides Two Stages of Authorization.

C. *Measurement of Similarity Value*

We used the following formulae to measure the similarity value of the pair of rules[27]. The similarity value (S) of the pair of rules R_1 and R_2 for an attribute ‘atr’ is calculated using formula-1.

$$S_{atr}\ (R_1, R_2) = \frac{NSV}{NDV} \quad (1)$$

Where NSV (Number of Same Values) is the number of the same values of the attribute ‘reappeared in both the rules R_1 and $R_{2.}$ The variable NDV (Number of Distinct Values) is the number of distinct values of the attribute ‘atr’ in the rules R_1 and R_2. The similarity-value (S) of two rules for category C (Subject, Object, and Environment) is measured using the formula-2.

$$S_C\ (R_1, R_2) = \sum_{atr \in \{ATS\,(R1) \cap ATS\,(R2)\}} P_{atr}\, S_{atr}\ (R_1, R_2) \quad (2)$$

where ATS(R_1) and ATS(R_2) are a set of attributes of R_1 and $R_{2,}$ respectively. ‘$P_{atr'}$ is the probability of the attribute (P_{atr} =1/number of the common attribute in both R1 and R2). The variable ‘atr’ is the set of attributes that appeared in both R1 and R2 under the category C. The similarity-value (S) of the pair of the rule is measured by the formula-3.

$$S\ (R_1, R_2) = \begin{cases} P_{subject}\, S_{subject}(R_1, R_2) + \ P_{object}\, S_{object}(R_1, R_2) + \\ P_{environment}\, S_{environment}(R_1, R_2) \end{cases} \quad (3)$$

The variable P is the probability of the category (subject, object, and environment). As equal probability is applied, the probability of each category is 1/3. The similarity value of the above-mentioned rules $Rule_1$ and $Rule_2$ is calculated as,

$$S(Rule1,Rule2) = \frac{1}{3}S_{subject}\ (R_1,\ R_2) + \frac{1}{3}S_{object}(R_1,R_2) + \frac{1}{3}S_{environment}(R_1,R_2)$$

$$= \frac{1}{3}\times (0.25+0.5) + \frac{1}{3}\times 1 + \frac{1}{3}\times 1$$

IV. System Architecture

Fig. 5 shows the system architecture of our proposed method. It contains the rules-clusters storage and management module, rule generation module, and rule clustering module. The functional operation, task, requirements of each module, and relationship among the modules are described in the following sections.

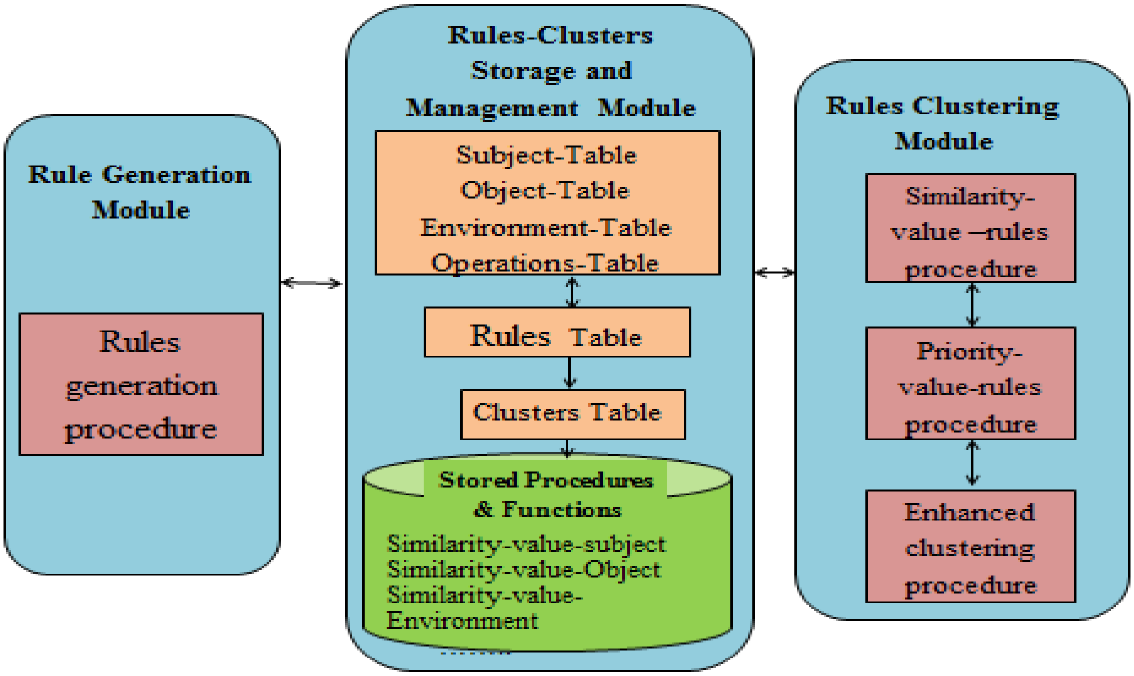

Rule Generation Module
Rules generation procedure
Rules-Clusters Storage and Management Module
Subject-Table
Object-Table
Environment-Table
Operations-Table
Rules Table
Clusters Table
Stored Procedures & Functions
Similarity-value-subject
Similarity-value-Object
Similarity-value-Environment
Rules Clustering Module
Similarity-value –rules procedure
Priority-value-rules procedure
Enhanced clustering procedure

Fig. 5 The System Architecture of Our Approach.

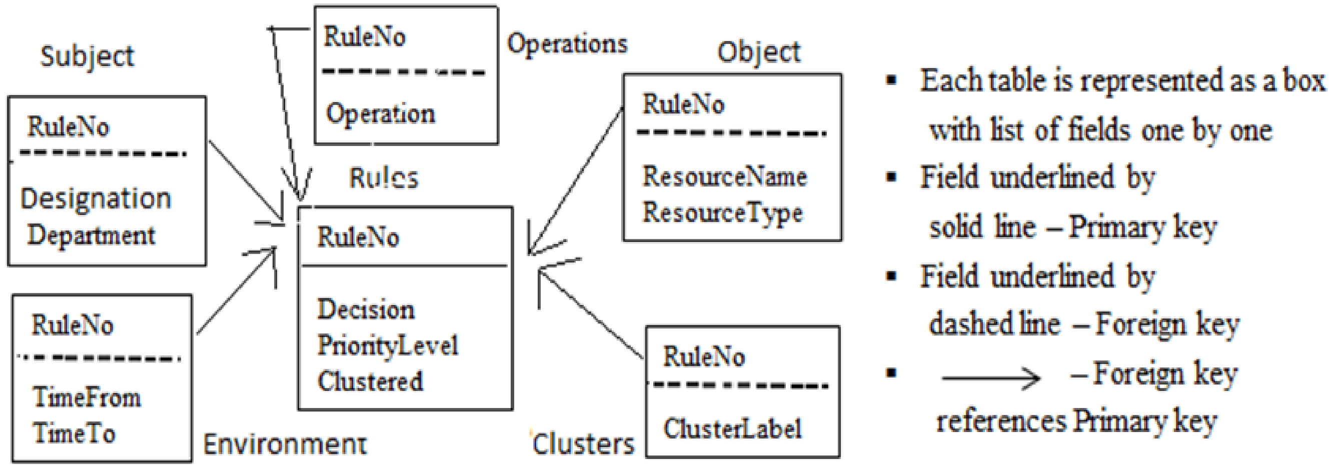

Subject
RuleNo
Designation
Department
RuleNo
Operation
Operations
Rules
RuleNo
Decision
PriorityLevel
Clustered
Object
RuleNo
ResourceName
ResourceType
RuleNo
TimeFrom
TimeTo
Environment
Clusters
RuleNo
ClusterLabel
▪ Each table is represented as a box with list of fields one by one
▪ Field underlined by solid line – Primary key
▪ Field underlined by dashed line – Foreign key
▪ ⟶ – Foreign key references Primary key

Fig. 6 Schema Diagram of Specification of ABAC Policies and Clusters.

A. Rules-Clusters Storage and Management Module (RCSMM)

We created relations (tables) in Oracle11 (Oracle 12c is for an in-memory database environment). The schema diagram of RCSMM is shown in Fig. 6. The relation 'Subject' is used to store and manage the attributes and values of the subject. Likely the relation 'Object' and the relation 'Environment' store and manage the attributes of the resources and the environmental conditions respectively. The generated security policies are stored and maintained in the table 'Rules'. The information about the request of operation for a resource is maintained in the relation 'Operations'. The generated security rules are clustered and stored in the relation 'Clusters'. Table III, Table IV, Table V, Table VI show the sample records of the relations Subject, Object, Environment, and Clusters.

Table VI illustrates that rules 1 and 2 are clustered in cluster-1 and cluster-11 consists of rules 32 and 33. We created a relation for each category (Subject, Object, and Environmental Conditions) to store and manage the value of all attributes. We have written stored functions to measure the similarity value of pair of rules for an attribute (e.g. Department), the similarity value of pair of rules for each category (e.g. Subject) and finally to measure the similarity value of pair of rules.

TABLE III. ATTRIBUTES OF SUBJECT

Relation: Subject		
RULENO	RULENO	RULENO
9	9	9
35	35	35
501	501	501
720	720	720

TABLE IV. ATTRIBUTES OF OBJECT

Relation: Object	
RULENO	RESOURCENAME
9	Pat00005_blood_report
35	Pat00007_urine_report
501	Pat00008_thyroid_report
720	Pat00039_blood_report

TABLE V. ATTRIBUTES OF ENVIRONMENT

Relation: Environment		
RULENO	TIMEFROM	TIMETO
9	8	18
35	6	14
501	8	18
720	7	19

TABLE VI. CLUSTERS OF RULES

Relation: Clusters	
RULENO	CLUSTERLABEL
1	1
2	1
32	11
33	11

The above-stored function measures the similarity value of pair of rules for the attribute Department. The stored function call nsv_department(r1,r2) returns the number of values that are the same for the attribute 'Department' in both r1 and r2. The stored function call ndv_department(r1,r2) returns the number of distinct values for the attribute 'Department' in both r1 and r2.

Stored function sv_department(r1 number,r2 number)

```
create or replace function sv_department(r1 number, r2 number)
return float is
sv float;
nsv number(3);
ndv number(3);
begin
nsv:=nsv_department(r1,r2);
ndv:=ndv_department(r1,r2);
if nsv=0 or ndv=0 then
sv:=0;
else
sv:=nsv/ndv;
end if;
return sv;
end sv_department;
```

We have created stored functions to measure the similarity values of all attributes. The stored function sv_rules(r1,r2) measures the final similarity value of two rules r1 and r2. In this procedure p is the probability where we use equal probability of all the categories (Subject, Object, and Environment) and sv_subject(r1,r2), sv_object(r1,r2), sv_environment(r1,r2) are the similarity values of Subject, Object and Environment.

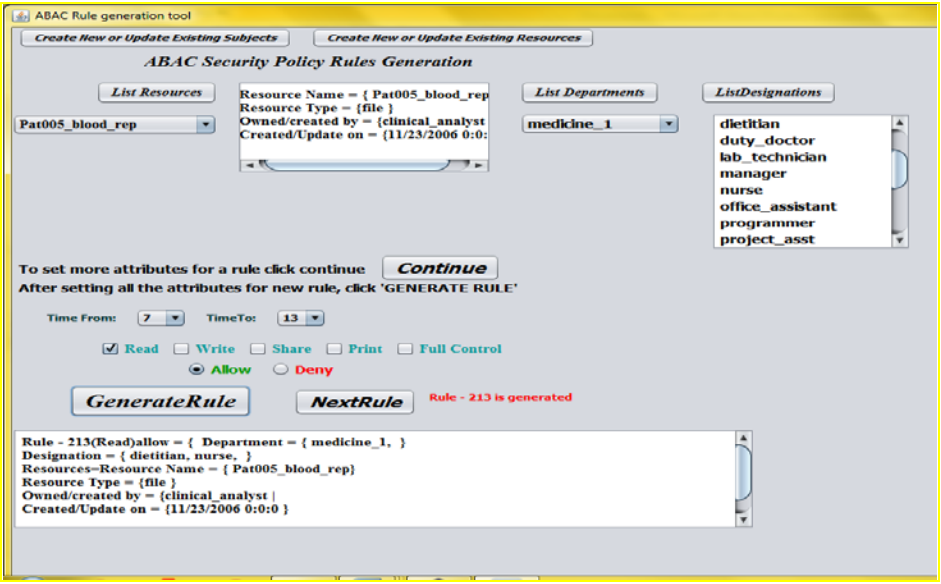
Stored procedure function sv_rulet(r1 number,r2 number)

```
create or replace function sv_rule(r1 number,r2 number)
return float is
sv float;
p float;
sv1 float;
sv2 float;
sv3 float;
begin
p:=1/3;
sv1:=sv_subject(r1,r2);
sv2:=sv_object(r1,r2);
sv3:=sv_environment(r1,r2);
sv:=(p*sv1)+(p*sv2)+(p*sv3);
return sv;
end sv_rule;
```

B. Rule Generation Module (RGM)

We designed the module RGM in java for automated rule generation. RGM is an admin interface in java to generate or specify the security policies for all shared resources. The admin can easily set the rules for a resource with the attributes of the categories Subject, Object, and Environment. This module is flexible for the admin to update the rules easily. This module allows the admin to create, or delete or update the entities of the subject, object, and environmental conditions. The priority of the rule is determined based on the role of the subject. The subject-attribute 'grade' is assigned with the value (low or high or middle) based on the attribute designation. The priority of the rule or request is determined based on the attribute 'grade'. We have written a rule generation procedure in Java to generate ABAC rules.

Fig. 7 shows the interface of rule generation. The generated policies are stored in the relations Rules, Subject, Object, Environment, and Operations. Fig. 8 shows the sample of generated rules.



ABAC Rule generation tool
Create New or Update Existing Subjects
Create New or Update Existing Resources
ABAC Security Policy Rules Generation
List Resources
List Departments
ListDesignations
Continue
GenerateRule
NextRule

Pat005_blood_rep

Resource Name = { Pat005_blood_rep
Resource Type = {file }
Owned/created by = {clinical_analyst
Created/Update on = {11/23/2006 0:0:

medicine_1

dietitian
duty_doctor
lab_technician
manager
nurse
office_assistant
programmer
project_asst

To set more attributes for a rule click continue
After setting all the attributes for new rule, click 'GENERATE RULE'

Time From: 7 TimeTo: 13

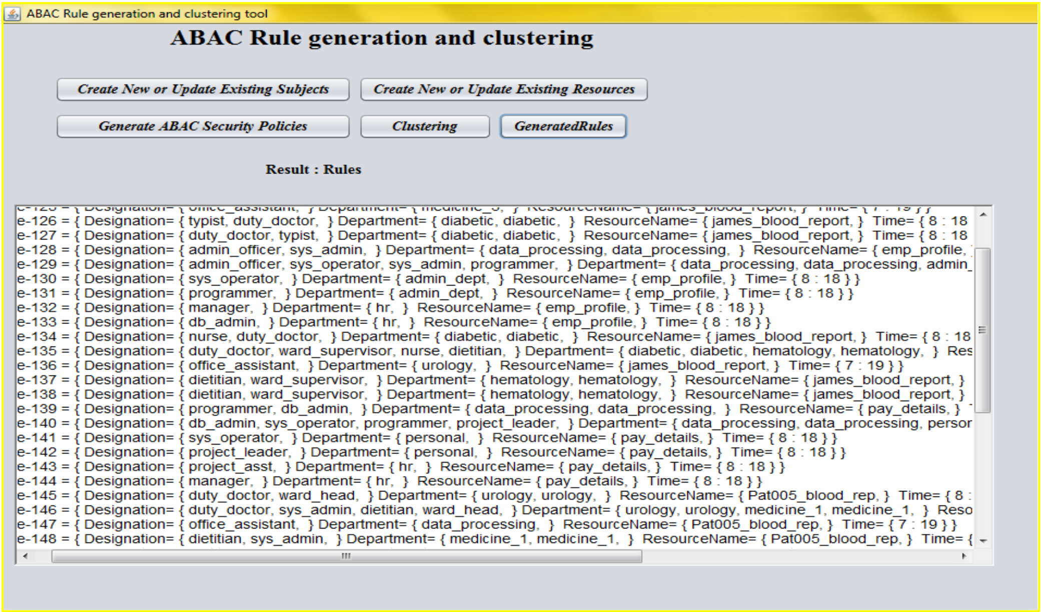
☑ Read □ Write □ Share □ Print □ Full Control

● Allow ○ Deny

Rule - 213 is generated

```
Rule - 213(Read)allow = {  Department = { medicine_1,  }
Designation = { dietitian, nurse,  }
Resources=Resource Name = { Pat005_blood_rep}
Resource Type = {file }
Owned/created by = {clinical_analyst |
Created/Update on = {11/23/2006 0:0:0 }
```

Fig. 7 ABAC Rule Generation Tool.


ABAC Rule generation and clustering tool
ABAC Rule generation and clustering
Create New or Update Existing Subjects
Create New or Update Existing Resources
Generate ABAC Security Policies
Clustering
GeneratedRules
Result : Rules

```
e-125 = { Designation= { office_assistant, } Department= { medicine_5, } ResourceName= { james_blood_report, } Time= { 7 : 19 } }
e-126 = { Designation= { typist, duty_doctor,  } Department= { diabetic, diabetic,  }  ResourceName= { james_blood_report, }  Time= { 8 : 18
e-127 = { Designation= { duty_doctor, typist,  } Department= { diabetic, diabetic,  }  ResourceName= { james_blood_report, }  Time= { 8 : 18
e-128 = { Designation= { admin_officer, sys_admin,  } Department= { data_processing, data_processing,  }  ResourceName= { emp_profile,
e-129 = { Designation= { admin_officer, sys_operator, sys_admin, programmer,  } Department= { data_processing, data_processing, admin_
e-130 = { Designation= { sys_operator,  } Department= { admin_dept,  }  ResourceName= { emp_profile, }  Time= { 8 : 18 } }
e-131 = { Designation= { programmer,  } Department= { admin_dept,  }  ResourceName= { emp_profile, }  Time= { 8 : 18 } }
e-132 = { Designation= { manager,  } Department= { hr,  }  ResourceName= { emp_profile, }  Time= { 8 : 18 } }
e-133 = { Designation= { db_admin,  } Department= { hr,  }  ResourceName= { emp_profile, }  Time= { 8 : 18 } }
e-134 = { Designation= { nurse, duty_doctor,  } Department= { diabetic, diabetic,  }  ResourceName= { james_blood_report, }  Time= { 8 : 18
e-135 = { Designation= { duty_doctor, ward_supervisor, nurse, dietitian,  } Department= { diabetic, diabetic, hematology, hematology,  }  Res
e-136 = { Designation= { office_assistant,  } Department= { urology,  }  ResourceName= { james_blood_report, }  Time= { 7 : 19 } }
e-137 = { Designation= { dietitian, ward_supervisor,  } Department= { hematology, hematology,  }  ResourceName= { james_blood_report, }
e-138 = { Designation= { dietitian, ward_supervisor,  } Department= { hematology, hematology,  }  ResourceName= { james_blood_report, }
e-139 = { Designation= { programmer, db_admin,  } Department= { data_processing, data_processing,  }  ResourceName= { pay_details, }
e-140 = { Designation= { db_admin, sys_operator, programmer, project_leader,  } Department= { data_processing, data_processing, persor
e-141 = { Designation= { sys_operator,  } Department= { personal,  }  ResourceName= { pay_details, }  Time= { 8 : 18 } }
e-142 = { Designation= { project_leader,  } Department= { personal,  }  ResourceName= { pay_details, }  Time= { 8 : 18 } }
e-143 = { Designation= { project_asst,  } Department= { hr,  }  ResourceName= { pay_details, }  Time= { 8 : 18 } }
e-144 = { Designation= { manager,  } Department= { hr,  }  ResourceName= { pay_details, }  Time= { 8 : 18 } }
e-145 = { Designation= { duty_doctor, ward_head,  } Department= { urology, urology,  }  ResourceName= { Pat005_blood_rep, }  Time= { 8 :
e-146 = { Designation= { duty_doctor, sys_admin, dietitian, ward_head,  } Department= { urology, urology, medicine_1, medicine_1,  }  Reso
e-147 = { Designation= { office_assistant,  } Department= { data_processing,  }  ResourceName= { Pat005_blood_rep, }  Time= { 7 : 19 } }
e-148 = { Designation= { dietitian, sys_admin,  } Department= { medicine_1, medicine_1,  }  ResourceName= { Pat005_blood_rep, }  Time= {
```

Fig. 8 Sample Generated Rules.

C. Rule Clustering Module (RCM)

We implemented our enhanced clustering algorithm in java to cluster the security policies. In our previous work [28] [29], we applied a direct clustering approach that every rule R_i (from top to bottom) is paired with the rule R_j (from the next successive rule of Ri to the last rule in the database). The pair of rules are clustered if the similarity value of the pair of rules is above the threshold value and the difference between the priority-level of those rules is zero (priority-level of two rules are equal). The introduction of the parameter priority-level reduces the size of the clusters [30].

Although the previous approach clusters the rules efficiently, it produces more clusters, and the same rule is clustered in multiple clusters. In our enhanced approach, we follow the top-to-bottom approach thus we start from the first rule to last (1,2,3,..,n) and for each rule R_i we made the pairing of the rule R_i with every R_j (j=i+1,i+2,..,n}. If two rules R_i and R_j are similar rules based on the condition (similarity-value(R_i, R_j)>0.8 and priority-level(R_i) is equal to priority-level(R_j)), then we follow the following criteria to cluster the pair of rules R_i and $R_{j.}$

- The rules R_i and R_j are stored in a new cluster if R_i is not yet clustered. (first similar rule of R_i)
- In the case of the rule, R_i is already clustered in $C_{k,}$ and R_j is not clustered: (rule-specific-cluster-merging technique)
    - the rule R_j is merged with the cluster C_k if R_j is similar to all the rules in the cluster $C_{k,}$
    - otherwise, R_j and R_i are stored in a new cluster.
- Every rule should be clustered, and any cluster should not be empty.

The enhanced clustering algorithm is written as follows:

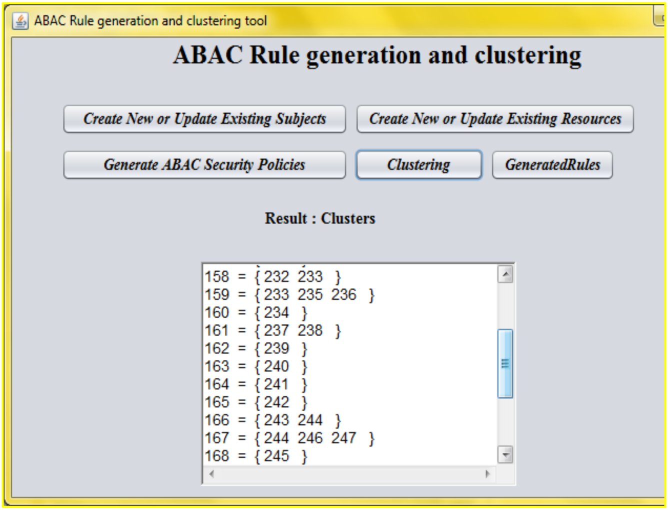
Algorithm: Rule-Specific-Cluster-Merging Approach (RSCA)

```
Input : Security rules $R_1$, $R_2$,…, $R_{n\,,}$ // n is number of rules
 Output: Cluster of Rules $C_1$, $C_2$,.., $C_k$ // k is the number of clusters
K=0; // K is the number of clusters
T={$R_1$, $R_2$,……$R_n$} /* T is the table of records and each record
$R_i$ contains information of single rule */
For each rule $r_1$ in $R_1$ to $R_n$ loop   //form first record to last
record
    For each rule $r_2$ in $R_{i+1}$ to $R_n$ loop
        Compute Similarity-value($r_{1\,,}$ $r_2$)
        If Similarity-value ($r_1$ $_,r_2$)>0.8 and  priority-level($r_1$) = 
            priority-level($r_2$) then
            If $r_1$ is not yet clustered then
                K=K+1; $C_k$={ $r_1$, $r_2$}
            Else if $r_1$ is already clustered and  $r_2$is not clustered
            then
                Found the cluster $C_F$ such that $r_2$ is similar to
                all the  rules in $C_F$ where $r_1$ is the member of
                this cluster
                If such cluster $C_F$ is found then
                    $C_F$= $C_F$ U {$r_2$}
                // is merged with the existing cluster.
                Else If such cluster $C_F$ is not found then
                    K=K+1;$C_k$={ $r_{1,}$ $r_2$}
                // new cluster is created to store $r_1$ and $r_2$
                End If
            End If
        End If
    End Loop
    If $r_1$ is not yet clustered then
        // none of the rules is similar to $r_1$
        K=K+1;$C_k$={ $r_1$} // $r_1$ is clustered in a new cluster
    End If
End Loop
End
```

The implementation of our enhanced clustering approach clusters the given set of rules. Fig. 9 shows the results of our implementation. We managed six tables Rules, Operations, Subject, Object, Environment, and Clusters to store the detail of rules, operations on objects, subjects, objects, environments, and clusters respectively. The creation of an individual table for every entity avoids the redundancy of data and null values.

Rule clustering module (RCM) proposes an approach to cluster ABAC Policies before policy validation to improve the performance of the model (Resolving the policy errors in every rule is a time-consuming and complex process). We used only the basic technique of the hierarchical clustering algorithm. We introduce a novel clustering (rule-specific-cluster-merging) and resolved the two major problems of clustering: avoidable-redundancy (the same rule is contained in many clusters), and avoids the conflict clusters (not all rules in clusters are similar). The scopes of our approach are:

- We used a rule-specific-cluster-merging approach (instead of using the rule-sub-module-reduction method or cluster-merging method). (time complexity is low).
- If two rules are already clustered, our approach will not entertain the clustering once again (reduces the number of clusters and a rule is not stored in multiple clusters unnecessarily).
- If a pair of rules are similar and anyone rule is already clustered, then we are seeking to cluster the next one with the specific clusters where the clustered rule is a member. (improves the performance of the process).
- A rule is merged with the existing cluster if all the rules of the existing clusters are similar only (avoids conflict-clusters-not all rules in a cluster are similar).


ABAC Rule generation and clustering tool
ABAC Rule generation and clustering
Create New or Update Existing Subjects
Create New or Update Existing Resources
Generate ABAC Security Policies
Clustering
GeneratedRules

Result : Clusters

```
158 = { 232 233  }
159 = { 233 235 236  }
160 = { 234  }
161 = { 237 238  }
162 = { 239  }
163 = { 240  }
164 = { 241  }
165 = { 242  }
166 = { 243 244  }
167 = { 244 246 247  }
168 = { 245  }
```

Fig. 9 Sample Result of Clustering.

V. Results and Discussion

We described and compared our approach with the hierarchical clustering algorithm. Fig. 10 illustrates the example of the hierarchical clustering algorithm.

We took eight rules {R1,R2,R3,R4,R5,R6,R7,R8} for clustering. In the first step, the hierarchical algorithm considers each rule as a cluster. At each iteration, every cluster is compared with other clusters and the similar clusters are merged. This is a continuous process while there are comparable clusters. Let we consider.

C is the cluster to be compared,

k is the number of current clusters,

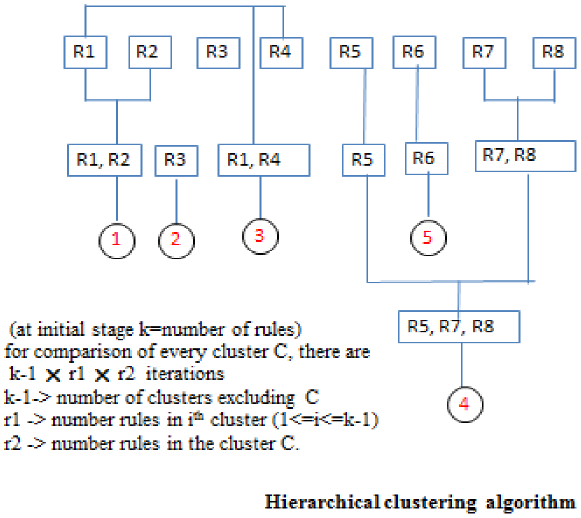
k-1 is the number of clusters excluding C,

r1 is the number of rules in the ith cluster (1≤ i≤ k-1) and

r2 is the number of rules in C.

The comparison of every cluster C, this approach performs k-1 × r1 × r2 iterations. Thus each cluster is compared with all other clusters at every iteration and similar clusters are only merged. The clusters {R1} and {R2} are merged. The cluster {R3} and {R6} are not similar with all other clusters. The clusters {R1} and {R4} are merged. The clusters {R5}, {R7} and {R8} are merged in a cluster. The generated clusters are {R1,R2}, {R3}, {R1,R4}, {R5,R7,R8}, and {R6}.

Fig. 11 illustrates the same example of clustering eight rules with our proposed approach. In the first step of our approach, we compare every rule R_i (1≤i≤n-1, n is the number of rules) with Rj (i+1 ≤ j ≤ n). At each iteration, the un-clustered rule is compared only with the specific clusters, not all the clusters. Thus if R1 and R2 are similar rules, R1 is clustered and R2 is yet to clustered, then R2 is compared with the clusters that contain R1. If k is the number of specific clusters (where the feasibility of similarity occurs for the un-clustered rule) and t is the number of rules in the kth cluster, then our approach requires k×t comparisons only. Hence our approach reduces the number of comparisons and consumes less time than the other approaches. For the discussion of our proposed approach, we took five rules as a sample (listed in Table VII) to make the discussion easy and understandable.



(at initial stage k=number of rules)
for comparison of every cluster C, there are
k-1 × r1 × r2 iterations
k-1-> number of clusters excluding C
r1 -> number rules in ith cluster (1<=i<=k-1)
r2 -> number rules in the cluster C.

Fig. 10 Example of the Hierarchical Clustering Approach.

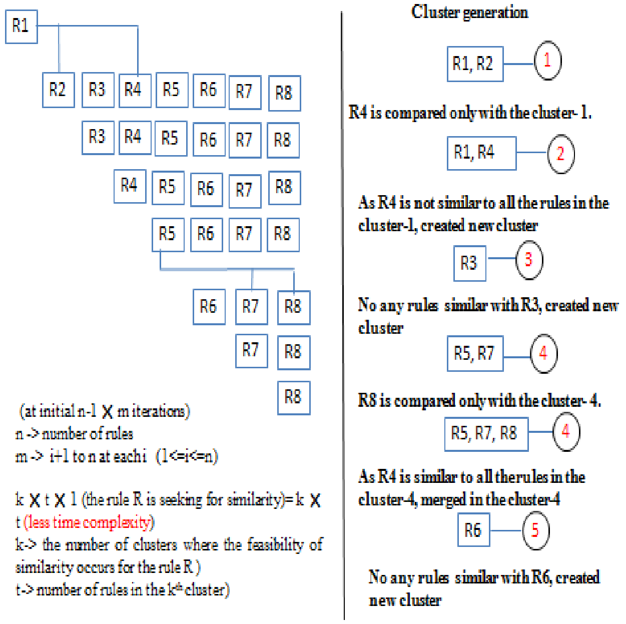

R1
R2 R3 R4 R5 R6 R7 R8
R3 R4 R5 R6 R7 R8
R4 R5 R6 R7 R8
R5 R6 R7 R8
R6 R7 R8
R7 R8
R8
(at initial n-1 X m iterations)
n -> number of rules
m -> i+1 to n at each i (1<=i<=n)
k X t X 1 (the rule R is seeking for similarity)= k X t (less time complexity)
k-> the number of clusters where the feasibility of similarity occurs for the rule R)
t-> number of rules in the k^th cluster)
Cluster generation
R1, R2
1
R4 is compared only with the cluster- 1.
R1, R4
2
As R4 is not similar to all the rules in the cluster-1, created new cluster
R3
3
No any rules similar with R3, created new cluster
R5, R7
4
R8 is compared only with the cluster- 4.
R5, R7, R8
4
As R4 is similar to all the rules in the cluster-4, merged in the cluster-4
R6
5
No any rules similar with R6, created new cluster

Fig. 11 Example of Our Proposed Clustering Approach.

The similarity value of each pair of rules are (1,2)= 0.83, (1,3)= 0.33, (1,4)= 0.33, (2,3)= 0.33, (2,4)=0.83, (2,5)=0.83, (3,4)=0.33, (3,5)=0.33, (4,5)=1. The workflow and result of the above sample rules are described in Table VIII. Our novel clustering approach produces three clusters C1={2,5,4}C2={1,2} C3={3} for the above five rules. In our proposed approach, all rules in every cluster are similar. Thus a rule is merged if it is similar to all the rules already exist in that cluster. In some previous researches, the number of clusters is very high. Also, only minimum rules are similar in a cluster, and this leads to complexities in detecting and removing anomalies [22]. Unlike the rule-sub-module-reduction method [21], we merge the cluster at the time of the creation of the cluster itself, which increases the performance of the approach. In our previous work [28], we introduced the parameter priority-level to avoid the anomaly conflict-demand and reduce the more number of clusters. Although our previous research reduced the number of clusters, this enhanced cluster with the cluster-merging technique decreases the creation of more clusters and also maintains the constraints that all rules in a cluster are similar only.

Fig. 12 illustrates the comparison of our novel rule-specific clustering approach (RSCA) with the previous approaches 'Cluster-based approach' (CBA) and 'Log-based clustering approach' (LBCA) [21][22] (Maryem Ait El Hadj, Mohammed Erradi). CBA generates more clusters than LBCA and RSCA. Comparing to the hierarchical clustering algorithm(HCA) and other discussed previous approaches, our proposed clustering approach reduces the generation of more clusters and the number of comparisons, hence this result shows that our proposed approach consumes less time and increases the performance of the clustering technique and this helps to improve the efficiency of the ABAC model. Fig. 13 gives the comparative analysis of the existing approaches with our proposed approach based on the time complexity (Time is represented as Nanoseconds).

Table IX shows the qualitative analysis of clustering approaches. We used rule-redundancy-clusters (the same rule is clustered in multiple clusters), conflict-clusters (not all the rules in a cluster are similar), high-cluster-generation (more number of clusters are generated), and less time-consuming (measured based on the number of comparisons and iterations) to perform this qualitative analysis.

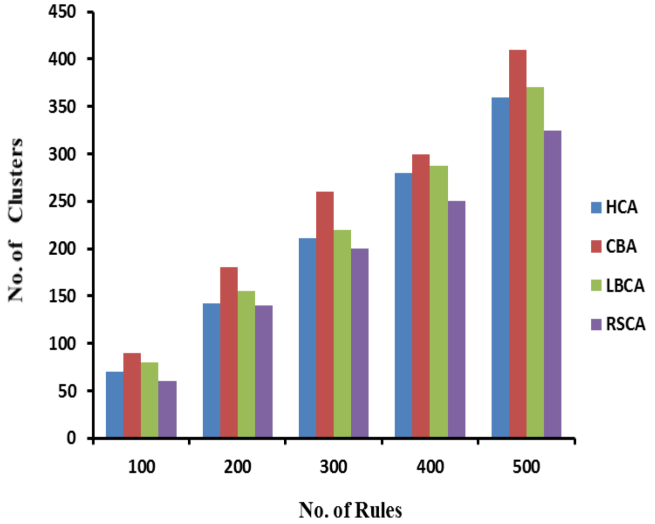


Fig. 12 A Comparative Study based on the Size and Number of Clusters.

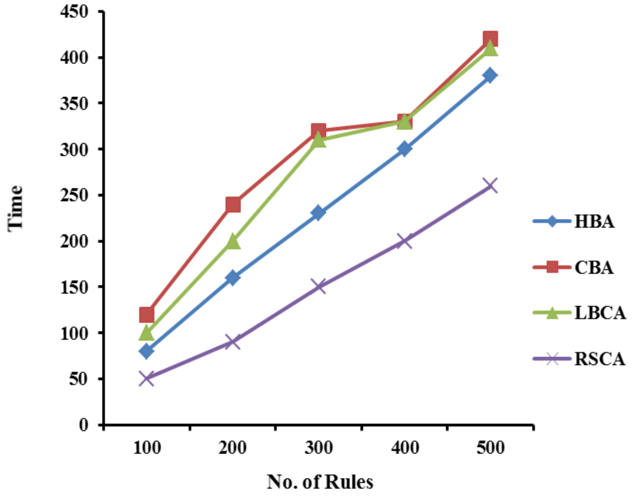


Fig. 13 A Comparative Study based on Time Complexity.

TABLE VII. SAMPLE ABAC RULES

RuleNo	Decision	PriorityLevel	Designation	Department	ResourceName	Time
2	allow$_{read}$	1	db_admin, typist, sys_asst, duty_doctor	diabetic, hematology	james_blood_report	8:18
3	allow$_{read}$	0	Office assistant	medicine_3	james_blood_report	7:19
5	allow$_{read}$	1	typist	diabetic	james_blood_report	8:18
1	allow$_{read}$	1	sys_asst, db_admin	hematology	james_blood_report,	8:18
4	allow$_{read}$	1	typist, duty_doctor	diabetic	james_blood_report	8:18

TABLE VIII. RESULT OF OUR PROPOSED CLUSTERING APPROACH WITH A SAMPLE OF THE ABOVE FIVE SECURITY RULES

Pair of rules	Clusters generated	Discussion

(2,3) (2,5) (2,1) (2,4)	C1={2,5}
	C2={2,1}
	C1={2,4,5}
		(2, 3) is not similar. (2,5) is similar and clustered in C1. (2,1) is similar and tried to merge with the existing clusters (C1) where rule 2 is a member. But the rule 1 is not similar to all the rules in C1, so (2,1) is stored in a new cluster C2={2,1}. As (2,4) is similar and matched with all the rules in the existing cluster C1where rule 2 is a member, and rule 4 is merged with C1. C1={2,5,4}

(2,3) (2,4) (2,5)	C1={2,5,4} C2={1,2}	(2, 3) is not similar. (2,4) and (2,5) are similar pairs but already both are clustered
(3,4) (3,5)	C1={2,5,4}C2={1,2} C3={3}	(3,4) and (3,4) are not similar pairs. Hence 3 is clustered in new cluster C3={3}
(4,5)	C1={2,5,4}C2={1,2} C3={3}	(4,5) is similar but already it is clustered.

TABLE IX. QUALITATIVE ANALYSIS OF EXISTING ABAC POLICIES CLUSTERING AND OUR APPROACH

References	Technique	Reduces rule-redundancy-clusters	Avoids Conflict-clusters	Reduces high cluster-generation	Less time-consuming
M. A. El Hadj et al. [22]	A rule is clustered with all similar rules, comparable clusters are merged.	No	No	Yes	No
M. A. El Hadj et al. [21]	Enhanced approach and used rule-merge sub-module clustering	No	No	Yes	No
Hierarchical Clustering [24]	No centroid. All elements are considered clusters. At each cycle, comparable clusters are merged	Yes	Yes	Yes	No
Our approach	Rule-specific-cluster-merging approach	Yes	Yes	Yes	Yes

VI. CONCLUSION

Cloud service providers use many access control models in the implementation of the intrusion detection system to secure big data and other shared resources in the distributed environment. ABAC model meets the security requirements of advanced computing technologies like cloud computing, fog computing, and the Internet of Things (IoT). Detection and Removal of Anomalies (DRA) in the policies improve the efficiency and reliability of the model. Applying DRA in every cluster of rules rather than in every rule enhances the performance of the approach highly. Existing DRA approaches have the inability to prove the efficiency of the model, due to applying the poor clustering approach or the approach which not properly detect and resolve all considerable anomalies. Our research contribution consists of three modules: 1) developing a tool to generate ABAC policies; 2) Storing and managing ABAC policies in the database; 3) applying a novel rule-specific-cluster-merging algorithm to cluster ABAC policies. In previous researches, clustering methods produce more clusters and cluster a rule in multiple clusters. The previous approaches generate conflict clusters thus not all the rules in a cluster are similar, which results that in increases the complexity of the approach and degrades the performance too. Our novel approach with the rule-specific-cluster-merging technique reduces the generation of more clusters, avoids the clustering of the same rule in multiple clusters and conflict clusters. With our RSCA algorithm, we decrease the number of comparisons. Hence our approach gives high performance and reduces the complexity in applying DRA. The additional parameter priority-level in every rule avoids the anomaly conflict-demand. This is only our part of research towards improving the performance and efficiency of the ABAC model. Our future work is to detect and resolve the rule redundancy at the time of clustering to decrease the complexity in the clustering mechanism. In the future, we will propose an approach to detect and resolve all considerable policy errors to improve the efficiency of the ABAC model and intrusion detection system.

REFERENCES

[1] Q. Zhang, L. Cheng, and R. Boutaba, “Cloud computing: State-of-the-art and research challenges,” J. Internet Serv. Appl., vol. 1, no. 1, pp. 7–18, 2010, doi: 10.1007/s13174-010-0007-6.

[2] P. B. Tarigan, “Encyclopedia of Cryptograpghy and Security, Second Edition, Springer,” Journal of Chemical Information and Modeling, vol. 53, no. 9. pp. 1689–1699, 2013, doi: 10.1017/CBO9781107415324.004.

[3] C. Liang et al., “Intrusion detection system for the internet of things based on blockchain and multi-agent systems,” Electron., vol. 9, no. 7, pp. 1–27, 2020, doi: 10.3390/electronics9071120.

[4] C. Prakash and S. Dasgupta, “Cloud computing security analysis: Challenges and possible solutions,” Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016, pp. 54–57, 2016, doi: 10.1109/ICEEOT.2016.7755626.

[5] M. Mukherjee et al., “Security and Privacy in Fog Computing: Challenges,” IEEE Access, vol. 5, pp. 19293–19304, 2017, doi: 10.1109/ACCESS.2017.2749422.

[6] R. Zhang, H. Ma, and Y. Lu, “PT US CR,” J. Syst. Softw., 2016, doi: 10.1016/j.jss.2016.12.018.

[7] E. Conrad, S. Misenar, and J. Feldman, “Domain 5: Identity and Access Management (Controlling Access and Managing Identity),” CISSP Study Guid., pp. 293–327, 2016, doi: 10.1016/b978-0-12-802437-9.00006-0.

[8] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges,” Cybersecurity, vol. 2, no. 1, 2019, doi: 10.1186/s42400-019-0038-7.

[9] K. Vijayalakshmi and V. Jayalakshmi, “Shared Access Control Models for Big data : A Perspective Study and Analysis,” I Pandian A.P., Palanisamy R., Ntalianis K. Proc. Int. Conf. Intell. Comput. Inf. Control Syst. Adv. Intell. Syst. Comput. vol 1272. Springer, Singapore. https//doi.org/10.1007/, 2021.

[10] A. Markandey, P. Dhamdhere, and Y. Gajmal, “Data access security in cloud computing: A review,” 2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018, pp. 633–636, 2019, doi: 10.1109/GUCON.2018.8675033.

[11] E. Sahafizadeh, “Survey on Access Control Models,” pp. 1–3, 2010.

[12] S. Salloum, J. Z. Huang, and Y. He, “Random Sample Partition: A Distributed Data Model for Big Data Analysis,” IEEE Trans. Ind. Informatics, vol. 15, no. 11, pp. 5846–5854, 2019, doi: 10.1109/TII.2019.2912723.

[13] R. S. Sandhu et al., “Role Based Access Control Models,” IEEE, vol. 6, no. 2, pp. 21–29, 1996, doi: 10.1016/S1363-4127(01)00204-7.

[14] K. Soni and S. Kumar, “Comparison of RBAC and ABAC Security Models for Private Cloud,” Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 584–587, 2019, doi: 10.1109/COMITCon.2019.8862220.

[15] M. B. and A. C. A. Nascimento, “Information Security Example Policy,” "Uni-ARBAC A Unified Adm. Model Role-Based Access Control. Int. Publ., vol. 1, pp. 218–230, 2016, doi: 10.1007/978-3-319-45871-7.

[16] E. Franco and D. C. Muchaluat-saade, “ACROSS : A generic framework for attribute-based access control with distributed policies for virtual organizations,” Futur. Gener. Comput. Syst., vol. 78, pp. 1–17, 2018, doi: 10.1016/j.future.2017.07.049.

[17] C. Morisset, T. A. C. Willemse, and N. Zannone, “A framework for the extended evaluation of ABAC policies,” Cybersecurity, vol. 2, no. 1, 2019, doi: 10.1186/s42400-019-0024-0.

[18] M. Yahiaoui, A. Zinedine, and M. Harti, “Deconflicting Policies in Attribute-Based Access Control Systems,” Colloq. Inf. Sci. Technol. Cist, vol. 2018-Octob, pp. 130–136, 2018, doi: 10.1109/CIST.2018.8596576.

[19] N. Bhatia and Vandana, “Survey of Nearest Neighbor Techniques,” vol. 8, no. 2, pp. 302–305, 2010.

[20] G. Ahalya and H. M. Pandey, “Data clustering approaches survey and analysis,” 2015 1st Int. Conf. Futur. Trends Comput. Anal. Knowl. Manag. ABLAZE 2015, pp. 532–537, 2015, doi: 10.1109/ABLAZE.2015.7154919.

[21] M. Ait, E. Hadj, M. Erradi, and A. Khoumsi, “Validation and Correction of Large Security Policies : A Clustering and Access Log Based Approach,” 2018 IEEE Int. Conf. Big Data (Big Data), no. 1, pp. 5330–5332, 2018, doi: 10.1109/BigData.2018.8622610.

[22] M. A. El Hadj, M. Ayache, Y. Benkaouz, A. Khoumsi, and M. Erradi, “Clustering-based approach for anomaly detection in XACML policies,” ICETE 2017 - Proc. 14th Int. Jt. Conf. E-bus. Telecommun., vol. 4, no. Icete, pp. 548–553, 2017, doi: 10.5220/0006471205480553.

[23] S. Guo, “Analysis and Evaluation of Similarity Metrics in Collaborative Filtering Recommender System,” 2014.

[24] M. S. Yang and K. L. Wu, “A similarity-based robust clustering method,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 4, pp. 434–448, 2004, doi: 10.1109/TPAMI.2004.1265860.

[25] M. Pratap, S. Sural, and J. Vaidya, “Managing attribute-based access control policies in a unified framework using data warehousing and in-memory database,” Comput. Secur., vol. 86, pp. 183–205, 2019, doi: 10.1016/j.cose.2019.06.001.

[26] V. C. Hu et al., “Guide to attribute based access control (abac) definition and considerations,” NIST Spec. Publ., vol. 800, p. 162, 2014, doi: 10.6028/NIST.SP.800-162.

[27] D. Lin, P. Rao, R. Ferrini, E. Bertino, and J. Lobo, “A similarity measure for comparing XACML policies,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 9, pp. 1946–1959, 2013, doi: 10.1109/TKDE.2012.174.

[28] V. Vijayalakshmi, K. and Jayalakshmi, “A Priority-based Approach for Detection of Anomalies in ABAC Policies using Clustering Technique,” no. Iccmc, pp. 897–903, 2020, doi: 10.1109/iccmc48092.2020.iccmc-000166.

[29] K. Vijayalakshmi and V. Jayalakshmi, “Analysis on data deduplication techniques of storage of big data in cloud,” Proc. 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021, IEEE, pp. 976–983, 2021.

[30] K. Vijayalakshmi and V. Jayalakshmi, “Identifying Considerable Anomalies and Conflicts in ABAC Security Policies,” Proc. Fifth Int. Conf. Intelligent Computing and Control Systems. ICICCS 2021, IEEE, pp. 1286–1293, 2021.

Multi-Robot based Control System

Atef Gharbi[1]
Faculty of Computing and Information Technology[1]
Northern Border University
Rafha, KSA
Institut National des Sciences Appliquées et de Technologie (INSAT), LISI[1]
Université de Carthage
Tunisia

***Abstract*—One of the most important challenge in Robotic Flexible Manufacturing Systems (RFMS) is how to develop a Multi-Robot based control system in which the robot is able to take intelligent decision to a changing environment. The problematic is how to ensure the flexibility with the proposed multi-robot based control system based on triggering strategies. The flexibility of the whole system is expanded by the capacity of the flexible robots to effectively ensure tasks assigned to it. Through this paper, three contributions can be presented: (i) the RFMS based Control Architecture by presenting in details the main components and methods, (ii) the planning model, and (iii) the different levels of flexibility in RFMS.**

***Keywords*—Robotic Flexible Manufacturing Systems (RFMS); multi-robot based control system; RFMS control architecture; planning model; flexibility**

I. Introduction

Nowadays, Flexible Manufacturing Systems (FMS) are facing widely frequent market changes determined by world-wide competition, new customers' requirements, continuous evolution of software and hardware, and the rapid introduction of new products [1]. Flexible Manufacturing Systems must ensure high quality products at acceptable costs and react rapidly to new market and products changes [2]. FMS can meet product changes, but they cannot respond to structural changes. In fact, the manufacturing systems are not able to face the dynamic changing environment due to their static control structure [3-5]. To react rapidly to the quickly changing environment, Robotic Flexible Manufacturing Systems (RFMS) based on multi-robot control system is considered as a good solution having the properties such as adaptability and flexibility [6]. Multi-robot based control system is an emerging solution which is becoming more and more popular as it helps to decentralize the decision in the control system [7]. Nowadays, there is a huge number of research activities that has been approved in this sense [8-10]. It is basic for Robotic Flexible Manufacturing Systems to have capacities such as autonomy, flexibility and adaptability. The RFMS framework based on Multi-Robot based control is intended to meet these criteria. A flexible manufacturing system can be applied either on static or dynamic system [11]. Robotic Flexible Manufacturing Systems can be used in many fields such as: medicine [12, 13], thermodynamic domain [14], optimal numerisation [15], motion [16], assembly system [17], automotive [18], fuzzy system [19].

The Robotic Flexible Manufacturing Systems can be based on either customisation or design. Robotic Flexible Manufacturing Systems based on customisation means satisfy the customers' needs during the design process of manufacturing systems leading to more time and effort before completing it [20].

Robotic Flexible Manufacturing Systems based on design means the ability of the change to obtain new robotic manufacturing systems based on existing ones as required, simply and economically [21]. The flexibility in design permits generating new manufacturing systems effortlessly by ensuring the required modifications from the existing ones, and the development cost can be significantly decreased [22].

In this paper, the architecture as well as the behaviour of intelligent flexible robots are presented. Therefore, the contributions are based on the following operations: (1) Firstly, design of flexible software architecture especially for a flexible robot. (2) Secondly, specification of the planning model ensured by the flexible robot. (3) Thirdly, the different levels of flexibility in Robotic Flexible Manufacturing System. To approve these contributions, the proposed methodology is applied to a benchmarking system.

Step 1: Define the high-level architecture of the RFMS.

On the basis of the control and flexibility objectives, the multi-robot control system is designed till the single controlled device and the related automation tasks. It is important for the system to be designed in a way ensuring capabilities such as flexibility. To do so, the multi-robot control system is conceived to incorporate several self-flexible levels to respond quickly to any changes occurring in the environment.

Step 2: Define the planning ability.

In the Multi-Robot Based Control System, the planning ability is considered as a very important point to study that's why it is defined how it is implemented. In fact, a plan is considered as a state-transition model.

Step 3: Define the flexibility ability.

In the Multi-Robot Based Control System, a Flexible Robot is defined what means. After that, a study on how the Flexible Robot ensures the flexibility. In order to cover a wide range of the production policies, the flexible robot must ensure several flexibility levels that can be categorized in the following ways: the product family, the product variant, the plan, the task, the skill, the failure, the production control, the adaptation and the configuration. These different levels of flexibility will be detailed later in the paper.

The remainder of this paper is organized as follows: Section 2 introduces the state of art. Section 3 presents the production system benchmark used as running example. Section 4 describes the Multi-Robot based Control Architecture. Section 5 defines the planning model. The Section 6 presents the different levels of flexibility in RFMS. Finally, the conclusion and future work are summarized in the last section.

II. State of the Art

To define well a robot control system, a special attention is given to its architecture. A huge number of research papers was presented to define it. The first classification is based on Knowledge Utilisation based on the way the robot uses its knowledge to perform action. In this first classification, there are Competitive approach and Collective approach. The competitive approach is based on the use of a single criterion to take decision [23]. The Collective approach enables to take in consideration many criteria to make decision [24]. The competitive approach can be categorized into five types, namely lookup-based [25], finite state machine [26-27], priority-based or hierarchical-based [28], goal-based [29] and utility-based competitive approaches [30].

The second classification is based on Knowledge Design. The intelligent robot can be represented through a defined architecture that can be deliberative, reactive or hybrid. The deliberative architecture (called also hierarchal architecture, top-down, knowledge-based approach, or explicit-based approach) is the most used in the artificial intelligence [31]. The deliberative architecture consists of vertical layers where each layer is based on the data sent by the previous one. In general, a robot senses the environment, plans and executes to achieve a goal.

The reactive architecture (called also bottom-up, behaviour-based architecture or implicit-based approach) is based on a mapping between perception (provided by sensors) and action. The reactive architecture is considered as horizontal architecture where the different behaviors can be executed in parallel [32].

The hybrid architecture is more commonly used especially to control robot as the reactive aspect permits to take action in real-time to the perception of the dynamic environment and the deliberative aspect enables to plan future actions to satisfy a goal [33].

Each architecture has its own strengths and weaknesses. The deliberative architecture is likely to have higher computational cost than the reactive approach due to data sent between vertical layers. In addition, the deliberative approach is more complicated due to a complete knowledge has to be provided. However, the reactive architecture is less flexible than the deliberative one because behaviors cannot be modified as much as in the deliberative architecture (although it is considered easy to implement). Therefore, purely deliberative and reactive architecture are not considered suitable for a complex system. In this context, the hybrid architecture gets strengths as well as weaknesses of the two approaches, thus why we focus on how to balance the two approaches in this paper.

III. Production System Benchmark

As much as possible, the contribution will be illustrated with a simple current example called RARM [34]. It is described informally, but it will be used as an example of the various formalisms presented in this article. The production system benchmark RARM is depicted in Fig. 1.

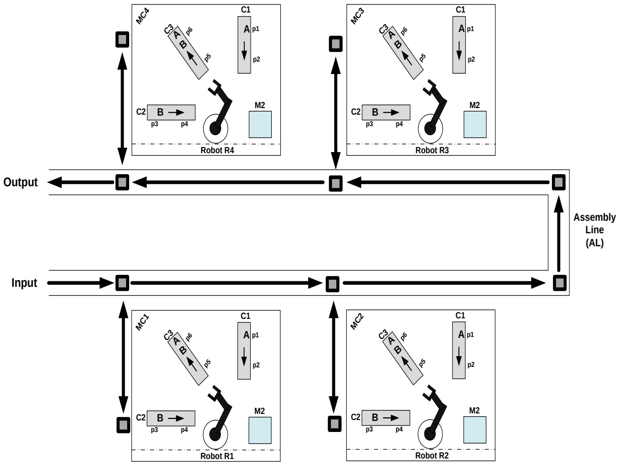


Fig. 1. The Production System Benchmark RARM.

The whole manufacturing system is divided into two main parts (Fig. 1): Assembly Line (AL) and Manufacturing Cell (MC). The Manufacturing Cell named as MC1, MC2, MC3 and MC4 are connected through the Assembly Line (AL) to enable the flexible robots moving from one MC to another.

Each Working Place contains three conveyors (C1, C2 and C3), a processing–assembling unit (machine M), a flexible robot R and additional sensors. Workpieces to be treated as they come irregularly one by one. The workpieces of Type A are carried via the conveyor C1, and the workpieces of Type B, via the conveyor C2. Only one workpiece can be on the input conveyor. The flexible robot is used to load and unload workpieces between the processing–assembling unit and the storage equipment (Input/Output). Firstly, the flexible robot carries a workpiece from the Input storage to the processing–assembling unit, which is processed by the machine M. After processing, the flexible robot transports the finished workpiece to the Output storage.

IV. High-Level Robot-based Control Architecture

The traditional methodology in designing robots has been to design the hardware and the software according to what it should do. Traditional robots can execute specific tasks, but they are not flexible, and therefore applications assigned to them depend on their physical structure and their controller abilities. Creating Flexible Robotic Manufacturing System is facing hardware and software challenges.

While existing survey papers on Flexible Robotic Manufacturing System have studied the architecture and hardware feature of robots, in this paper, a special attention is given to the challenging issues emerged when developing flexible robots. Thus, the main problem to resolve is arising when the flexible robot perform tasks through some flexibility abilities.

To react rapidly to the quickly changing environment, it is basic for the framework to have such capacities as adaptability and flexibility. The Multi-Robot based control framework is designed to meet these needs. To ease the system flexibility, robots would be itself flexible. To do so, every Flexible Robot belonging to the Multi-Robot system has its own Goal to achieve and can generate a plan associated to this goal. This policy helps the Flexible Robot to select an appropriate plan of tasks to be executed.

In Fig. 2, a Flexible Manufacturing System Meta-Model is presented. Each Product is assumed to have its own Family. A Product_Variant is considered as a specific case of Product_Family. To ensure a Product_Variant, a list of Plan has to be executed and is composed of Task_Manager. Each Task_Manager has some inputs which are Events and uses some Resources, perceive data through Receptors and execute commands by Effectors. The flexible robot needs to have some specific Skill_Manager to execute well a task.

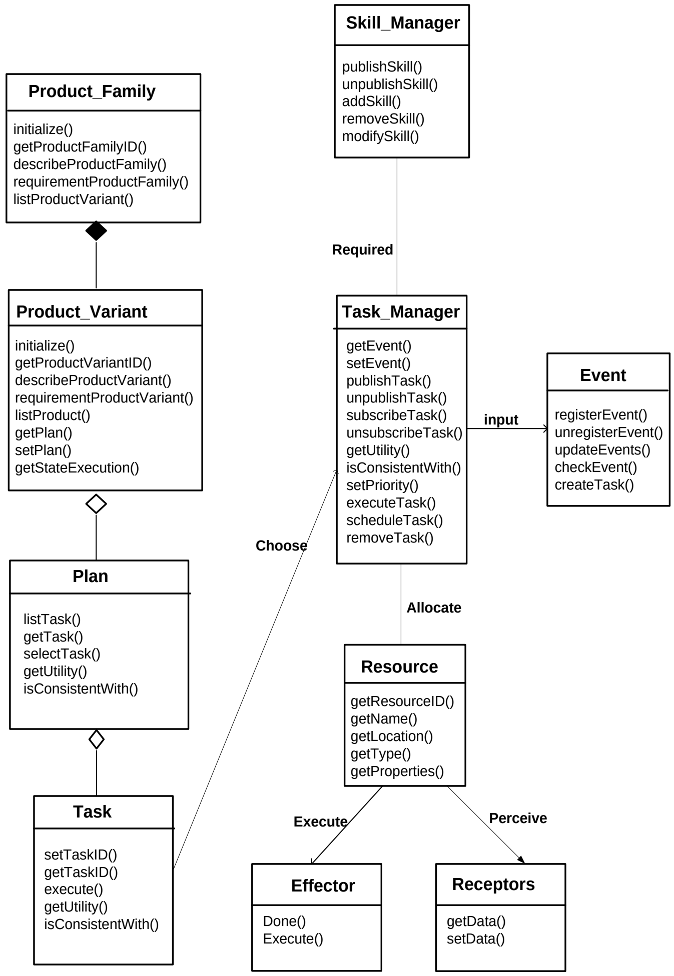


Fig. 2. Conceptual Meta-Model for a Flexible Manufacturing System.

Where

- Product_Family: It is a set of similar products having common tasks. In the production system RARM, it is possible to produce two potential product families simultaneously. To produce these product families, three types of machines need to be installed: Drill, Load, and Assembly. All of these machine types have a modular structure that allows adding/removing services. Based on these services, each machine type can have different configurations with different abilities and/or skills. Therefore, the flexible robot chooses the right machine configuration and decides the best production policy based on the machine availability and their cost structures.

Running example: In the benchmarking production system, two types of product families are defined: product family type and product family type. It is possible to process the two potential product families simultaneously. The product family type is treated firstly by RMC1 and then by RMC2. The product family type is handled firstly by RMC2 and then by RMC3. As for RMC1 (resp. RMC2, RMC3), it consists of a drilling machine (resp. milling machine, assembling machine). The buffer can store the finished workpieces and workpieces waiting to be processed.

- Product_Variant: is the same product but having different size, color, materials.

Running example: For the product family type , there are three possible product variant (i) the first production variant consists of inserting an A-work piece (through the conveyor C1) into the processing center M to be treated, then it is evacuated by the robot to the output conveyor C3; (ii) the second production variant consists of inserting a B-work piece (through the conveyor C2) into the processing center M to be treated, then it is evacuated by the robot to the output conveyor C3; (iii) the third production variant consists of inserting an A-work piece into the processing center M to be treated, then a B-work piece is added in the center and the two work pieces are finally assembled.

- Plan: The Flexible robot may have many tasks which are inconsistent. Therefore, the plan is composed of consistent tasks. The plan is composed of a set of tasks for a given product variant that can be either fixed or variable. The aim is to regroup as much as possible of tasks to be included in the same plan. Thus, the ***selectTask*** method is used to add a task in the Plan and the ***getTask*** method permits to return all the tasks related to the same Plan. Similar to Task, the ***isConsistentWith*** method is used to verify the consistency between two plans and the ***getUtility*** method enables to choose a plan among many concurrent existing Plans. If a plan has been chosen, the ***commitGoal*** method is used to create a new Goal instance.

- Task: In general the task objects are associated to the event object. There is a relation one-to-many between task and event. The methods ***setEvent*** and ***getEvent*** are used to set a link between task and associated events. To check the consistency between two tasks, the

isConsistentWith method is used. The Flexible robot has to choose between several tasks existing at the same time (which some of them can be inconsistent) based on the task's utility through the ***getUtility*** method. In some circumstance, the Flexible robot decides to select one or several consistent tasks to constitute a Plan with the use of ***selectAsPlan*** method. For each task, all exceptions are enumerated and trigger events are determined.

Running example: The set of actions is {Ci1_left, Ci1_right, Ri1_left, Ri1_right, Ci2_left, Ci2_right, Ri2_left, Ri2_right, Ci3_left, Ci3_right, Ri3_left, Ri3_right, takei1, takei2, takei3, loadi1, loadi2, loadi3, puti1, puti2, puti3, processi1, processi2}

Where:

- Ci1_left (resp. Ci1_right) means a workpiece of type A is moved to the left of conveyor Ci1 from position p1 (resp. p2) to position p2 (resp. p1).
- Ri1_left (resp. Ri1_right) means the Robot ri taking a workpiece of type A is moving to the left (resp. to the right) from the position p2 of conveyor Ci1 (resp. the processing unit Mi) to the processing unit Mi (resp. the position} p2 of conveyor Ci1).
- takei1 (resp. takei2, takei3) means the Robot ri is currently taking a workpiece of type A (resp. B , AB).
- loadi1 (resp. loadi2, loadi3) meansthe fact of loading a workpiece of type A (resp. B , AB).
- puti1 (resp. puti2, puti3) means the Robot ri is currently putting the workpiece of type A (resp. B , AB).
- processi1 (resp. processi2) means the fact of processing a workpiece of type A (resp. B).
- Event: the Flexible robot can update its knowledge about the environment through sensors. The Flexible robot can register to a specific event (this is done by the ***registerEvent*** method) or unregister (through the ***unregisterEvent*** method). The events associated to the robot are considered independent. Whenever the Flexible robot receives a new event, it checks firstly if there is a need to create a new task (this is done by the ***checkEvent*** method). If the condition is satisfied, a new task is created (this is ensured by the ***createTask*** method). All tasks arise from the Flexible robot's perception.
- Resource: Each manufacturing system is composed of a set of resources (e.g., machines, tools, grippers, conveyors, transport devices, etc.). Each resource performs a distinct function. It is possible to find a pool of more than one resource that has the same function.
- Receptor: the flexible robot knows its environment through sensors. Thus, the data provided by the sensors present the robot's vision of its environment. The perception parameters have to be defined and the robot must know how to interpret the data.

Running example

- The sensor sens1 (respectively sens2) is used to verify if there is a workpiece at the position p1 (respectively the position p2) on the conveyor C1;
- The sensor sens3 (respectively sens4) checks for the existence of a workpiece at the position p3 (respectively the position p4) on the conveyor C2.
- Effector: the flexible robot can execute the task using the effector. For each effector, a behavior is proposed to judge the requests to it.

Running example

- The effector act1 (respectively act2, act3) ensures the movement of the conveyor C1 (respectively C2, C3);
- The effector act4 rotates a robotic agent;
- The effector act5 elevates the robotic agent arm vertically.

V. Planning Model

The planning model means how the flexible robot should act to decompose the problems into subproblems to obtain the whole solution that the flexible robot must apply. The planning model of the flexible robot leads to a very huge number of possibilities which the flexible robot will have to take in consideration again in order to retain only the valid possibilities that should be kept.

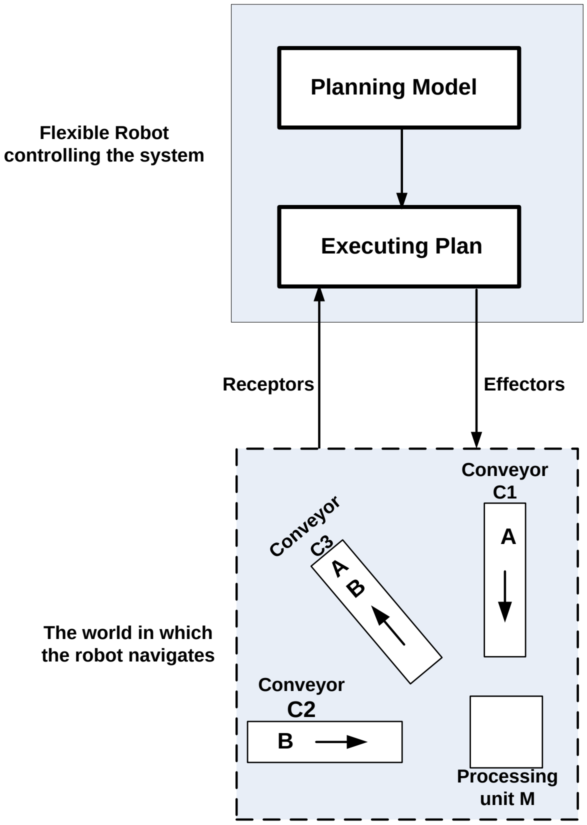


Fig. 3. The Conceptual Planning Model.

The planning model is based on two necessary elements: the receptors (i.e. a set of sensors to get data about the external environment in which the flexible robot is existing) and the effectors (i.e. a set of actuators to realize the flexible robot's tasks) [35]. Fig. 3 shows a conceptual model of the flexible robot including two components: the planning model, and the executing plan [36, 37]).

To be more specific, the planning model is based on state-transition model where Σ representing the world is a finite state-transition system, i.e., a triple Σ = (S; A; γ), where S is a finite set of states, A is a finite set of actions, $\gamma : S \times A \rightarrow 2^S$ is a state-transition function. If (s, a) = $\varnothing$; then it is said that a is not applicable to s or not executable in s.

Given a state transition system Σ , the aim of planning is to determine which actions to execute to which states in order to realize some objectives when starting from a given situation. A plan is a solution that determines the appropriate actions to reach the goal. The objective can be specified by a goal state s_g or a set of goal states Sg. The objective can be obtained by any sequence of state transitions that ends at one of the goal states. The planning model necessitates the descriptions of , the initial state before applying the plan, and the desired objectives (e.g., to reach a set of states that satisfies a given goal condition). Therefore, the planning model's objective is to produce a plan (i.e., an ordered finite sequence of actions) that puts Σ into any one of some finite set of states Sg.

More formally, the plan π is any sequence of actions π = (a1 , … , ak), where k $\geq$ 0.

The length of the plan is $|\pi|$ = k is equal to the the number of actions. If π 1 = (a1 , … , ak) and π 2 = (a'1, … , a'j) are plans, then their concatenation is the plan π 1. π 2 = (a1 , … , ak, a'1, … ,a'j). The state produced by applying π to a state s is the state that is produced by applying the actions of π in the order given.

The plan π is executable in a state s0 if there is a sequence of states (s0; s1; … ; sn) such that for i = 1; … ; n,

$s_i = \gamma(s_{i-1}, a_2)$. In this case it is said that (s0; s1; … ; sn) is π's execution trace from s0, and $\gamma(s_0, \pi) = s_n$ is defined. If sn satisfies the goal g, then it is said that π is a solution for the planning problem P = (O; s0; g).

The quality of a plan is measured by length, where the shorter of two plans is better under the same satisfaction degrees.

An action or a plan posts a set of goals G = {g1; g2; ...; gn}. This invokes the following process:

1) Loop over each goal gi:

a) Determine the set of plans filling the goal gi.

b) Keep only the plans which pre-conditions are satisfied.

c) For each remaining plan, check its mandatory resources.

2) Order the different goals gi according to its priority.

3) For every goal gi in order of priority.

a) If only one plan P realizes gi

then apply P

else // several plans

b) order each plan achieving the goal gi based on the length of the plan and how many resources it uses.

c) Choose the plan having the highest scoring.

Running example

Giving (S, A, Gs) where S = {si , i=1...n } is a set of states, A= {Ci1_left, Ci1_right, Ri1_left, Ri1_right, Ci2_left, Ci2_right, Ri2_left, Ri2_right, Ci3_left, Ci3_right, Ri3_left, Ri3_right, takei1, takei2, takei3, loadi1, loadi2, loadi3, puti1, puti2, puti3, processi1, processi2 i=1...n } is a set of actions, and Gs is the problem goal.

If the goal g = {workpiece in the processing unit} and the robot is at the initial position s1. Let:

- π 0 = (Ci1_left, takei1).
- π 1 = (load_i1, put_i1, process_i1, Ci1_right).
- π 2 = (Ci1_left, take_i1, load_i1, put_i1, process_i1, Ci1_right).

Then

π 0 is not considered as a solution because the final state is not a goal state;

π 1 is not a solution because it is not applicable to s1;

π 3 is the only solution because it is applicable to s1 and the final state is a goal state.

VI. How does the RFMS Ensure Flexibility?

The flexible robot controls the system through an event-triggering policy which means whenever an event related to the system (for example a resource failure), the flexible robot decides to take the right decision. In order to cover a wide range of the production policies, the flexible robot must ensure several flexibility levels that can be categorized in the following ways:

- Product Family flexibility means that the flexible robot is able to change-over production families to produce a new Product family which is feasible in terms of requirements (that means manufacturing facility procedure to ensure the production of each product family). In fact, Flexible Manufacturing Systems are designed to achieve several operations on many products grouped in families according to their operational requirements.

As it is illustrated in Fig. 4, each Flexible Robot has some Product to achieve. To do so, it determines the appropriate production plan and the tasks that can be executed. The internal behaviour of the Robot is defined as follow: firstly, the Flexible Robot evaluates the feasibility of the new product. If it is not feasible, then the Flexible Robot generates error, else it determines the list of tasks to be executed. Each task needs some resources. If the Flexible Robot fails in achieving this task due to some missing resources, it can ask help from other robots able to provide the requested resources which are considered as Helping Robots.

Running example the flexible robot can switch from the product family type α to the product family type β.

In [38], the authors present a methodology to group products into families depending on similarities through a modified Jaccard similarity index.

$$S_{ij} = \frac{\sum_{m \in i} \sum_{n \in j} S_{mn}}{N_i N_j} \tag{1}$$

Where

i,j families,

m (resp. n) products of family i (respectively j)

Sij degree of similarity between families i and j

Smn degree of similarity between products m and n

Ni (resp. Nj) number of products in the family i (resp. j)

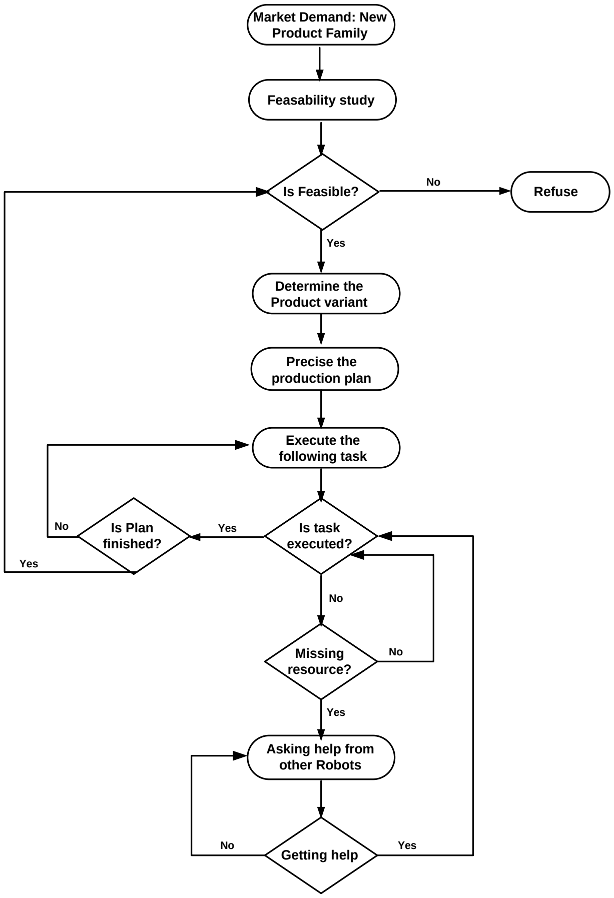

Market Demand: New Product Family
Feasability study
Is Feasible?
No
Refuse
Yes
Determine the Product variant
Precise the production plan
Execute the following task
Is task executed?
Yes
Is Plan finished?
No
Yes
No
Missing resource?
No
Yes
Asking help from other Robots
Getting help
No
Yes

Fig. 4. Flexible Robot Behavior.

- Product Variant flexibility means that the flexible robot is capable to define the different possible configurations for the same product family (i.e. alternative configurations for each product family).

Running example the flexible robot can switch from the first production variant to the second product variant in case of shortage of workpiece A.

To measure the similarity degree between two products m and n, the Jaccard similarity coefficient Smn [38] is used which is defined.

$$S_{mn} = \frac{a}{a+b+c}, 0 \le S_{mn} \le 1 \tag{2}$$

Where a represents the number of common machines used to produce both the products m and n; b defines the number of machines used to produce the product m; and c represents the number of machines used to produce the product n.

- Plan flexibility means that the flexible robot is able to define the order of execution of the tasks to ensure the same plan.

Running example the flexible robot can switch from the following plan {C2 left, take2, load2, process2} to a new plan {load1, put1, process1, C1 right}. The following algorithm obtains a valid plan constituted by a task sequence (T1, .., Ti, Tj, .., Tn) where Ti.post-condition = Tj.pre-condition.

```
Algorithm Graph_generation()
Input: Node t(acti, pre-condi, post-condi), Precedence Graph G = (T, R)
Output: Precedence Graph G = (T, R)
Add t into T
For j in 1 to length(T) do
    If (post-condj = post-condi )
            Add r = (actj , acti) into R
    End if
End for
```

- Task flexibility means that the flexible robot has the ability to manage the task switching with minimal effort.

A Task Precedence Graph G = (T, R) is used to provide a simple visual representation of a complex system by modelling the interactions and precedence relations among the different tasks where T is the set of tasks and R is the relationship between Tasks (precedence order).

Fig. 5 shows Task Precedence Graph composed of night Tasks (from T1 to T9), and illustrates how is the final production system configuration.

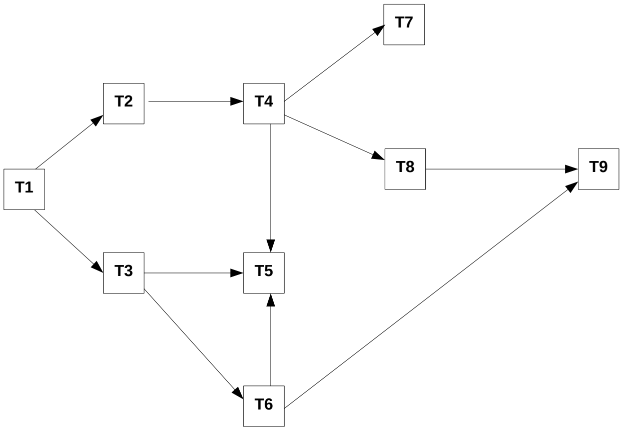


Fig. 5. Task Precedence Graph.

The list of tasks that can be executed is determined through the following algorithm.

```
***Algorithm*** Task_Choice()
Input: Precedence Graph G = (T, R)
// T is the whole tasks
// R is the relationship between Tasks (precedence order)
Output: set of tasks can be executed by end actuators
Repeat
    For each t in T do
        If (indegree(t)=0) & (t. actuator = available) then
                Return t
        End if
    End for
    If (t.state = executed) then
      t.actuator ← available
      T ← T – {t}       //Remove t from the whole tasks T
      R ← R – t.outgoingEdge //remove its outgoing edges from R
    End if
Until all tasks are executed
```

- Skill flexibility means that the flexible robot has the ability to change its different skills i.e. adding new skills, removing others, and modifying of several services composition, e.g., redesigning the services by adding a new one and eliminate others to be more flexible with the environment evolution.

Running example the flexible robot has the moving skill (forward/backward), it is possible to add on it the new skill turning (left/right).

- Failure flexibility describes the aptitude of flexible robot to deal with breakdowns and consequently guaranteeing continuation of production.

Running Example. The flexible Robot controlling the production system RARM consider many scenario in case of faults happen to physical components such as actuators, conveyors or machines.

- The first scenario involves a single conveyor C1 that transports A-work pieces to be processed by the machine unit.
- The second scenario involves a single conveyor C2 that transports B-work pieces to be processed by the machine unit.
- The third scenario involves two conveyors C1 and C2 that transports A and B-work pieces to be processed by the machine unit.

If the conveyor C1 is broken in the RARM Production System, then the flexible Robot has to apply the second scenario. If the conveyor C2 is broken in the RARM Production System, then the flexible Robot has to follow the first scenario. If the conveyor C1 and C2 are functioning well, then the flexible robot can apply the third scenario (Fig. 6).

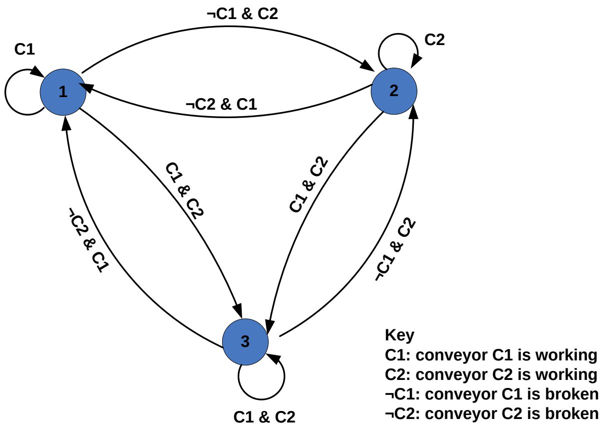

¬C1 & C2
C1
C2
1
2
¬C2 & C1
C1 & C2
C1 & C2
¬C2 & C1
¬C1 & C2
3
C1 & C2
Key
C1: conveyor C1 is working
C2: conveyor C2 is working
¬C1: conveyor C1 is broken
¬C2: conveyor C2 is broken

Fig. 6. Flexible Robot Behaviour in Case of Failure.

- Adaptation flexibility: Fig. 7(a), represents the normal case where there are: (i) two flexible robots, each one is existing in a station, (ii) Robot repository to help other robot facing problems and (iii) Gripper repository which permit robot to accommodate the product family (depending on the form and the geometry of the workpiece to be handled). This normal case is considered as the starting point upon which all adaptation scenarios are based.

In Fig. 7(b), rather than stopping the manufacturing system to repair the broken robot, the robot facing a problem can be substituted by another one. In Fig. 7(c), if the robot is broken and there is no other available robot, then the workpiece can be transferred to the second workstation to be processed.

In Fig. 7(d), a new suitable gripper is used to be convenient with the form of workpiece. The use of new grippers for the transfer of the workpiece facilitates the adaptation flexibility.

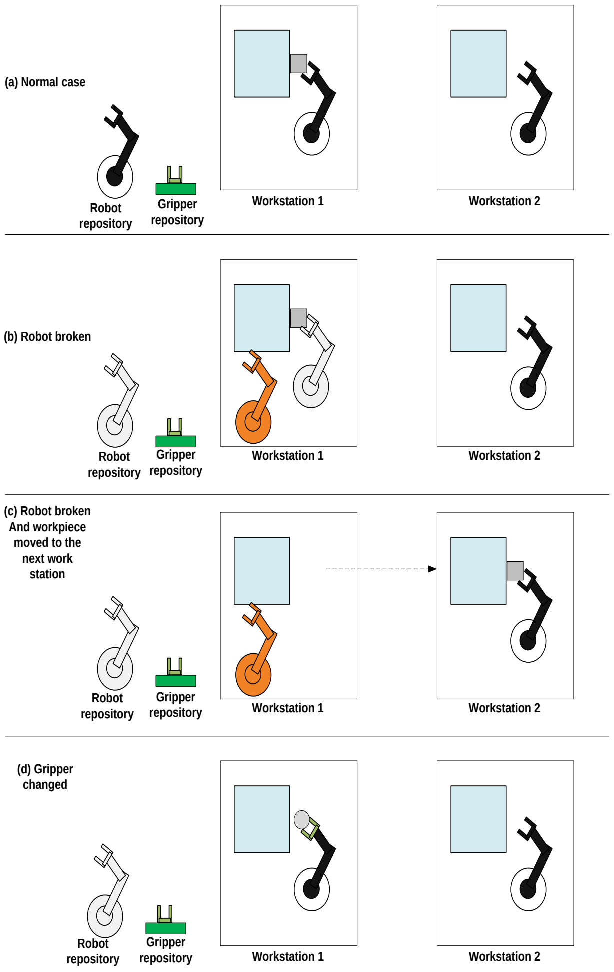


Fig. 7. Adaptation Scenarios.

- Configuration flexibility: Fig. 8 illustrates a decision graph representing all the set of possible configurations that can be executed by the flexible robot. It comprises six levels: (i) the first level represents the different product families (for example here, there are only two); (ii) the second level defines whether the Product Family is possible or not, (for each product family, there are two alternatives Possible or Not Possible); (iii) the third level specifies the different product variants that can be executed related to a specific Product Family if it is possible of course; (iv) the fourth level defines the availability to execute a product variant (for each product variant, the input indicates if is available or not); (v) the fifth level represents the different plans that can be executed for each product variant if it is available; (vi) the sixth level represents the different tasks to be executed for each plan. Each node of the tree graph is a decision point.

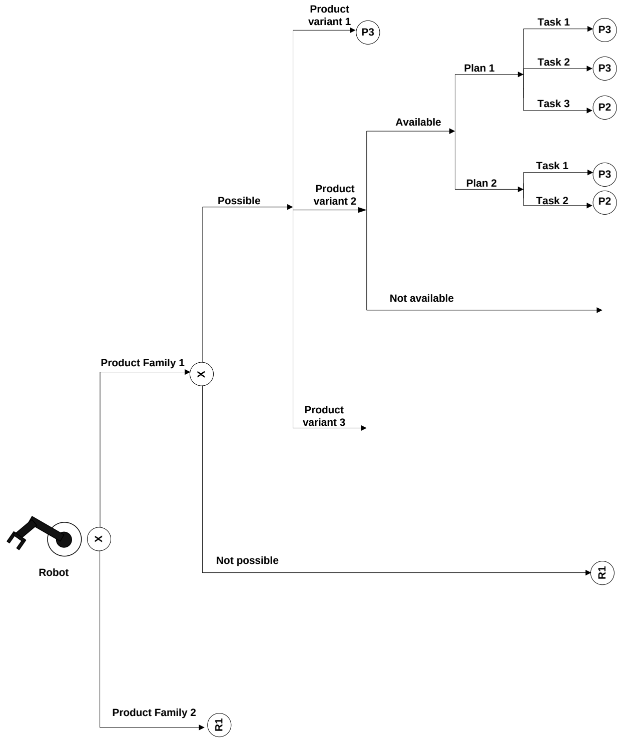


Fig. 8. Intelligent Robot Control Decision Making..

The total number of alternative solutions for each intelligent robot can be represented as:

$$R = \prod_{i} \prod_{s} \prod_{j} \prod_{a} \prod_{p} \sum_{t} R_{i,s,j,a,p,t} \qquad (3)$$

Where

- R is the total number of possible configurations for the intelligent robot,
- i, is the product family index,
- s, Boolean parameter to indicate if the product family is possible (true or false),
- j, is the product variant index,
- a, Boolean parameter to indicate if the product variant is available (true or false),
- p, is the product variant plan,
- t, is the task index.
- Ri,s,j,a,p,t, is the complete configuration that the intelligent robot can choose. Based on the above analysis for each Product family, Product variant, Plan and the identification of suitable tasks to be executed in order.

VII. Conclusion

Robotic Flexible Manufacturing Systems (RFMS) is a suitable solution to accommodate changes and meet customers' needs such as autonomous decision, control, and flexibility to react rapidly to the quickly changing environment.

Through this paper, we consider the challenge how to implement RFMS, the proposed approach presents the following contributions: (i) firstly, the general approach is designed to define the basic architecture of RFMS by presenting in details the main components and methods, (ii) secondly, the control robot is defined how to deal with the planning, (iii) thirdly, the different manners in which the flexible robot can adapt the system are proposed. The robots behave more like they are thinking, by making a decision about action selection and predicting the effects of actions. Therefore, the problem is divided into three parts: the robot-based architecture, the planning model and the flexible robot behaviour.

The future work will be as the following. The methodology can be expended to include human-computer interaction. The Multi-Robot based control system can be ameliorated to allow robots to participate in multiple collaboration at the same time.

Acknowledgment

The authors gratefully acknowledge the approval and the support of this research study by the grant no -7436-CIT-2017-1-8-F- from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A.

References

[1] Silva, A.; Ribeiro, R.; Teixeira, M. Modeling and control of flexible context-dependent manufacturing systems, Information Sciences, Volume 421, 2017, Pages 1-14, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2017.08.084.

[2] Liu, H.; Wu, W.; Su, H.; Zhang, Z. Design of optimal Petri-net controllers for a class of flexible manufacturing systems with key resources, Information Sciences, Volume 363, 2016, Pages 221-234, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2015.11.021.

[3] Gao, G.; Wang, J.; Yue, W.; Ou, W. Structural-vulnerability assessment of reconfigurable manufacturing system based on universal generating function, Reliability Engineering & System Safety, Volume 203, 2020, 107101, ISSN 0951-8320, https://doi.org/10.1016/j.ress.2020.107101.

[4] Zhang, Y.; Zhao, M.; Zhang, Y.; Pan, R.; Cai, J. Dynamic and steady-state performance analysis for multi-state repairable reconfigurable manufacturing systems with buffers, European Journal of Operational Research, Volume 283, Issue 2, 2020, Pages 491-510, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2019.11.013.

[5] Mpofu, K.; Tlale, N.S. Multi-level decision making in reconfigurable machining systems using fuzzy logic, Journal of Manufacturing Systems, Volume 31, Issue 2, 2012, Pages 103-112, ISSN 0278-6125, https://doi.org/10.1016/j.jmsy.2011.08.006.

[6] Chen Zheng, Xiansheng Qin, Benoît Eynard, Jing Bai, Jing Li, Yicha Zhang, SME-oriented flexible design approach for robotic manufacturing systems, Journal of Manufacturing Systems, Volume 53, 2019, Pages 62-74, ISSN 0278-6125, https://doi.org/10.1016/j.jmsy.2019.09.010.

[7] Saliba, M. A.; Zammit, D.; Azzopardi, S. Towards practical, high-level guidelines to promote company strategy for the use of reconfigurable manufacturing automation, Robotics and Computer-Integrated Manufacturing, Volume 47, 2017, Pages 53-60, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2016.12.002.

[8] Ferreras-Higuero, E.; Leal-Muñoz, E.; García de Jalón, J.; Chacón, E.; Vizán, A. Robot-process precision modelling for the improvement of productivity in flexible manufacturing cells, Robotics and Computer-Integrated Manufacturing, Volume 65, 2020, 101966, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2020.101966.

[9] Kontovourkis, O.; Phocas, M. C.; Katsambas, C. Digital to physical development of a reconfigurable modular formwork for concrete casting and assembling of a shell structure, Automation in Construction, Volume 106, 2019, 102855, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2019.102855.

[10] Björnsson,A.; Jonsson,M.; Johansen,K. Automated material handling in composite manufacturing using pick-and-place systems – a review, Robotics and Computer-Integrated Manufacturing, Volume 51, 2018, Pages 222-229, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2017.12.003.

[11] M.G. Abou-Ali, M.A. Shouman, Effect of dynamic and static dispatching strategies on dynamically planned and unplanned FMS, Journal of Materials Processing Technology, Volume 148, Issue 1, 2004, Pages 132-138, ISSN 0924-0136, https://doi.org/10.1016/j.jmatprotec.2004.01.054.

[12] Jason Y.K. Chan, et Al. Prospective clinical trial to evaluate safety and feasibility of using a single port flexible robotic system for transoral head and neck surgery, Oral Oncology, Volume 94, 2019, Pages 101-105, ISSN 1368-8375, https://doi.org/10.1016/j.oraloncology.2019.05.018.

[13] Michael Z. Lerner, Michael Tricoli, Marshall Strome, Abrasion and blunt tissue trauma study of a novel flexible robotic system in the porcine model, American Journal of Otolaryngology, Volume 38, Issue 4, 2017, Pages 447-451, ISSN 0196-0709, https://doi.org/10.1016/j.amjoto.2017.04.002.

[14] Jair Carlos Dutra, et Al. Development of a flexible robotic welding system for weld overlay cladding of thermoelectrical plants' boiler tube walls, Mechatronics, Volume 24, Issue 5, 2014, Pages 416-425, ISSN 0957-4158, https://doi.org/10.1016/j.mechatronics.2014.03.002.

[15] Forbes, J.; Damaren, C. Design of optimal strictly positive real controllers using numerical optimization for the control of flexible robotic systems, Journal of the Franklin Institute, Volume 348, Issue 8, 2011, Pages 2191-2215, ISSN 0016-0032, https://doi.org/10.1016/j.jfranklin.2011.06.013.

[16] Gabriel G. Kost, Ryszard Zdanowicz, Modeling of manufacturing systems and robot motions, Journal of Materials Processing Technology, Volumes 164–165, 2005, Pages 1369-1378, ISSN 0924-0136, https://doi.org/10.1016/j.jmatprotec.2005.02.186.

[17] Giulio Rosati, Simone Minto, Fabio Oscari, Design and construction of a variable-aperture gripper for flexible automated assembly, Robotics and Computer-Integrated Manufacturing, Volume 48, 2017, Pages 157-166, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2017.03.010.

[18] Yin, G. et Al. Flexible punching system using industrial robots for automotive panels, Robotics and Computer-Integrated Manufacturing, Volume 52, 2018, Pages 92-99, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2017.11.002.

[19] G. Nagamani, Young Hoon Joo, G. Soundararajan, Reza Mohajerpoor, Robust event-triggered reliable control for T-S fuzzy uncertain systems via weighted based inequality, Information Sciences, Volume 512,

[20] 2020, Pages 31-49, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.09.034.

[21] Qiao G, Lu RF, McLean C. Flexible manufacturing systems for mass customisation manufacturing. International Journal of Mass Customisation. 2006 Jan 1;1(2-3):374-93.

[22] Zheng, C. et Al. Survey on Design Approaches for Robotic Manufacturing Systems in SMEs, Procedia CIRP, Volume 84, 2019, Pages 16-21, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2019.04.183.

[23] Wilms, M. et Al. Development of a decision logic for the selection of a flexible robotic system for the automated manufacturing in tooling, Procedia CIRP, Volume 81, 2019, Pages 435-440, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2019.03.075.

[24] Ajeil, F. ; et al. Multi-objective path planning of an autonomous mobile robot using hybrid PSO-MFB optimization algorithm, Applied Soft Computing **2020**, 89, Article 106076

[25] Seeja, G.; et al. A Survey on Swarm Robotic Modeling, Analysis and Hardware Architecture. Procedia Computer Science **2018** , 133, 478-485

[26] Erdem, O.; Carus, A.; Erdem, H.; Carus, A.; Le, H .Large-scale SRAM-based IP lookup architectures using compact trie search structures. Computers & Electrical Engineering **2014**, 40, 1186-1198.

[27] Graaf, B.; Weber, S.; Deursen, A. Model-driven migration of supervisory machine control architectures. Journal of Systems and Software **2008**, 81, 517-535.

[28] Yulan, H. ; Qisong, Z.; Pengfei, X. Study on Multi-Robot Cooperation Stalking Using Finite State Machine. Procedia Engineering **2012**, 29, 3502-3506.

[29] Dai, Y.; et al. A switching formation strategy for obstacle avoidance of a multi-robot system based on robot priority model. ISA Transactions **2015** , 56, 123-134

[30] Kuhner, D.; et al. A service assistant combining autonomous robotics, flexible goal formulation, and deep-learning-based brain–computer interfacing. Robotics and Autonomous Systems **2019**, 116, 98-113

[31] Romero, A.; et al. Simplifying the creation and management of utility models in continuous domains for cognitive robotics. Neurocomputing **2019**, 35311, 106-118.

[32] Lemaignan, S.; Warnier, M. ; Sisbot, E. ; Clodic, A. ; Alami, R. Artificial cognition for social human–robot interaction: An implementation Artificial Intelligence **2017**, 247, 45-69.

[33] Baklouti, E.; Ben Amor, N.; Jallouli, M. Reactive control architecture for mobile robot autonomous navigation, Robotics and Autonomous Systems, Volume 89, 2017, Pages 9-14, ISSN 0921-8890, https://doi.org/10.1016/j.robot.2016.09.001.

[34] Yu, C.; et al. Onboard system of hybrid underwater robotic vehicles: Integrated software architecture and control algorithm. Ocean Engineering **2019**, 187, Article 106121.

[35] Gharbi, A. A Social Multi-Agent Cooperation System Based on Planning and Distributed Task Allocation, Information, 11(5) 271; doi:10.3390/info11050271 (2020).

[36] Gharbi, A. Five Capabilities Model Applied to Multi-Robot Systems. International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC), 7(1), pp.57-88, 2015

[37] Gharbi, A.; Gharsellaoui, H.; Ben Ahmed, S. Multi-Agent Control System, ICSOFT, EA, 2014:117-124.

[38] Au, T.C., Kuter, U. and Nau, D., 2008, May. Planning for interactions among autonomous agents. In International Workshop on Programming Multi-Agent Systems (pp. 1-23). Springer, Berlin, Heidelberg.

[39] Anunciene Barbosa Duarte, et al. Genetic diversity between and within full-sib families of Jatropha using ISSR markers, Industrial Crops and Products, Volume 124, 2018, Pages 899-905, ISSN 0926-6690, https://doi.org/10.1016/j.indcrop.2018.08.066.

ICS: Interoperable Communication System for Inter-Domain Routing in Internet-of-Things

Bhavana A[1]
Research Scholar, Department of Computer Science and Engineering, VTU, Belagavi, Karnataka, India

Nandha Kumar A N[2]
Professor, Department of Computer Science and Engineering, GSSS, Mysuru, Karnataka, India

***Abstract*—The Internet-of-Things consists of heterogeneous smart appliances connected by global network with self-configuring capabilities requiring interoperable communication schemes while performing inter-domain routing. A review of existing interoperable approaches shows that there is still a large scope of improving IoT interoperability. The proposed system introduces Interoperable Communication System (ICS) by developing a novel inter-domain routing in IoT using two schemes. Preemptive and Non-Preemptive Communication scheme targets mainly emergency-based routing, which demands faster transmission, and dedicated transmission, demanding accountability in communication. A simulation study carried out for the proposed system shows that it offers approximately 90% reduced delay, 57% increased packet delivery ratio, and 98% faster processing time when compared with existing approaches to accomplish interoperability in IoT.**

***Keywords*—Internet-of-Things (IoT); interoperability; heterogeneous; gateway protocol; inter-domain routing**

I. Introduction

The term Internet-of-Things (IoT) refers to technological advancement, which connects various virtual and physical things using the internet's existing infrastructure [1]. It consists of network infrastructure globally-connected, characterized by self-configuring capabilities considering interoperable communication systems [2]. IoT devices are generally considered smart appliances integrated with different platforms to assist in operating various applications, e.g., transportation, utilities, agriculture, healthcare, commerce, and industrial buildings [3]. In this aspect, the concept of interoperability is a much-discussed topic under the roof of research and development in IoT. If two IoT platforms are incompatible, then an application's operational features cannot be fully established [4]. The concept of IoT interoperability is perceived in the form of a platform, semantic, syntactic, networking, and device interoperability [5]. In the present time, the state of interoperability in IoT is managed using different approaches. The primary approach is to work on gateways and adapters that mechanizes intermediate software and tools to bridge the communication among the IoT devices with respect to different standards and data [6]. Existing gateway protocol in IoT offers one gateway to communicate with others. Still, it suffers from significant scalability issues, especially in multiple and massive smart appliances, which results in higher design complexity. The second approach uses overlay-based techniques to integrate the actuators/sensors with other IP-based objects to carry out seamless operation [7]. It can also be carried out using a virtual networking system that permits device-based communication with an end-to-end approach; however, it also suffers from scalability issues. The third approach uses different networking technologies using IP-based approaches, Software-defined approaches, network function virtualization, fog computing, etc. [8]-[10]. The fourth approach uses service-oriented architecture to offer syntactic interoperability among all smart appliances [11]. The fifth approach is to make use of semantic web technologies [12]. Apart from this, various other research work is being carried out to investigate interoperability issues in IoT. It is still in progress, and no definitive solution has yet arrived. There are various problems associated with interoperability from existing research trends which are as follows: i) existing trends of the solution is more inclined towards considering specific case study which is somewhat unpractical from practical IoT deployment, ii) in the presence of a large number of smart appliances, it is eventual that the communication has to be carried out using the application-domain approach, which is not considered much in existing trends of research, iii) although, all the major investigation towards interoperability of gateway node, they have ignored the consideration of amending the search and data translational services when gateway communicates with multiple application domains.

Existing studies in IoT towards interoperability focuses on specific communication environment without considering capabilities of devices included in it. This is a significant limitation of existing system. Therefore, this paper introduces a simple yet efficient inter-domain routing scheme in IoT that emphasizes achieving interoperability. The investigation contributes to evolving a new methodology where heterogeneous communication systems are considered while performing inter-domain routing. This paper's contribution is mainly to introduce a simplified framework that can support interoperability of communication systems in IoT, not current times. The organization of the paper is as follows: Section II discusses existing research approaches towards interoperability in IoT, Section III discusses identified research problems, Section IV discusses proposed research methodologies, Section V discusses system design along with algorithm discussion. Result analysis is carried out in Section VI, while the paper's conclusive remarks are provided in Section VII and future work description in Section VIII.

II. Existing Approaches

At present, various approaches have been implemented to address multiple ranges of issues in IoT [13]. This section specifically discusses the issues about interoperability in IoT. The recent work carried out by Xu et al. [14] has presented a technique for improving throughput performance, specifically concerning industrial IoT systems using cognitive networks. The study has used a convex optimization process in order to address resource allocation issues in industrial IoT. The technique also considers scheduling of IoT devices in order to achieve better fairness control. A similar direction of study towards resource provisioning to achieve interoperability in IoT is also presented by Zhou et al. [15]. This approach makes use of the Lyapunov optimization scheme with delay awareness. Without having any form of dependencies towards predefined information of statistical data of cloud system, this technique reports provisioning of IoT application of different types. The approach finally claims a cost-effective operation considering practical world traces of traffic information. The current study also emphasizes security and interoperable conditions in IoT where signcryption is used for the multi-receiver system over the mobile system in IoT (Qiu et al. [16].

Another study carried out by Behera et al. [17] has developed an enhanced communication scheme focusing on the heterogeneous network in IoT. Considering the wireless sensor network case study, this implementation mainly deals with selecting cluster-head for performing data aggregation in the heterogeneous network in IoT. However, the scheme is restricted to addressing issues associated with energy efficiency only in IoT. Considering the case study of healthcare in IoT, the work carried out by Ray et al. [18] has used broker services for message queueing in order to perform service provisioning. A unique gateway testbed has been constructed, which includes various wireless networks for usage in translational services. The complete study has been carried out in prototyping form using real-time sensors. The work carried out by Diez et al. [19] has presented a validation framework in order to assess the fact of proper alignment of data with standards of the data model. It has also been seen that messaging protocols play an essential role in interoperability in IoT. Al-Masri [20] has discussed the dependency of multiple messaging protocols to establish proper communication among heterogeneous devices in IoT.

There is also dedicated research being carried out towards assessing the testbed of IoT. One such research work is carried out by Lanza et al. [21] where experimentation in the form of services has been presented for IoT. Existing studies have also been carried out towards assessing the trustworthiness of gateway systems in IoT, as seen in the work of Fraile et al. [22]. The approach discusses security architecture for device drivers along with faster visibility of the node. Yang et al. [23] have presented an optimization scheme towards industrial IoT operation where the Poisson process is used for developing the work distribution in IoT. The adoption of convex optimization principle is used for performing analysis. A combined study towards network security and quality of service is carried out by Sood et al. [24] in order to enhance the heterogeneity in the serviced rendered by Software Defined Network in IoT. The technique also performs transformation of the controllers of heterogeneous form to the homogeneous form using mathematical modelling. Heterogeneity is also studied with respect to security over physical layers in IoT as seen in work of Wang et al. [25]. The work carried out by Leu et al. [26] have emphasized over achieving stability in messaging services in IoT in order to enhance the heterogeneity in Service-Oriented Architecture. The technique makes use of shortest processing time in order to carry out scheduling of the IoT messages in web-based services.

Study carried out by Nguyen et al. [27] has used network codes in order to investigate the impact of energy consumption over IoT performance. In this study, the author has investigated the use of random linear network coding targeting to improve the throughput performance in IoT. A unique study presented by Wu et al. [28] discusses the use of tree concept in presenting sensor network over IoT considering case study of healthcare assessment. In this work, a unique tree structured is developed with indexing policies where a neural network is applied in order to perform feature engineering for facilitating interoperable operation in IoT. Such learning based approaches towards heterogeneous network are also witnessed in work of Yang et al. [29]. This approach makes use of radio frequency in order to cater up the QoS requirement using Markov decision process. The study also uses reinforcement learning scheme for further optimizing the outcome. Adoption of swarm intelligence is also reported to carry out optimization operation in IoT. Work of Ni et al. [30] have used dragonfly algorithm in order to plan the communication path considering multiple robotic system of heterogeneous form. An indepth investigation towards various optimization algorithm towards improving hardware architecture for heterogeneous IoT is carried out by Krishnamoorthy et al. [31]. Noori et al. [32] have carried out investigation towards slotted ALOHA protocol for improving the communication performance in heterogeneous IoT. Finally, Asad et al. [33] have improved the quality of service using provisioning of quality of service over multiple radio access technology using game theory concept. The idea is to offer a seamless access among heterogeneous devices in IoT.

Hence, it is seen that existing system mainly make use of heterogeneity concept in order to achieve better interoperability in IoT. Heterogeneity doesn’t necessarily address interoperability issues in IoT. A closer look into existing system shows beneficial approaches and outcome when it comes to addresses considered problem. However, all the essential problems are not considered in generalized environment of IoT. The next section discusses about research problem.

III. Limitation of Existing Studies

After reviewing the existing approaches of interoperability in IoT, following research problems has been noticed:

- Although, there are certain degree of work being carried out towards addressing interoperability problem in IoT, but they are very much specific to singular environment of communication. However, practical application in IoT demands more towards inter-domain communication scheme, which unfortunately is very less. Existing solution are highly inclined towards either network layer or device level. This is not sufficient to deal with various emergency communication systems in IoT.

- Existing system offers minimal emphasis over capabilities of smart appliances or IoT devices while working on interoperability. There is a need of evolving up with standard of communication over IoT which are not present at this moment in order to cater up the demands of communication associated with the low end devices in IoT. It is essential to understand that a practical form of solution for interoperability should not be purely dependent on available network entity. Another biggest challenge is also associated with the existing gateway system in IoT which is not compatible with the changes when a new smart appliance is added or new services are edited. Apart from this, there is an emergent need of flexible interoperability within the gateway system for boosting up machine-to-machine communication system with higher range of scalability.
- The third challenge explored in existing studies is that they consider single of two platform in the form of application domain. In the practical concept of large scale scenario of IoT deployment, there are possibilities of multiple number of application domain and IoT nodes within this domain are required to establish communication process dynamically. A better and effective approach should be highly practical with facilitation of scalable operation over the multiple number of application domains. Adding of new domain should not affect the scalable performance of the gateway node.
- There are various routing scheme in IoT; however all the existing protocols suffers from the challenges associated with fault tolerance, context awareness, presence of multiple network standards, node deployment, intermittent connectivity, multihop communication scheme. Apart from this issue, existing routing scheme cannot actually differentiate dynamic demands of the large scale distributed communication system. It works in similar way for all kinds of communication demands. Interoperability is achieved by a predefined environment in IoT which is impractical.

Hence, the statement of problem is "It is challenging to develop scalable, interoperable routing scheme which justifies all forms of traffic demands in unbiased fashion in large scale IoT environment."

IV. Research Methodology

The proposed system adopts an analytical research methodology where the idea is to construct a framework which can jointly address the both preemptive and non-preemptive communication system in IoT. The novelty of the proposed methodology is that it offers higher degree of flexibility to achieve interoperability in achieving communication in IoT. The prime aim of the proposed study is to offer a significant range of interoperability among the heterogeneous smart appliances, also called as IoT nodes. It is a continuation of our prior implementation where scalability factor has been achieved [34]. Fig. 1 highlights the architecture implemented for this purpose to develop Interoperable Communication System (ICS).

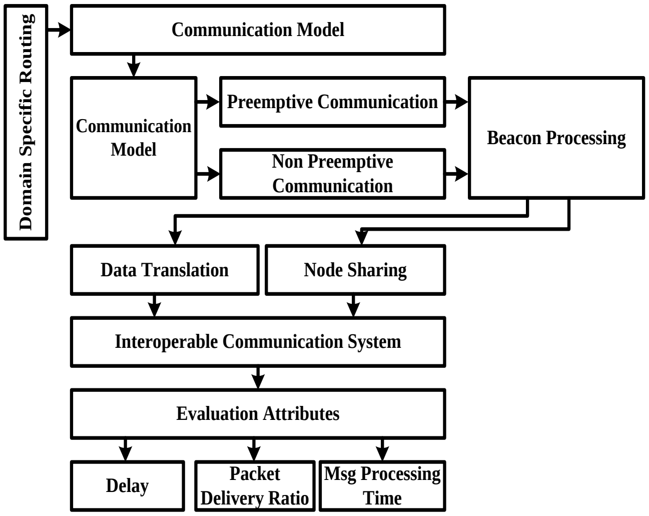


Fig. 1. Architecture of ICS.

Fig.1 highlights the composition of prominent block of operation of ICS. The first block of IoT Topology mainly incorporates IoT environment with presence of IoT nodes deployed in the form of group. The second block of Application Domain clusters all the IoT nodes within the IoT environment. The common block of IoT Routing considers the deployment of any communication protocol within application domain. Each domain is assumed to execute discrete IoT local routing. The next important block of operations is preemptive and non-preemptive communication scheme. The former scheme is used for processing emergency transmission while the later scheme is used for dedicated transmission in IoT. The idea is to develop a framework which can cater up communication demands of maximum situation in IoT. The next block of operation is Gateway Node Management which is mainly used for translational services for facilitating interoperability in the system. In case of preemptive approach, the gateway node performs data translation while in case of non-preemptive approach; addresses of the node are shared. The study outcome is finally. The further discussion of methodology of proposed system is illustrated in next section with respect to system design.

V. System Design

This part of the study illustrates about the system design as well as implementation being carried out towards developing Interoperable Communication System (ICS) in IoT. The complete implementation is carried out considering the practical environmental scenario of using different forms of smart appliances deployed in IoT environment. The study considers heterogeneity in the devices being connected where achieving interoperability in communication system is still an open-end problem. For this purpose, the complete ICS is designed considering two states of communication demands i.e. preemptive and non-preemptive stage. This section elaborates about the system design deployed for this purpose.

A. Preemptive Communication Scheme

The prime purpose of the preemptive communication scheme is to offer instantaneous communication establishment during any form of emergency circumstances (Fig. 2). It is based on the hypothesis that when a large number of heterogeneous smart appliances are connected in an IoT, either in small and large scale, achieving faster communication is quite a challenging aspect. Hence there is a need of preemptive mechanism which can ensure that an emergency communication always takes place irrespective of any situation of some of the IoT nodes owing to any reason. This scheme also considers that only few IoT nodes could possibly exhibit degradation in its performance of data forwarding either due to resource depletion or due to any miscellaneous reasons. Apart from this, it should also offer supportability of mobility condition of IoT nodes.

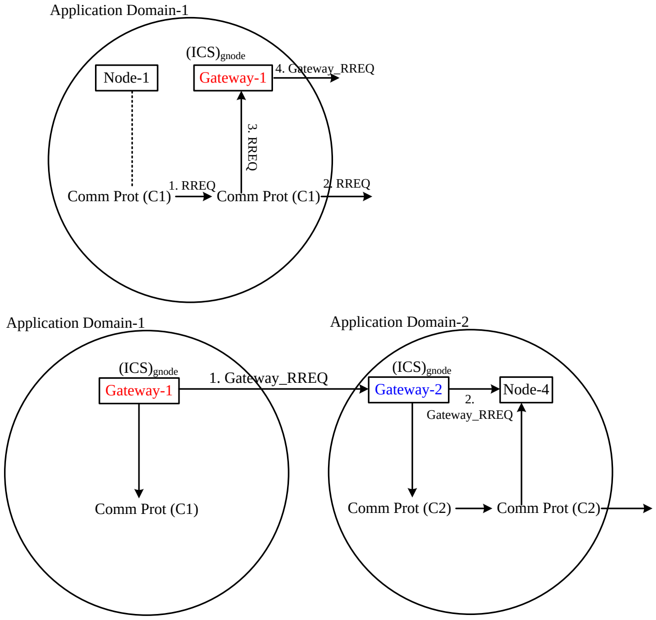


Fig. 2. Preemptive Routing Scheme in IoT.

Fig. 2 highlights a practical implementation environment where a smart appliance $node_1$ is a part of application domain α_1 which sensed an event-data and now it is required to transmit this data to $node_4$ that is part of different application domain α_2. The environment also considers usage of two different communication scheme β_1 and β_2 followed by all IoT nodes in domain α_1 and α_2 respectively. The transmitting $node_1$ broadcast the beacon for route discovery first considering the presence of an IoT gateway node γ. It should be noted that the study considers a single gateway node in one deployment zone γ which offers translation services for two application domain by offering two discrete sub-gateway services i.e. $\gamma=(\gamma_1, \gamma_2)$. This gateway node carries the addresses associated with identity and position of their respective registered smart appliances followed by establishment of connection. The proposed system constructs an algorithm which initiates processing of the request of data forwarding where gateway protocols assists in establishing preemptive routing. The algorithm implemented for this purpose is shown as follows:

Algorithm for Preemptive Communication Scheme

```
Input: $n, \gamma, \beta, Cv, f$
Output: $d$
Start
1. For i=1:$n$
2.   Formulate $\alpha$=1: $m$
3.   $\alpha_i$=[ $\alpha_1, \alpha_2, \ldots \alpha_m | m<n$]
4.   alloc $n_i \rightarrow \alpha_m, \gamma \rightarrow \alpha$
5.   For $\alpha$=1:$m$
6.     $n_{1i}$ forwards msg($C_{v1}$)
7.     $C_{v1}$ forwards msg($\gamma(\alpha)$)
8.     $\gamma(\alpha_i) \rightarrow msg_1(\gamma(\alpha_{i+1}))$
9.     $\gamma(\alpha_{i+1}) \rightarrow f(C_{v2}) \rightarrow (d)n_{2j}$
10.  End
11. End
```

The above mentioned algorithm takes the input of n (number of smart appliances), γ (gateway node), α (application domain), Cv (Communication vector), and f (function for data translation), which after processing leads to generation of d (transmitted data). The algorithm considers all the sensor nodes (Line-1) and formulates m number of application domain α (Line-2 and Line-3). It should be noted that number of application domain m is considerably less compare to number of IoT nodes. The algorithm further allocates specific number of sensors n_i to all m number of application domain (Line-4). Considering all the number of available application domain α (Line-5), the IoT node initially forwards the message in the form of communication vector C_v (Line-6). The variable msg will mean control message or beacon used for route discovery process (Line-6). This is carried out in the form of broadcasting and this information is then passed on primary gateway node γ of the parent application domain (Line-7). The parent domain accesses the information from the message to find out the location of the destination node within different application domain (Line-8). The implementation considers the fact that each application domain exercises a unique routing scheme where the gateway node assists in data translational services in cross-application domains. It is also assumed that each gateway node possess information of its respective IoT nodes. Therefore, when a request to forward data to a specific node arrives for a gateway node from different gateway node, there are two further possibilities viz. i) if the destination node is present than the data d is subjected to data translation using function f(x) (Line-9) and finally data is forwarded, and ii) if the destination node is not present than the gateway node passes the route request message to the next gateway node and this search continues until and unless the destination node is found. It should be also noted that all the gateway nodes performs an exchange of their information after a certain interval of time. This will reduce the search time as updated information will be always available.

B. Non-Preemptive Communication Scheme

This is another alternative of the routing scheme where the circumstances demands dedicated routing of the data among heterogeneous IoT devices.

This part of the implementation considers exactly the similar deployment of application domain, gateway node, and IoT devices. According to the representation shown in Fig. 3, the transmitting node node$_1$ transmits the control message for the purpose of route discovery process. When the parent gateway node receives this request, it is forwarded to neighboring gateway node of different application domain. This is done if the primary gateway node is located far from destination gateway node which has destination node. If the target gateway node is located near to primary gateway node than the message is forwarded directly to the destination gateway node. Different from preemptive communication scheme, this scheme basically emphasizes on the address information of all the nodes involved in communication process in order to offer more accountability in communication process. Another uniqueness of this technique is that it offers bidirectionality of the communication process. Apart from this, the study also contributes towards assisting in exchanging routing table information between the transmitting and destination node located in different application domain. This makes the process of updating routing operation quite faster.

The algorithm responsible for this purpose is as follow:

Algorithm for Non-Preemptive Communication Scheme

```
Input: $n$, $\gamma$, $\alpha$, $C_v$, $\lambda$, $\tau$.
Output: $d$
Start
1. For i=1:$n$
2.   Formulate $\alpha$=1: $m$
3.   $\alpha_i$=[ $\alpha_1$, $\alpha_2$, …. $\alpha_m$| $m$<$n$]
4.   alloc $n_i \rightarrow \alpha_m$, $\gamma \rightarrow \alpha$
5.   For $\alpha$=1:$m$
6.     $n_{1i}$ forwards msg($C_{v1}$)
7.     response of $C_{v1}$ is forwarded to $n_{1i}$.
8.     $C_{v1} \rightarrow \tau(\gamma(\alpha_i))$
9.     $\gamma(\alpha_i) \rightarrow \tau_1(\tau(\gamma(\alpha_i)))$
10.     $\gamma(\alpha_i) \rightarrow \lambda(C_{v2})$
11.     $C_{v2} \rightarrow \tau(n_{1i})$
12.     transmit $d$
13.   End
14. End
```

The algorithm takes the input of n (IoT nodes), γ (gateway node), α (application domain), C_v (Communication vector), λ (process of sharing address), and τ (information of address), which after processing yields an outcome of the d (transmitted data). The unique aspect of this routing scheme is that it mechanizes a method which performs embedding of the routing information to the IoT devices present in the network. In order to facilitate a dedicated communication path, the proposed algorithm ensures that each communicating nodes should have an access to the address information of all the nodes involved in path establishment via different gateways. The gateway node is responsible for sharing the address information among each other. Therefore the non-preemptive process of communication differs from preemptive communication scheme by facilitating the gateway node to carry out mutual exchanging of positional information among each other. The preliminary steps of operation are nearly equivalent to preemptive operation. Considering all the IoT nodes and application domain involved in it (Line-1 and Line-2), the proposed system initiates it process when node$_i$ in domain α_i attempts to communicate with node$_j$ in application domain α_j. The IoT node nodei transmit its control message using communication vector C_v (Line-6). The respective response is forwarded to node node$_i$ and instantly transmits the information about the address along with it to node$_i$ via the gateway node γ_i in application domain α_i. The request for routing is not forwarded to the application domain but they only share information of its addresses. The gateway node associated with the α_j share its own information as well as adjacent IoT devices after it receives the information of address for the node$_i$ in the form of response. Once this information about address is received, the gateway node share the equivalent information with an assistance of communication vector C_v that make use of method of sharing address which is further forwarded to transmitting IoT node. One of the significant characteristic of this algorithm is that it is capable of generating maximum number of addresses of all connected IoT nodes with the gateway node. Therefore, there are minimal utilization of the network resources as well as there are lesser dependencies in processing this algorithm in contrast to preemptive routing scheme. Another potential contribution of this algorithm is that it facilitates inter-domain routing specific to those nodes which are registered under established entries in more secured manner. The proposed system also performs specific way to exchange the message. The proposed approach carries out forwarding of the beacon for collecting the associated information of the IoT nodes followed by sharing of the control message along with the gateway nodes. In this process, the sequence number as well as address of IoT node is used for formulating the structure of beacon used in proposed study. The proposed system considers IP address as the address of the IoT nodes while freshness of data is represented by the sequence number.

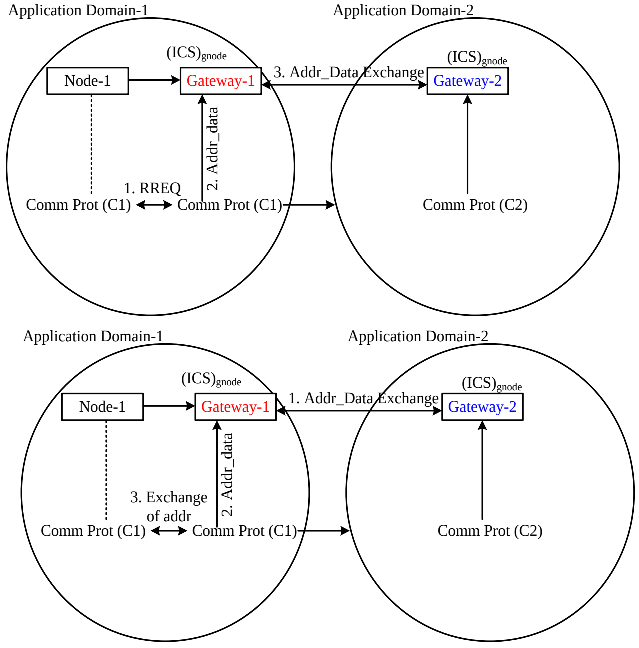


Fig. 3. Non-Preemptive Communication Scheme.

VI. Result Analysis

This section discusses about the outcome obtained after implementing the proposed ICS. The implementation of the proposed system is carried out considering 500 IoT nodes deployed in 1000 x 1000 m^2 simulation area where the transmission range is considered as 200m with assessment period of 1000s of iteration. As the proposed system mainly presents a modelling for communication scheme; therefore, it considers the equivalent standard performance parameters of delay, packet delivery ratio, and overall processing time. In order to perform a measurable assessment, the outcome of the proposed system is compared with the existing standards which are related to IoT. The first existing standard considered for processing is RFC 3221 which is responsible for performing inter-domain routing in IoT [35]. The second existing standards which are related to IoT is RFC 8352 which is responsible for implementing gateway scheme in IoT [36]. This section illustrates all outcome achieved.

A. Outcome of Delay

Delay is one of the prominent performance parameters to signify the communication effectiveness. Communication carried out in large scale scenario in presence of gateway and uneven traffic can significantly increase delay. Hence, the assessment of delay is carried out by forwarding 2500 bytes as experimental data. Fig. 4 highlights that the proposed system excels better outcome in contrast to existing approaches. The prime reason behind this is the underlying mechanism of message control in proposed ICS with availability of both the schemes. Presence of preemptive scheme always results in lower delay but at the same time, non-preemptive scheme too assists in reducing delay as once the path is set via accessing addresses, then data transmission becomes almost seamless in this dedicated path of inter-domain routing.

It is to be noted that the gateway node of proposed system contributes towards collecting individual address of the IoT devices and it performs periodic exchange of information with its adjacent gateway node. Therefore, proposed system can ascertain routing of maximum quantity of data over shorter range of time. Apart from this, the duration for obtaining information of communication vector from other IoT devices is also reduced. There is no potential overhead in the gateway as it only permits the exchange of the address information of IoT device. On the other hand, RFC 3221 offers overhead control in IoT, but there is no updating process towards the routing information in the memory of routing table. Apart from this, no information is shared with its adjacent nodes. The existing approach of RFC 8352 assists in performing gateway protocol implementation standard by opting for highly stabilized communication channel which is carried out using index of mobility criterion. It should be noted that gateway node is tracked via index of mobility to find out if they resides within the coverage of network. This calls for increasing search process which further increases the delay to a higher value. Therefore, the proposed system offers approximately 90% improved delay reduction when compared with existing standards of interdomain routing in IoT for targeting better interoperability.

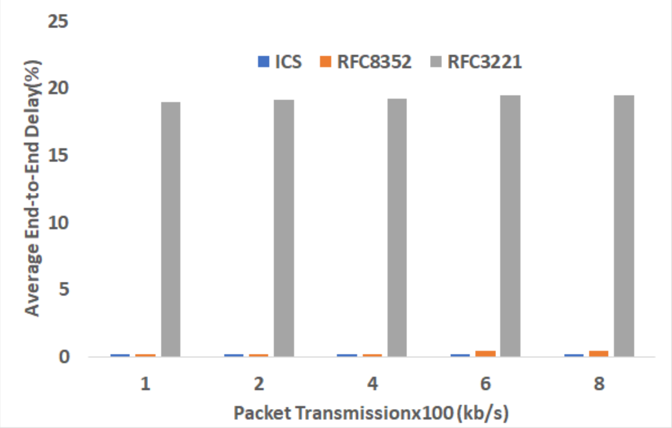


Fig. 4. Analyzed Outcome of Delay.

B. Outcome of Packet Delivery Ratio

Packet delivery ratio is considered another performance parameter which is assessed by considering amount of data which is already being forwarded to the amount of the data received by the recipient. Fig. 5 highlights that proposed system offers approximately 57% of improved packet delivery ratio in contrast to existing system of inter-domain routing. The prime reason behind this outcome is that it offers better form of data processing over the gateway node which makes the system characterized by higher interoperability. The routing tables are generated as well as updated by proposed data translational services into multiple structured which has solution to all the possible alternative communication path between transmitting and receiving IoT nodes.

Apart from this, preemptive scheme assists in node-by-node communication system where data is instantly forwarded to next node during route discovery process itself. Hence, it constructs a stabilized path whereas the path stability increases multifold when non-preemptive communication scheme is applied. However, no such operation takes place in existing system resulting in inferior packet delivery ratio.

C. Outcome of Processing Time

Processing time is one of the essential performance parameter to measure the computational complexity as well as to assess the overall response time of the system. It gives a fair idea of effectiveness of proposed model over practical network environment. The proposed evaluation considers processing time as the cumulative amount of time needed for the gateway node in order to perform processing of the data transmission from source node of one application domain to destination node of another application domain. Hence, there is higher feasibility of different ranges of processing time for different set of communicating IoT nodes. Considering multiple rounds of simulation, the assessment of the processing time is carried out to evaluate the effect of various traffic behaviour, its load, and variable data over the processing time.

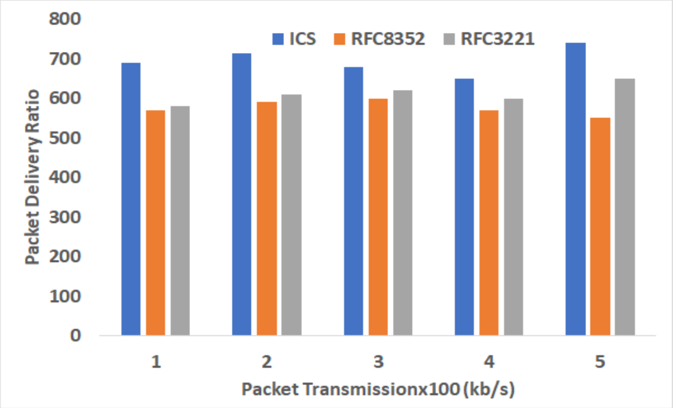


Fig. 5. Analyzed Outcome of Packet Delivery Ratio.

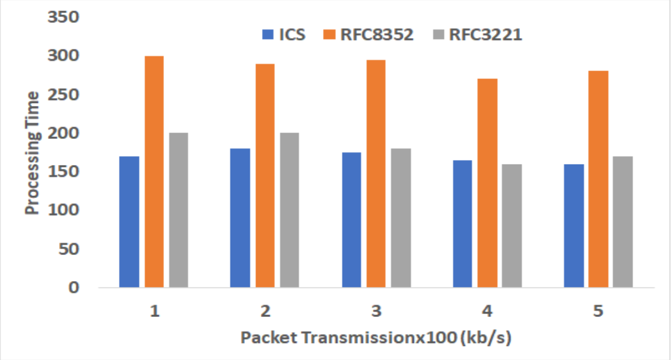


Fig. 6. Analyzed Outcome of Processing Time.

Fig.6 showcase that the proposed system accomplishes approximately 98% of reduced processing time in comparison to existing system. The justification behind this outcome is that the developed gateway system of ICS can cater up the demands of any form of dynamic environment. The proposed ICS solve the problem associated with intermittent breakage of link in dynamic topology of IoT. This is made possible when there is communication exchange between gateway nodes of different application domain. However, such forms of route management is not taken care of by existing standards in IoT. Another essential contribution of proposed system is that IoT nodes are freed from any form of duty towards strengthening the connectivity of IoT devices that is also the reason for faster response time.

VII. Conclusion

This paper presents a discussion about a novel framework of facilitating an inter-domain routing in IoT with core emphasis to accomplish full-fledge interoperability. The study uses an analytical approach where an IoT environment is deployed in such a way that the device-to-device (heterogeneous) communication is carried out using gateway. The protocol implemented is executed on gateway node where it facilities two forms of routing scheme apart from its data translation services. Preemptive scheme assists in emergency communication while non-preemptive scheme assists in offering dedicated line of communication in IoT. The study outcome shows better communication performance in contrast to existing schemes deployed in achieving interoperability. The definitive findings of the measure of its significance in the research community is that proposed system offers 90% reduced delay performance, 57% of more packet delivery ratio, and 98% of reduced processing time in comparison to existing system.

VIII. Future Work

The present research work will be continued towards achieving secure communication in data transmission in IoT in compliance of scalability and interoperability aspects too. It is necessary as the proposed system is capable of connecting a vast set of network with another network via a gateway system and hence massive number of data as well as services are exchanged. Hence, future work could be carried out using trust-based data integrity scheme along with privacy preservation.

References

[1] Alam, Sarfraz, Mohammad MR Chowdhury, and Josef Noll. "Interoperability of security-enabled internet of things." Wireless Personal Communications 61, no. 3 (2011): 567-586.

[2] Ahlgren, Bengt, Markus Hidell, and Edith C-H. Ngai. "Internet of things for smart cities: Interoperability and open data." IEEE Internet Computing 20, no. 6 (2016): 52-56.

[3] Elkhodr, Mahmoud, Seyed Shahrestani, and Hon Cheung. "The internet of things: new interoperability, management and security challenges." arXiv preprint arXiv:1604.04824 (2016).

[4] Di Martino, Beniamino, Massimiliano Rak, Massimo Ficco, Antonio Esposito, Salvatore Augusto Maisto, and Stefania Nacchia. "Internet of things reference architectures, security and interoperability: A survey." Internet of Things 1 (2018): 99-112.

[5] Ahmad, Awais, Salvatore Cuomo, Wei Wu, and Gwanggil Jeon. "Intelligent algorithms and standards for interoperability in Internet of Things." (2019): 1187-1191.

[6] Mukherjee, Saswati, and Susan Elias. "An applications interoperability model for heterogeneous internet of things environments." Computers & Electrical Engineering 64 (2017): 163-172.

[7] Jabbar, Sohail, Farhan Ullah, Shehzad Khalid, Murad Khan, and Kijun Han. "Semantic interoperability in heterogeneous IoT infrastructure for healthcare." Wireless Communications and Mobile Computing 2017 (2017).

[8] Qin, Zhijing, Grit Denker, Carlo Giannelli, Paolo Bellavista, and Nalini Venkatasubramanian. "A software defined networking architecture for the internet-of-things." In 2014 IEEE network operations and management symposium (NOMS), pp. 1-9. IEEE, 2014.

[9] Kalkan, Kubra, and Sherali Zeadally. "Securing internet of things with software defined networking." IEEE Communications Magazine 56, no. 9 (2017): 186-192.

[10] Babiceanu, Radu F., and Remzi Seker. "Cyber resilience protection for industrial internet of things: A software-defined networking approach." Computers in Industry 104 (2019): 47-58.

[11] Noura, Mahda, Mohammed Atiquzzaman, and Martin Gaedke. "Interoperability in internet of things: Taxonomies and open challenges." Mobile Networks and Applications 24, no. 3 (2019): 796-809.

[12] Rahman, Hafizur, and Md Iftekhar Hussain. "A comprehensive survey on semantic interoperability for Internet of Things: State-of-the-art and research challenges." Transactions on Emerging Telecommunications Technologies 31, no. 12 (2020): e3902.

[13] Bhavana, A., and AN Nandha Kumar. "An Analytical Modeling for Leveraging Scalable Communication in IoT for Inter-Domain Routing." In Proceedings of the Computational Methods in Systems and Software, pp. 1-11. Springer, Cham, 2018.

[14] L. Xu, W. Yin, X. Zhang and Y. Yang, "Fairness-Aware Throughput Maximization Over Cognitive Heterogeneous NOMA Networks for Industrial Cognitive IoT," in IEEE Transactions on Communications, vol. 68, no. 8, pp. 4723-4733, Aug. 2020, doi: 10.1109/TCOMM.2020.2992720

[15] Z. Zhou, S. Yu, W. Chen and X. Chen, "CE-IoT: Cost-Effective Cloud-Edge Resource Provisioning for Heterogeneous IoT Applications," in IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8600-8614, Sept. 2020, doi: 10.1109/JIOT.2020.2994308.

[16] J. Qiu, K. Fan, K. Zhang, Q. Pan, H. Li and Y. Yang, "An Efficient Multi-Message and Multi-Receiver Signcryption Scheme for Heterogeneous Smart Mobile IoT," in IEEE Access, vol. 7, pp. 180205-180217, 2019, doi: 10.1109/ACCESS.2019.2958089.

[17] T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand and A. H. Gandomi, "I-SEP: An Improved Routing Protocol for Heterogeneous WSN for IoT-Based Environmental Monitoring," in IEEE Internet of Things Journal, vol. 7, no. 1, pp. 710-717, Jan. 2020, doi: 10.1109/JIOT.2019.2940988.

[18] P. P. Ray, N. Thapa and D. Dash, "Implementation and Performance Analysis of Interoperable and Heterogeneous IoT-Edge Gateway for Pervasive Wellness Care," in IEEE Transactions on Consumer Electronics, vol. 65, no. 4, pp. 464-473, Nov. 2019, doi: 10.1109/TCE.2019.2939494.

[19] L. Diez, J. Choque, L. Sánchez and L. Muñoz, "Fostering IoT Service Replicability in Interoperable Urban Ecosystems," in IEEE Access, vol. 8, pp. 228480-228495, 2020, doi: 10.1109/ACCESS.2020.3046286.

[20] E. Al-Masri et al., "Investigating Messaging Protocols for the Internet of Things (IoT)," in IEEE Access, vol. 8, pp. 94880-94911, 2020, doi: 10.1109/ACCESS.2020.2993363.

[21] J. Lanza et al., "Experimentation as a Service Over Semantically Interoperable Internet of Things Testbeds," in IEEE Access, vol. 6, pp. 51607-51625, 2018, doi: 10.1109/ACCESS.2018.2867452.

[22] F. Fraile, T. Tagawa, R. Poler and A. Ortiz, "Trustworthy Industrial IoT Gateways for Interoperability Platforms and Ecosystems," in IEEE Internet of Things Journal, vol. 5, no. 6, pp. 4506-4514, Dec. 2018, doi: 10.1109/JIOT.2018.2832041.

[23] J. Yang, C. Ma, B. Jiang, G. Ding, G. Zheng and H. Wang, "Joint Optimization in Cached-Enabled Heterogeneous Network for Efficient Industrial IoT," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 5, pp. 831-844, May 2020, doi: 10.1109/JSAC.2020.2980907.

[24] K. Sood, K. K. Karmakar, S. Yu, V. Varadharajan, S. R. Pokhrel and Y. Xiang, "Alleviating Heterogeneity in SDN-IoT Networks to Maintain QoS and Enhance Security," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 5964-5975, July 2020, doi: 10.1109/JIOT.2019.2959025

[25] D. Wang, B. Bai, K. Lei, W. Zhao, Y. Yang and Z. Han, "Enhancing Information Security via Physical Layer Approaches in Heterogeneous IoT With Multiple Access Mobile Edge Computing in Smart City," in IEEE Access, vol. 7, pp. 54508-54521, 2019, doi: 10.1109/ACCESS.2019.2913438.

[26] J. Leu, C. Chen and K. Hsu, "Improving Heterogeneous SOA-Based IoT Message Stability by Shortest Processing Time Scheduling," in IEEE Transactions on Services Computing, vol. 7, no. 4, pp. 575-585, Oct.-Dec. 2014, doi: 10.1109/TSC.2013.30.

[27] V. Nguyen, J. A. Cabrera, G. T. Nguyen, D. You and F. H. P. Fitzek, "Versatile Network Codes: Energy Consumption in Heterogeneous IoT Devices," in IEEE Access, vol. 8, pp. 168219-168228, 2020, doi: 10.1109/ACCESS.2020.3023639.

[28] C. K. Wu et al., "An IoT Tree Health Indexing Method Using Heterogeneous Neural Network," in IEEE Access, vol. 7, pp. 66176-66184, 2019, doi: 10.1109/ACCESS.2019.2918060.

[29] H. Yang, A. Alphones, W. Zhong, C. Chen and X. Xie, "Learning-Based Energy-Efficient Resource Management by Heterogeneous RF/VLC for Ultra-Reliable Low-Latency Industrial IoT Networks," in IEEE Transactions on Industrial Informatics, vol. 16, no. 8, pp. 5565-5576, Aug. 2020, doi: 10.1109/TII.2019.2933867.

[30] J. Ni, X. Wang, M. Tang, W. Cao, P. Shi and S. X. Yang, "An Improved Real-Time Path Planning Method Based on Dragonfly Algorithm for Heterogeneous Multi-Robot System," in IEEE Access, vol. 8, pp. 140558-140568, 2020, doi: 10.1109/ACCESS.2020.3012886.

[31] R. Krishnamoorthy et al., "Systematic Approach for State-of-the-Art Architectures and System-on-Chip Selection for Heterogeneous IoT Applications," in IEEE Access, vol. 9, pp. 25594-25622, 2021, doi: 10.1109/ACCESS.2021.3055650.

[32] M. Noori, S. Rahimian and M. Ardakani, "Capacity Region of ALOHA Protocol for Heterogeneous IoT Networks," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8228-8236, Oct. 2019, doi: 10.1109/JIOT.2019.2920161.

[33] M. Asad, S. Qaisar and A. Basit, "Client-Centric Access Device Selection for Heterogeneous QoS Requirements in Beyond 5G IoT Networks," in IEEE Access, vol. 8, pp. 219820-219836, 2020, doi: 10.1109/ACCESS.2020.3042522.

[34] Bhavana, A., and AN Nandha Kumar. "An Analytical Modeling for Leveraging Scalable Communication in IoT for Inter-Domain Routing." In Proceedings of the Computational Methods in Systems and Software, pp. 1-11. Springer, Cham, 2018.

[35] Huston, Geoff. Commentary on inter-domain routing in the internet. RFC 3221, December, 2001.

[36] Gomez, Carles, M. Kovatsch, H. Tian, and Z. Cao. "Energy-Efficient features of internet of things protocols." draft-ietf-lwigenergy-efficient-06 (trabajo en curso) (2017).

Security and Threats of RFID and WSNs: Comparative Study

Ghada Hisham Alzeer[1], Ghada Sultam Aljumaie[2], Wajdi Alhakami[3]
Taif University, College of Computers and Information Technology, Taif, KSA

***Abstract*—The Internet of Things (IoT) has garnered significant attention from people with growing changes in human life over the last few years. IoT is a network of a group of smart devices that use sensors to collect information and conduct events in their environments. The information can then be shared on the Internet. IoT uses a range of technologies and finds various applications such as smart homes, environmental monitoring, and healthcare. In this paper, we conducted a comparative study to analyze the difference between two technologies—Wireless Sensor Networks (WSNs) and Radio Frequency Identification (RFID). It is pertinent to note that these technologies would not be effective without incorporating security aspects due to a potential number of threats and attacks on the network. This paper provides a comprehensive review of the recent approaches to securing RFID and WSNs. We have carefully chosen most of these studies to investigate only the recent technique from 2017 to 2020. The paper also highlights common attacks on RFID and WSNs and the secure authentication mechanisms on these technologies. It further provides a different way of detecting varying attacks in RFID and WSNs.**

Keywords—Security; IoT; WSN; RFID

I. Introduction

The Internet of Things (IoT) is a network of a group of smart devices that use sensors to collect information and conduct events in their environments. The information can then be shared on the Internet. IoT has witnessed rapid growth recently; Cisco reported a remarkable increase in the number of IoT devices to nearly 50 billion in 2020 [1]. IoT is used in several areas such as industrial automation (Industrial IoT), sensing applications in smart homes, traffic control, and other applications that deal less with sensors and more with data analysis. Industrial IoT and smart homes deal more with sensors and less with data analysis. The IoT that focuses more on data analysis is used in the transformation of business processes (BPs) such as banking, organizational operations, and healthcare optimization [2][3].

IoT uses a wide range of technologies such as Wireless Sensor Networks (WSNs), Radio Frequency Identification (RFID), and Near Field Communication (NFC), as shown in Fig. 1 [4].

Among these technologies, WSNs and RFID are mainly used and have become the two main pillars [4].

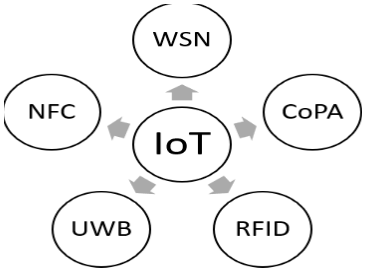


Fig. 1. IoT Technologies.

II. Radio Frequency Identification

RFID can be defined as the nonlinear network system that replaces barcodes and QR codes for a rapid response and relies on radio waves to capture and disseminate information [5]. It was first designed in 1948 and took many years to mature and become affordable and reliable for widespread use. Some considered RFID as the most widespread computing technology in history [6]. Today, it has become an important and integral part of current technologies such as computing and IoT [7], [8]. RFID is composed of four parts: tag, an antenna and transceiver tag processor, a database, and a backend. Tags are connected to items to store their information, the RFID reader reads the data coming from the tag and writes it to the transponder, and the backend database links that data with records. See Fig. 2 [6], [5], [9].

Active tags include a battery that allows automatic data transfer to the readers. On the other hand, passive tags are triggered by the electromagnetic waves of the reader.

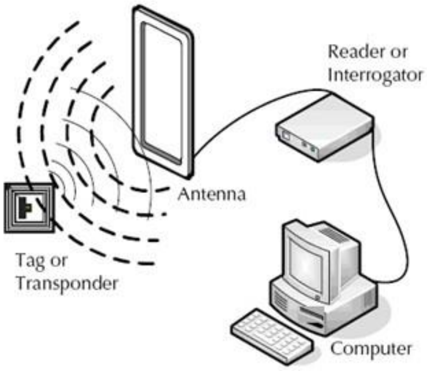


Fig. 2. Components of RFID.

These tags are more commonly used than their active counterparts on the account of their low cost and infinite life. Tags contain read-only memory (ROM), that stores data classified as security data, system ID, and OS instructions and volatile read/write or random access memory (RAM) that stores data during transmission and response [6], [5]. They are used in various applications such as transportation, logistics, manufacturing, healthcare/pharmaceutical industry, processing, and security [9], [7]. With the advent of IoT technology and the development of signal processing technology and distributed network technology for IoT nodes to acquire signals, a model has been established to acquire radio frequency signals within an IoT environment to add more features that are important in many fields [10].

III. Wireless Sensor Networks

WSNs have been becoming the area of interest for various researchers due to the rapid development of wireless technology and embedded electronics. WSN contains node sensors – small devices used to sense their current environment [11]. It is a distinct type of network containing small distributed devices called sensor nodes. They are considered low-power devices that communicate with each other without infrastructure and used for sensing and collecting data through wireless communication [12]. The basic components of sensor nodes include microcontrollers that perform data processing and control other components to perform their functions [13]. Transmitter and receiver use radio waves to send and receive data over wireless networks. Wireless sensors are powered by batteries or a power source. The choice of power source depends on the deployment environment and energy availability of the applications [14]. As provided in Fig. 3, EEPROM or Flash memory [15] are also the key components of sensor nodes.

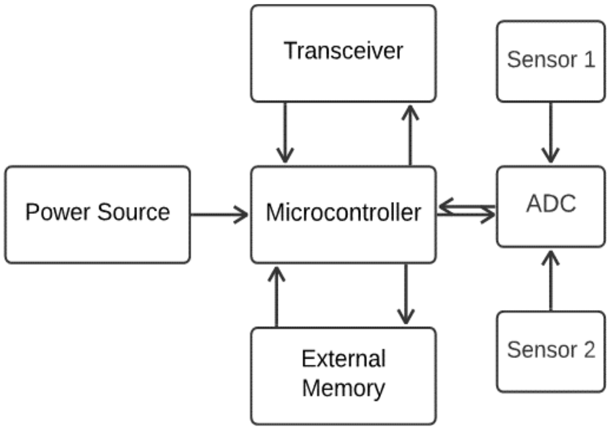


Fig. 3. Basic Components of WSN Node.

IoT model enables computers to access data about objects and the environment without human interaction [10]. Such model involves the integration of ‘physical things’ and IT infrastructure to transfer and collect data through a wireless network. It further allows to understand, interpret, communicate, and exchange data without any communication units and human participation [16], [10]. WSN plays an important role in IoT applications [17], as it provides IoT applications with high sensing and operational capabilities. WSNs are the eyes and ears of IoT; they convert physical phenomena into digital signals and transmit these signals for processing and analysis [18]. Today, there is a myriad of applications that depend on WSN and IoT technology, such as patient monitoring (measuring blood pressure, heart rate, and oxygen concentration) [19] and smart homes and buildings [17]. With the tremendous growth of IoT devices with high connectivity, there has been an increasing concern about their security and the data they store and transmit across various devices. Moreover, there has been an increase in the number of attacks on these devices. The current security challenges of IoT devices are generally due to their limited capacity, processing power, and battery life [20]. These limitations have made IoT devices a target for attackers such as hackers, hacktivists, and cybercriminals. Cybersecurity is therefore important to secure IoT and ensure protection from malicious activities such as data theft, modification, unauthorized access attempt, or network attack [20].

RFID and WSN technologies are widely used in many applications, such as in the scientific or medical fields and even in our home life, so achieving security in them is very important because they may deal with very sensitive data. Therefore, security became sour main motivation in this paper, we discussed the security requirements and how to achieve them, the common attacks based on current research also discussed protection and detection mechanisms suggested by other researchers. Our research paper is one of the few that discusses both RFID and WSN in terms of security requirements and common attacks.

This paper is divided into nine sections: Section IV introduces the required security applied in RFID and WSNs. The common threats and attacks on RFID and WSNs are presented in Section V. The following Sections VI and VII, respectively focus on the security of RFID followed by WSNs for achieving secure authentication, ensuring confidentiality, and detecting common attacks on both. In Section VIII we discussed the papers mentioned in our paper from various aspects. Lastly, we mentioned our future work on RFID and WSNs in Section IX.

IV. Security Requirements of Internet of Things

To secure IoT deployment, we classified IoT security into three categories as listed in Fig. 4.

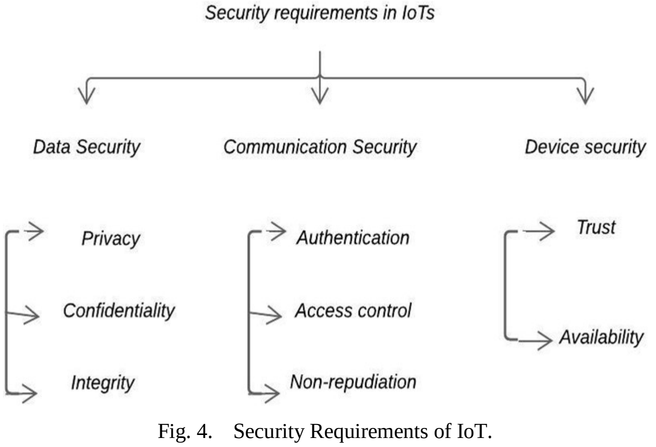


Fig. 4. Security Requirements of IoT.

A. Data Security (Privacy, Confidentiality, and Integrity)

Privacy includes the ability to hide personal information and control the use of that information [21]. There are several techniques to deal with data privacy, such as pseudo-random number generators, block and stream cipher, and anonymization [22]. Confidentiality, on the other hand, means that the communication between the sender and receiver must be protected from any malicious or unauthenticated users [23]. The integrity of data stored on remote servers must be protected on the IoT framework that preserves stored data, ensures its correctness, and provides protection from any loss or tampering. Many protocols were designed to achieve data integrity by using either encryption or anonymization techniques [24].

B. Communication Security (Authentication, Access Control, and Non-repudiation)

Authentication before communication is the key to the success of IoT and an important component of any security model [25]. The two parties must authenticate their communication [26]. It ensures the identification of these parties before making any contact [25]. Identity verification is carried out using several methods such as passwords, digital certificates, lightweight cryptography algorithms, or biometric identification [27]. IoT authentication is a complex process as it involves heterogeneous network authentication. Before joining a network, identification and authentication must be applied to all objects or sensors. It is imperative to note that IoT requires a unique code (UID) to identify each entity in the network [28]. Access control involves the authorization of users. A system administrator specifies access privileges for different users with which they can only access the relevant parts of system resources to protect their resources and information [29]. Access control algorithms can be divided into five types [22]:

1) Task-based Access Control (RBAC): RBAC manages all user-assigned access to roles and grants multiple user permissions to roles. For more efficiency, roles can be organized into a hierarchy, allowing some roles to inherit permissions from others. RBAC is generally used to simplify access control. It reduces complex protection management and endorses the analysis of user-assigned permissions [30].

2) Organization-based Access Control (OrBAC): An improved version of the RBAC model. However, it has a time limitation and supports the periodic activation of roles [31].

3) Capability-based Access Control (CapBAC): CapBAC gives each user a capability – a key that gives access rights. The admin then decides if the user can access the network by checking the validity of the key [32].

4) Attribute-based Access Control (ABAC): Depending on the characteristics of the requester and resource, users do not need to know the resources before they submit the request. ABAC has become significant recently, particularly in web service applications [33].

5) Trust-based Access Control (TBAC): It gives users a high level of trust to support dynamically changing permissions assigned to them [34]. Non-repudiation refers to a situation where data must be checked in a way that a sender has sent a message and it can be rejected or a receiver cannot refuse receipt of the message [35]. It can be achieved using Public Key Cryptography (PKC) and Digital signature [36].

C. Device Security (Trust and Availability)

Trust is critical to achieving security in an IoT system. Additionally, IoT devices must be trusted to prevent unwanted actions by malicious nodes [37]. The stages of trust-building start from the establishment stage to the operational and transmission stages of IoT. This trust is formed by two mechanisms – key generation and token. A key generated by the entitlement system is allocated to each new unit and introduced by a consumer device. Token, on the other hand, is generated by the owner or producer and coupled with an RFID indication of the device. [38]. In IoT, the availability of hardware and software remains essential. Hardware availability implies to the availability of devices for IoT applications at all times. Software availability is the ability to provide services at any place and time [39]. Moreover, in IoT devices, all data should be available to users whenever they need it. The devices and services must also be available and reachable whenever the users need them at the right time to achieve IoT expectations [38].

V. Attacks on Radio Frequency Identification and Wireless Sesnsor Network

In this section, we highlight some of the common attacks on RFID and WSNs.

A. Security Threats and Attacks on Radio Frequency Identification

The author in [40] summarized several threats directed towards RFIDs. A key reason behind most of these attacks is the security of the communication channel between the user and tags. A group of famous attacks on RFID is revealed below:

1) Action threat: In this type of threat, the tags possessed by an individual are monitored and predicted for his future intentions and actions.

2) Association threat: Electronic Product Code (EPC) tag is a unique number for each product. When a consumer purchases a product, a link between the consumer’s identity and the product is created.

3) Location threat: By tracking the tags associated with a user’s site, an attacker could obtain the exact location of the user.

4) Preference threat: It is possible to obtain consumer preferences illegally by tracking unique EPC tags for each product that identify company name and product type.

5) Constellation threat: It is one of the threats where the illegal parties track transactions between users.

6) Breadcrumb threat: Also known as electronic breadcrumbs, this threat occurs when a consumer buys a product that creates a link between his/her identity and EPC tag product number. Consequently, when the consumer gets rid of this product, the link is not broken and can be used.

Some common attacks on RFID systems mentioned in paper [41] are summarized in Table I.

TABLE I. SUMMARY OF SECURITY ATTACKS ON RFID TECHNOLOGY

Attacks	*Descriptions*
Temporariy disabling tags	The signs may be unintended – any event due to natural factors or interference of frequencies. They may also be intentional, such as Passive Interference and Active Jamming.
Removal or destruction of RFID readers	Because of its small pilgrimage, an RFID reader is vulnerable to attackers who use it to obtain data or modify it.
Relay attacks	Also known as MITMA; the intruder intercepts the radio signal between the sender and receiver and may modify it.
Attacks on the tags	Making a copy of the tag (Cloning) or impersonating the tag (Spoofing).
Reader attacks	Impersonating a legitimate reader (Impersonation) or recording the legitimate RFID tags (Eavesdropping).
Unauthorized tag reading	Since authentication protocol RFID tags are not supported, an attacker can read the contents of the RFID tags.
Tag modification	The data on RFID tags can be modified or deleted by the attacker.
Middleware attacks	The attacker uses RFID tags to either cause an attack (Buffer Overflows or end RFID middleware) or spread malicious code with an attack (Malicious Code Injection)
Covert channels	Using RFID tags, an attacker could create unauthorized channels for transmitting data.
DoS	The attacker blocks or disconnects RFID tags service from users.
Traffic analysis	Attacks by monitoring and analyzing traffic patterns
Crypto	Uses encryption methods to break encryption algorithms and access data
Side-channel	Leverages the physical application of encryption algorithms

B. *Security attacks and challenges of Wireless Sensor Networks.*

Fig. 5 [42] generally demonstrates the classification of the common security threats and attacks in WSNs.

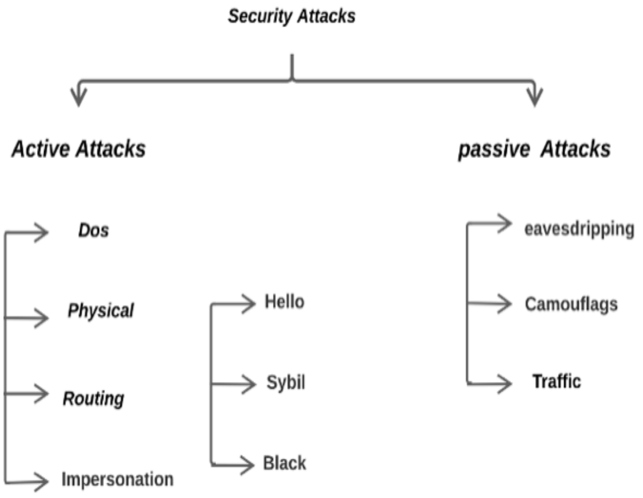


Fig. 5. Security Attacks in General.

The author in [43] notes Sybil attacks as the most common attacks observed in WSN, followed by wormhole and DoS attacks. DDoS attacks are relatively less on this type of network. The authors of paper [42] mentioned some common attacks on WSN systems, as shown in Table II.

TABLE II. SUMMARY OF SECURITY ATTACKS ON WSNS

Attacks	*Descriptions*
DoS	The attacker tries to sabotage the data and disable the system that reduces network efficiency
Sybil	In WSN networks, there are several sub-tasks such as duplicating information that you do not perform and assigning it to one node This node is attacked by Sybil Attacks, targeting the schemes of fault tolerance.
Blackhole	It is more severe than a Sally attack, as the attacker offers a shorter path to the nodes, acts as a black hole, and completely captures the data traffic. The attacker can also affect the data traffic.
HELLO Flood	This attack occurs in the network layer where the attacker fabricates hello, sends it to convince the sensor in WSN, and then changes the scenario
Wormhole	A common attack that occurs in two separate nodes carrying important parts of the message when a low-latency bandwidth is directed to them

VI. SECURITY IN RADIO FREQUENCY IDENTIFICATION TECHNOLOGY

This section includes an overview of previous works on RFID network security divided into several sections:

A. *Authentication Protocols for Radio Frequency Identification*

In [8], the authors introduced a new authentication protocol that offers an acceptable level of protection. It is also resistant to the risks reported in the article and evaluates the security of mutual authentication suggested by Wang and Ma. This review demonstrates the key security pitfalls of the protocol. Firstly, they presented two methods used by an opponent to make valid readers believe that they are dealing with a valid database. Next, they demonstrated how an adversary can turn an RFID reader into a legal database and introduced a new adversary model. Finally, they implemented an improved server method named ISMAP and demonstrated that this protocol provides sufficient protection against different types of attacks including the current adversary model discussed in the article. Additionally, the authors in [44] introduced a new lightweight RFID security authentication protocol (LRSAS). They analyzed the security properties of the protocol, containing data confidentiality and integrity (DCI), replay attack (RA), desynchronization attack (DA), impersonation attack (IA), tracking attack (TA), denial of service attack (DoS), and forward security (FS). Finally, they compared the LRSAS protocol with other protocols in terms of communications, computation, and storage. The authors also showed that the protocol is efficient in terms of security and cost requirements.

In [45], the authors presented two lightweight RFID protocols that provide security, identity authentication, and privacy and have multiple tag groups. They used a filtering process to decrease collision between tags, sleep activation mechanism, RFID system, and computing load. They also used a pseudorandom number generator (PRNG) and hash function to encrypt all sessions between the reader and tags. These protocols can resist eavesdropping, replay, and desynchronized attacks.

In [46], the authors introduced a group-based authentication protocol for the RFID system. It uses only mod operation and bitwise XOR. Additionally, two standard measures were used to measure the privacy of the system, resulting in anonymity when the opponent conducts numerous operations. Experimental results showed that their scheme maintains a high level of privacy when some tags are compromised. After the analysis, the authors proved that their protocol is safe and effective for a reduced RFID system.

B. *Security Communication in Radio Frequency Identification to Ensure Confidentiality*

In [47], the authors studied elliptical curve coding (ECC) protocol based on RFID security protocol, as it has several important features such as high strength ECC encryption that provides high security for communication and access to tag memory data. The new protocol relies on simple calculations such as XOR and bitwise AND which reduces complex calculations for low-cost tags. The authors analyzed their protocol for security and performance by using BAN logic. The analysis demonstrated that the protocol can provide mutual authentication of the tags and reader at the same time.

C. *Detection Mechanisms in Radio Frequency Identification*

In [48], the authors presented new effective research to preserve the privacy of cloning, as it is relevant and effective to preserve the privacy to explore cloning for all supplies that support RFID technology. They analyzed and evaluated the proposed mechanism through simulations which proved to be effective under various conditions. They then designed and implemented Multilateral Secure Computing (SMC) protocols to implement private-preserving for clone estimate that shows changes in efficiency regarding similar programs inside the existing SMC system. In [49], the authors discussed important problems associated with tag detection in RFID systems, including reader collision avoidance, optimal tag reporting, and optimal tag coverage problems. These issues occur due to the inability of collision intrusion detection and RFID readers that transmit packets created by other readers and poor access to resources in RFID tags on the account of severe limitations.

In [50], the authors presented an approach that implements MAC, routing, and application layer outlier detection processes in three different regions. Multiple invigilator regions executed internal or external detections after data collection. The proposed system has consequently been found to be efficient in terms of performance indicators. These indicators may be internal or external based on service quality. Various internal indicators used to measure the stability of structures are DI, RMSSDI, RSI, SI, CHI, and DBI. Additionally, various external indicators used to measure the stability of structures are FI, NMII, PI, and EI. Both internal and external indicators confirm the formation of structure and external detection processes. Furthermore, two indicators based on QoS (productivity and jitter) are used in this work.

The authors in [51] presented an efficient hash-based RFID authentication protocol that provides miss-tag detection. They presumed that for each user, an authentication system would validate large quantities with RFID tags inside its ranges. Their protocol can detect and reset lost tags if the missing tag can rejoin the system. After analyzing the protocol in terms of security, they proved that it can provide adequate security guarantees, resist various attacks, and offer better performance. Moreover, the protocol achieves both security and performance characteristics. See the summary of security in RFID technology in Table III.

TABLE III. DIFFERENT SECURITY TECHNIQUES ON RFID

Paper	Year	Techniques	Contribution
[8]	2020	GNY Logic and Scyther	The authors introduced a modern authentication protocol that offers an acceptable level of protection and is immune to security risks.
[44]	2020	Hash function, PRG, SKINNY encryption algorithm	The authors introduced a new lightweight RFID security authentication protocol (LRSAS). They analyzed security properties of the protocol that contain Data Confidentiality and Integrity (DCI), Replay Attack (RA), Desynchronization Attack (DA), Impersonation Attack (IA), Tracking Attack (TA), Denial of Service Attack (DoS), and Forward Security (FS).
[45]	2020	Hash function, PRG, activate-sleep mechanism,and filtering process	The authors presented two lightweight RFID protocols that provide security, identity authentication, and privacy.
[46]	2020	XOR operation	The authors introduced a group-based authentication protocol for the RFID system.
[47]	2016	XOR and bit wise AND	The authors studied elliptical curve coding (ECC) protocol based on RFID security protocol as it has several important features
[48]	2010	Algamal encryption system	The authors presented a novel efficient, private information mechanism to detect clones for RFID-enabled supply chain operations.
[49]	2009	Tree flow algorithm	The authors discussed many important problems associated with tag detection in RFID systems, such as reader collision avoidance, optimal tag reporting, and optimal tag coverage problems.
[50]	2019	DI, RMSSDI, RSI, SI, CHI, DBI, FI, NMII, PI, and EI	The authors presented an approach that implements MAC, routing, and application layer outlier detection processes in three different regions. The multiple invigilator region executes internal or external detections.
[51]	2018	Hash function	The authors presented an efficient hash-based RFID authentication protocol that provides miss-tag detection.

VII. Security in Wireless Sensor Networks

In this section, many papers written on the security of WSNs have been compiled and divided into several sections as shown in the following:

A. *Authentication Protocols for Wireless Sensor Networks*

In [52], the authors mentioned weaknesses in traditional authentication methods found in IoT and suggested the use of a system based on WSN identity authentication and blockchain technology. Blockchain is a book of accounts that cannot be modified or tampered with and where transactions or data are generally recorded. They integrated blockchain decentralization with the nodes that formed the IoT structure. In a public blockchain, several private blockchains are connected and each private blockchain is connected between the cluster heads of a WSN. In the end, we have a hybrid blockchain for the whole network. The authors also created a model where the identification data was recorded between cluster head nodes and ordinary nodes. Finally, a connection authentication is done between these nodes. After analyzing the model, it became clear that the system has a greater and more efficient level of safety.

The researchers in [53] submitted a proposal to make the use of authentication protocols in WSN more secure and focused on reducing the cost as compared to other conventional protocols. They used the Altera DE2 demo board and implemented several corresponding device structures such as the Altera Cyclone II field-programmable gate array. Finally, they showed the waves produced from this process – 16702A – a logic analysis device. Additionally, the process XOR was used for encoding the key. The results showed the effectiveness of the experiment.

Paper [54] also mentioned many concerns about the difficulty of preventing smartcard stolen and off-line guessing attacks. To prevent these attacks, the paper suggested using a protocol that uses honey-list technology and relies on three-factor authentication. As the sensor performance is limited, the protocol also encodes the elliptic curve that relies on the public key and uses only hash functions. The authors performed a formal security analysis using the real-or-random (ROR) and Burrows Abadi Needham (BAN) models. For verification, they used simulation software called Automated Validation of Internet Security Protocols and Applications (AVISPA) that resulted as a safe protocol.

The author in [55] focused on lightweight and cost as the two main features; the authors saw that WSN devices need strong and light authentication protocols that can withstand any difficult environment. They proposed a model that uses XOR and hash functions. This model was effective in terms of reducing the use of resources and speed while maintaining data security.

B. *Secure Communication in Wireless Sensor Networks to Ensure Confidentiality*

There has been a growing need to guarantee the high security of WSNs used in various applications such as home, industrial, and healthcare. Therefore, paper [56] proposed a protocol that improves the security of WSN by distributing the main keys, identifying the node, and verifying the identity of messages in WSN. The password is updated and changed for the message verifier and connected to the dynamic node on the network. The authors concluded that this method outperformed previous methods. Subsequently, paper [57] used a scheme based on additive homomorphic encryption algorithm in WSN, whereby a symmetric-key homomorphic is used to provide more protection for the confidentiality of data. This key also combines the data with a homomorphic signature to achieve integrity. After decoding, the data is classified according to various symmetric-key homomorphic. Furthermore, after analyzing the results, it became clear that using this method is effective in reducing cost and increasing effectiveness in terms of protecting the data from any tampering during its transmission and ensuring the accuracy of its collection.

The author in [58] suggested the use of hybrid technology from Diffie-Hellman key exchange and Elliptic Curve cryptography. The combination of these two technologies allowed for increased security of data traffic, confidentiality, authentication, and time savings. These techniques are simulated, applied to a Java platform, and implemented in a WSN environment. The authors of [59] directed their efforts towards solving the security problems of sensitive data, as it traveled through WSN for various applications, by applying new technologies. They integrated discrete chaotic map and genetic cryptography as 2DES and 3DES for WSN, which increased the security regardless of limited resources. For a text and visual data, Henon map encryption was used due to its strong encryption. They encoded these processes under the Arduino microcontroller and determined that the attacker might need time depending on the speed of his device. They concluded that using random numbers increases the robustness of the system and prevents attacks. They preserve the confidentiality of data from unauthorized disclosure and collect it with high accuracy, as shown in Fig. 6.

To increase security and make IoT devices more independent, the authors in the paper [60] suggested the use of blockchain security features such as availability to users, data integrity, and various cryptographic tools. The model was applied to the WSNs that were used to measure moisture and temperature. It was found that the transmission of information between the nodes became more secure, independent, and less vulnerable to various types of attacks such a DoS and MITM attacks.

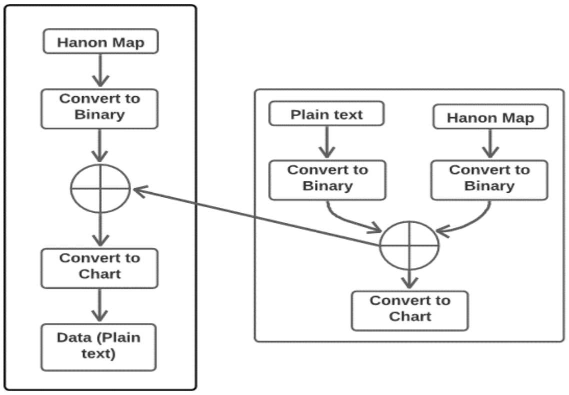


Fig. 6. General Block Diagram of Henon Map.

C. Detection Mechanisms in Wireless Sensor Networks

DoS jamming attack is one of the common attacks on WSN, as discussed in [61]. It aimed at sending many signals to jam the main signal. A denial of service occurred consequently and caused disruption of functions in the WSN nodes.

The authors in [61] proposed an exponentially weighted moving average (EWMA). They deployed an exponential moving variable that detects any change occurring in the traffic. The authors concluded that this model can accurately detect different jamming attacks and be used in situations where sensitive instantaneous information is transmitted.

Due to the sensitivity of information transmitted through WSN, a solution has been proposed in [62] to discover the unauthorized and intentional sequences of WSN. The sequence detection methodology in this paper relied on the use of MATLAB Simulink that uses an artificial neural network. In the first session, a large discrepancy in node values makes them a harmful contract for WSN. For the second session, the results of the regression of the artificial neural network for both packet delivery ratio (PDR) and energy consumption variables were analyzed. It was observed that ANN-based PDR is stronger and quicker than ANN-based energy usage. However, the results for both were good.

In the survey paper [63], a part of its objectives was to provide a comparison of different intrusion detection protocols of each WSN and IoT. It mentions the uses and efficiency of each type.

The authors in [65] aimed at using a new system that detects the sequence and has a longer residence time by adding a low-power resistance and survival continuity to IDS. The paper showed that nodes continue to work efficiently on algorithm strength, mobile nodes, and attack strength. See the summary of security in WSNs in Table IV.

TABLE IV. DIFFERENT SECURITY TECHNIQUES ON WSNs.

Paper	Year	Techniques used	Contribution
[52]	2020	Hybrid blockchain	The authors mentioned weaknesses in the traditional authentication methods of IoT and suggested the use of a system based on many WSN identity authentication with blockchain.
[53]	2015	XOR arithmetic	The researchers submitted a proposal to make the use of authentication protocols in WSN more secure and focused on reducing the cost as compared to other conventional protocols.
[54]	2020	Honey-list, three-factor authentication	The paper suggested using a protocol that uses honey-list technology and relies on three-factor authentication for preventing smartcard stolen and off-line guessing attacks.
[55]	2020	XOR and hash functions	The authors saw that WSN devices need strong and light authentication protocols and that can withstand any difficult environment.
[56]	2017	The protocol distributing the main keys, identifying the node and verifying the identity	The paper proposed a protocol that increases the security of the WSN by distributing the main keys, identifying the node also verifying the identity of the messages in the WSN.
[5]	2015	Symmetric-key homomorphic, homomorphic signature	The authors proposed a scheme based on an additive homomorphic encryption algorithm in WSN for the confidentiality of data.
[58]	2018	Diffie-Hellman, Elliptic Curve	Suggested the use of hybrid technology from Diffie-Hellman key exchange and Elliptic Curve cryptography. The combination of these two technologies allowed for increased security of data traffic, confidentiality, authentication, and time savings.
[59]	2020	Discrete chaotic map, genetic cryptography, Henon map	Solving the security problems of sensitive data, as it traveled through WSN for various applications, by applying new technologies. The authors integrated discrete chaotic map and genetic cryptography as 2DES and 3DES for WSN which increased security regardless of limited resources.
[60]	2020	Blockchain	To improve security and make IoT devices more independent, the authors suggested the use of blockchain security features such as availability to users, data integrity, and various cryptographic tools.
[61]	2018	Exponentially weighted moving average (EWMA)	The authors proposed an exponentially weighted moving average (EWMA). They deployed an exponential moving variable that detects any change occurring in the traffic. It can accurately detect different jamming attacks.
[62]	2020	Artificial neural network, MATLAB Simulink	Due to the sensitivity of information transmitted through WSN, a solution has been proposed by researchers to discover the unauthorized and intentional sequences of WSN. They relied on the use of MATLAB Simulink which uses an artificial neural network.
[63]	2019	Survey (e.g. Intrusion detection by cluster head, Hybrid anomaly detection..)	The paper provides a survey comparison of various intrusion detection protocols in both WSN and IoT. It mentions the uses and efficiency of each type.
[64]	2015	Low-power resistance and survival continuity to IDS	The paper aimed at using a new system that detects the sequence and has a longer residence time

VIII. DISCUSSION

This part focuses on discussing the above-mentioned recent techniques for protection and detection mechanisms. The discussion section has been divided into two parts:

A. Critical Review of Radio Frequency Identification Security

This section discusses the most important recent approaches which aim to secure an RFID environment. These approaches are selected as the most relevant and have novelty. For example, many research papers have discussed authentication and their protocols are distinguished using lightweight encryption algorithms. They consume fewer resources during calculation and are more efficient compared to traditional encryption algorithms. They are also suitable for devices with limited computing power such as RFID. The authors in papers [8] took a few measures to overcome some of the flaws and introduced an improved protocol using Scyther and GNY Logic. These are two excellent ways to assess security for the protocol of cryptography. However, the protocol has a drawback – it does not take into account multi-server or multi-reader environments. In two other researches [44] [45], the authors used hashing function to encrypt all session between tags and reader that ensures data integrity. They also used a pseudo-random number process to strengthen the encryption, making it difficult for the attacker to guess the key used. Both protocols proved effective in protecting against many types of attacks such as restart attack (RA), trace attack (TA), denial of service attack (DoS), and security forwarding (FS). The difference between the two papers is that the authors in [45] used the SKINNY encryption algorithm for the data used by the tag and reader to achieve mutual authentication. In our opinion, their protocol was good because it balanced security requirements and costs. Additionally, the use of SKINNY was well-suited for a scenario where the server is connected to numerous lightweight devices. In [44], we liked that they used the activate-sleep mechanism efficiently and filtering process which reduced collision on the tags. In the paper [46], the authors not only used a pseudo system that provides a feature on the side of the tag but also used that feature in the reader to generate the nonce. Their protocol only uses bit-wise XOR operation in the authentication stage along with symmetric encryption and decryption. It was an excellent protocol, as it uses fewer resources in the tags to achieve arithmetic work and store data. Moreover, it maintains a high level of privacy when attacking some tags.

In the next section, we discussed RFID security communication to ensure confidentiality. The authors in [47] presented a protocol based on an elliptical curve for coding. From our point of view, their protocol has several advantages. It provides mutual authentication for the tag and reader and is good at resisting some of the common attacks related to RFID technology. Additionally, their protocol only relies on a few simple operations such as XOR and bitwise AND which reduces the complexity of computation in low-cost tags. In the following section, we discussed several research papers regarding detection mechanisms in RFID. The researchers in the [48] presented a novel, reliable, privacy-preserving mechanism for detecting clones for RFID-enabled supply chain operations. They used the Algamal encryption system, which is an asymmetric encryption system, in their protocol that achieved both authentication and confidentiality. Their protocol has been effective at detecting RFID supply chain clones. However, from our point of view, their protocol has many weaknesses such as the need for more robust hardware. They also need to reduce the security level to n/2 to improve the performance of their protocols. The authors in [49] provide a distributed and localized algorithm. They used a tree flow algorithm centered on the recursive direction of the binary tree for tag identifiers and the problem of tag collision where the reader initially sends a broadcast including the string of "0". The ID of all these tags in the interrogation space starts with a "0" bit. When an answer is received or a collision of the tag is observed, the reader will iterate on both sub-trees "0" rooted at "00" and "01" However, if there is no answer, the reader assumes that there is no "0"-tags preceded in their interrogation region and sends a question "1" afterward. For the reader, the difficulty of TWA is proportional to the number of tags in TWA. The researchers introduced in [51] the protocol for dealing with lost tags. Their protocol depends on lightweight cryptographic techniques and the key size is taken into account. In their protocol, RFID tags and key size are the two main factors that affect the entire group authentication process. From our perspective, their protocol is unique because they considered the effect of key size on authentication efficiency, assuming the presence of a large number of RFID tags. They also proved their protocol efficient, as it requires less time to authenticate the tag, provides resistance to a replay attack, and all the tags are independently verified. In another in research [50], the authors suggested a scheme. It was found to be effective in terms of performance indicators. These indicators can be internal, external, or QoS-based. Internal indicators that have been used to measure structural stability are DI, RMSSDI, RSI, SI, CHI, and DBI. External indicators that were used to measure the stability of structures are FI, NMII, PI, and EI. Two additional indicators were also used based on service quality (productivity and jitter). One of the advantages of their model is that it observed an improvement of 0.15% in minimum and 14.9% in maximum in the case of network instability without outliers compared to that of the network with outliers. It has further proven high efficiency.

B. *A Critical Review of Wireless Sensor Networks Security*

After addressing the security requirements of WSN, we are reviewing various scientific papers focusing on their security. It was noted that in terms of authentication, blockchain technology is considered one of the leading modern techniques. Some researchers [52] used the technology in several ways; some used it as a blockchain structure linked to the head node. It is followed by the blockchain linked to sub-nodes that formed a structure distinguished by its effectiveness in the authentication. However, it can take more time in the case of a large number of nodes and become immune to the attacks of concurrent guessing against IDs and passwords. Some [60] have applied this technique to other ways on the IoT, but its effectiveness cannot be confirmed when applied to WSN, except through experience. It was also noticeable that the AVISPA tool was used [54], which aims to analyze the Internet safety protocols on a large scale and increases the strength of the experiment results. The authors were also focused on making security protocols that are lightweight, affordable, and offer high security in return. However, lightweight protocols may not be able to detect harmful nodes in WSN. Concerning secure communication in WSN, one of the research papers [57] suggested the use of homomorphic encryption. However, this type of encryption is known to be vulnerable to compromise attacks. To solve this dilemma, you can attempt to split the data into pieces and send them to different aggregators. Another type of encryption was also mentioned in one of the papers [58] that merges Diffie-Hellman key exchange and Elliptic Curve cryptography. It was an effective method in terms of time-saving, data security, and authentication. Another paper [59] proposed encryption based on data clutter that uses Hénon map to generate random numbers. However, it can break after several attempts, depending on the ability of the attacker. One work [51] mentioned updating the key periodically without the need for any sync, but the work did not mention time and efficiency factors. In the algorithms for the detection of attacks, Exponentially Weighted Moving Average (EWMA) [61] was used and the results were accurate, indicating the superiority of this method. Sequence attacks were detected in one work using algorithms [62] and in another work [64] using resistances placed on nodes. Both the studies gave positive results. We noticed a difference between RFID and WSN; in terms of security techniques, most of the references we discussed on RFID used a simple approach to achieve security, as the RFID tag has a short reading range from 5 meters (ideal conditions) to less than 1 meter (not ideal conditions). In contrast, a more sophisticated approach was used in most of the literature we discussed on WSNs. It is observed that authentication is different in these two technologies the reason behind that the different capabilities of them e.g., we can apply only lightweight approaches on RFID while we can apply the complex algorithms on WSN. This paper provides a comprehensive review of the recent approaches for securing RFID and WSNs.

IX. Conclusion

IoT technology has become an essential part of our era. It is defined as the set of devices connected for collecting and analyzing data from their environments. The types of technologies that use IoT are bifurcated. In this paper, we have highlighted the security and attacks of both WSNs and RFID since they are parts of the IoTs environments. The goal of providing a comprehensive study and investigate the recent research related to the security of WSN and RFID technologies in terms of security requirements, detection techniques, and prevention of attacks against them are accomplished. Thus the comprehensive discussion of these technologies of research observed in terms of efficiency, comparison of protocol security, cost, and weight is included.

In the future, we will keep up with the new approaches, further investigate, and compare the performance and security mechanisms of RFID and WSN.

References

[1] Bhabendu. Kumar. Mohanta,. Debasish. Jena,. Utkalika. Satapathy,. Srikanta. Patnaik,. Survey on iot security: Challenges and solution using machine learning, artificial intelligence and blockchain technology. Internet of Things, 11:100227, 2020.

[2] Daniel Minoli and Benedict Occhiogrosso. Blockchain mechanisms for iot security. Internet of Things, pages 1 – 13, 2018.

[3] Ahmet Arıs, Sema F Oktug, and Thiemo Voigt. Security of internet of things for a reliable internet of services. 2018.

[4] Yasmine Harbi, Zibouda Aliouat, Saad Harous, Abdelhak Bentaleb, and Allaoua Refoufi. A review of security in internet of things. Wireless Personal Communications, 108(1):325–344, 2019.

[5] You-Chiun Wang and Shu-Ju Liu.Minimum-cost deployment of adjustable readers to provide complete coverage of tags in rfid systems. Journal of Systems and Software, 134:228–241, 2017.

[6] Anas Mouattah and Khalid Hachemi. The feasibility of motion sensor-based smart rfid system in improving the power saving. 2020.

[7] Chris M Roberts. Radio frequency identification (rfid).Computers & security, 25(1):18–26, 2006.

[8] Mehdi Hosseinzadeh, Jan Lansky, Amir Masoud Rahmani, Cuong Trinh, Masoumeh Safkhani, Nasour Bagheri, and Bao Huynh. A new strong adversary model for rfid authentication protocols. IEEE Access, 8:125029–125045, 2020.

[9] Andreas Koschan, Suhong Li, John K Visich, Basheer M Khumawala, and Chen Zhang. Radio frequency identification technology: applications, technical challenges and strategies. Sensor Review, 2006.

[10] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. Future generation computer systems, 29(7):1645–1660, 2013.

[11] Mustafa Kocakulak and Ismail Butun. An overview of wireless sensor networks towards internet of things. In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pages 1–6. IEEE, 2017.

[12] Surbhi Gupta. Prosperity, vulnerabilities and security threats in wsn. International Journal of Advanced Research in Computer Science, 3(5), 2012.

[13] Marcos Augusto M Vieira, Claudionor N Coelho, DC jr da Silva, and Jose´ Monteiro da Mata. Survey on wireless sensor network devices. In EFTA 2003. 2003 IEEE Conference on Emerging Technologies and Factory Automation. Proceedings (Cat. No. 03TH8696), volume 1, pages 537–544. IEEE, 2003.

[14] Michal Prauzek, Jaromir Konecny, Monika Borova, Karolina Janosova, Jakub Hlavica, and Petr Musilek. Energy harvesting sources, storage devices and system topologies for environmental wireless sensor networks: A review. Sensors, 18(8):2446, 2018.

[15] Niels Reijers and Koen Langendoen. Efficient code distribution in wireless sensor networks. In Proceedings of the 2nd ACM international conference on Wireless sensor networks and applications, pages 60–67, 2003.

[16] Ameer Ahmed Abbasi and Mohamed Younis. A survey on clustering algorithms for wireless sensor networks. Computer communications, 30(14-15):2826–2841, 2007.

[17] Nayef Abdulwahab Mohammed Alduais, Jiwa Abdullah, and Ansar Jamil. Rdcm: An efficient real-time data collection model for iot/wsn edge with multivariate sensors. IEEE Access, 7:89063–89082, 2019.

[18] Xiaochen Lai, Quanli Liu, Xin Wei, Wei Wang, Guoqiao Zhou, and Guangyi Han. A survey of body sensor networks. Sensors, 13(5):5406–5447, 2013.

[19] Afsaneh Minaie, Ali Sanati-Mehrizy, Paymon Sanati-Mehrizy, and Reza Sanati-Mehrizy. Application of wireless sensor networks in health care system. age, 23(1), 2013.

[20] Gowthamaraj Rajendran, RS Ragul Nivash, Purushotham Parthiban Parthy, and S Balamurugan. Modern security threats in the internet of things (iot): Attacks and countermeasures. In 2019 International Carnahan Conference on Security Technology (ICCST), pages 1–6. IEEE, 2019.

[21] Sridipta Misra, Muthucumaru Maheswaran, and Salman Hashmi.Security challenges and approaches in internet of things. Springer, 2017.

[22] Arbia Riahi Sfar, Enrico Natalizio, Yacine Challal, and Zied Chtourou.

[23] A roadmap for security challenges in the internet of things. Digital Communications and Networks, pages 118–137, 2018.

[24] Carsten Maple. Security and privacy in the internet of things. Journal of Cyber Policy, 2(2):155–184, 2017.

[25] Israa Alqassem and Davor Svetinovic. A taxonomy of security and privacy requirements for the internet of things (iot). In 2014 IEEE International Conference on Industrial Engineering and Engineering Management, pages 1244–1248. IEEE, 2014.

[26] Isha Bhardwaj, Ajay Kumar, and Manu Bansal. A review on lightweight cryptography algorithms for data security and authentication in iots. In 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), pages 504–509. IEEE, 2017.

[27] Minhaj Ahmad Khan and Khaled Salah.Iot security: Review, blockchain solutions, and open challenges. Future Generation Com- puter Systems, 82:395–411, 2018.

[28] Mohammad Reza Sohizadeh Abyaneh. Security analysis of lightweight schemes for rfid systems. 2012.

[29] Shruti Jaiswal and Daya Gupta. Security requirements for internet of things (iot). In Proceedings of International Conference on Communication and Networks, pages 419–427. Springer, 2017.

[30] Mirza Abdur Razzaq, Sajid Habib Gill, Muhammad Ali Qureshi, and Saleem Ullah. Security issues in the internet of things (iot): a comprehensive study. International Journal of Advanced Computer Science and Applications, 8(6):383, 2017.

[31] D Richard Kuhn, Edward J Coyne, and Timothy R Weil. Adding attributes to role-based access control. Computer, 43(6):79–81, 2010.

[32] Anas Abou El Kalam, R El Baida, Philippe Balbiani, Salem Benferhat, Fre´de´ric Cuppens, Yves Deswarte, Alexandre Miege, Claire Saurel, and Gilles Trouessin. Organization based access control. In Proceedings POLICY 2003. IEEE 4th International Workshop on Policies for Distributed Systems and Networks, pages 120–131. IEEE, 2003.

[33] Yuta Nakamura, Yuanyu Zhang, Masahiro Sasabe, and Shoji Kasahara. Exploiting smart contracts for capability-based access control in the internet of things. Sensors, 20(6):1793, 2020.

[34] Bo Lang, Ian Foster, Frank Siebenlist, Rachana Ananthakrishnan, and Tim Freeman. A flexible attribute based access control method for grid computing. Journal of Grid Computing, 7(2):169, 2009.

[35] Rajiv Mishra and Rajesh Yadav. Access control in iot networks: Analysis and open challenges. Available at SSRN 3563077, 2020.

[36] Pol Van Aubel, Erik Poll, and Joost Rijneveld. Non-repudiation and end-to-end security for electric-vehicle charging. In 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), pages 1–5. IEEE, 2019.

[37] Edewede Oriwoh, Haider al Khateeb, and Marc Conrad. Responsibility and non-repudiation in resource-constrained internet of things scenarios. International Conference on Computing and Technology Innovation (CTI 2015), 2016.

[38] Sandro Etalle, Jeremy den Hartog, and S. Marsh. Trust and punishment. In Proceedings of the 1st International Conference on Autonomic Computing and Communication Systems, Autonomics, number 302 in ACM International Conference Proceeding Series, pages 5:1–5:6, Belgium, December 2007. ICST. http://eprints.ewi.utwente.nl/12914.

[39] Rwan Mahmoud, Tasneem Yousuf, Fadi Aloul, and Imran Zualkernan. Internet of things (iot) security: Current status, challenges and prospective measures. In 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), pages 336–341. IEEE, 2015.

[40] Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash. Internet of things: A survey on enabling technologies, protocols, and applications. IEEE communications surveys & tutorials, 17(4):2347–2376, 2015.

[41] Shantanu Rao, Nagaraja Thanthry, and Ravi Pendse. Rfid security threats to consumers: Hype vs. reality. In 2007 41st Annual IEEE International Carnahan Conference on Security Technology, pages 59–63. IEEE, 2007.

[42] Aikaterini Mitrokotsa, Melanie R Rieback, and Andrew S Tanenbaum. Classifying rfid attacks and defenses. Information Systems Frontiers, 12(5):491–505, 2010.

[43] Sayamuddin Ahmed Jilani, Chandan Koner, and Shovon Nandi. Security in wireless sensor networks: Attacks and evasion. In 2020 National Conference on Emerging Trends on Sustainable Technology and Engineering Applications (NCETSTEA), pages 1–5. IEEE, 2020.

[44] Adnan Ahmed, Kamalrulnizam Abu Bakar, Muhammad Ibrahim Channa, and Abdul Waheed Khan. A secure routing protocol with trust and energy awareness for wireless sensor network. Mobile Networks and Applications, 21(2):272–285, 2016.

[45] Liang Xiao, He Xu, Feng Zhu, Ruchuan Wang, and Peng Li. Skinny-based rfid lightweight authentication protocol. Sensors, 20(5):1366, 2020.

[46] Zhicai Shi, Xiaomei Zhang, and Jin Liu. The lightweight rfid grouping-proof protocols with identity authentication and forward security. Wireless Communications and Mobile Computing, 2020, 2020.

[47] Pramod Kumar Maurya and Satya Bagchi. Cyclic group based mutual authentication protocol for rfid system. Wireless Networks, 26(2):1005–1015, 2020.

[48] Quan Qian, Yan-Long Jia, and Rui Zhang. A lightweight rfid security protocol based on elliptic curve crytography. IJ Network Security, 18(2):354–361, 2016.

[49] Davide Zanetti, Leo Fellmann, Srdjan Capkun, et al. Privacy-preserving clone detection for rfid-enabled supply chains. In 2010 IEEE International Conference on RFID (IEEE RFID 2010), pages 37–44. IEEE, 2010.

[50] Bogdan Carbunar, Murali Krishna Ramanathan, Mehmet Koyut u¨rk, Suresh Jagannathan, and Ananth Grama. Efficient tag detection in rfid systems. Journal of Parallel and Distributed Computing, 69(2):180– 196, 2009.

[51] Adarsh Kumar and Alok Aggarwal. An efficient simulated annealing based constrained optimization approach for outlier detection mechanism in rfid-sensor integrated manet. Int. J. Comput. Inf. Syst. Ind. Manage. Appl., 11:55–64, 2019.

[52] Haowen Tan, Dongmin Choi, Pankoo Kim, Sungbum Pan, and Ilyong Chung. An efficient hash-based rfid grouping authentication protocol providing missing tags detection. Journal of Internet Technology, 19(2):481–488, 2018.

[53] Zhihua Cui, XUE Fei, Shiqiang Zhang, Xingjuan Cai, Yang Cao, Wensheng Zhang, and Jinjun Chen. A hybrid blockchain-based identity authentication scheme for multi-wsn. IEEE Transactions on Services Computing, 13(2):241–251, 2020.

[54] Shao-I Chu, Yu-Jung Huang, and Wei-Cheng Lin. Authentication protocol design and low-cost key encryption function implementation for wireless sensor networks. IEEE Systems Journal, 11(4):2718–2725, 2015.

[55] Joonyoung Lee, Sungjin Yu, Myeonghyun Kim, Youngho Park, and Ashok Kumar Das. On the design of secure and efficient three-factor authentication protocol using honey list for wireless sensor networks. IEEE Access, 8:107046–107062, 2020.

[56] Himani Sikarwar and Debasis Das. A lightweight and secure authentication protocol for wsn. In 2020 International Wireless Communications and Mobile Computing (IWCMC), pages 475–480. IEEE, 2020.

[57] Oguz Ata, Hasan H Balik, and Erdem Ucar. Protocol design for secure communication in wsn. TEM Journal, 6(2):192, 2017.

[58] Soufiene Ben Othman, Abdullah Ali Bahattab, Abdelbasset Trad, and Habib Youssef. Confidentiality and integrity for data aggregation in wsn using homomorphic encryption. Wireless Personal Communications, 80(2):867–889, 2015.

[59] Sunil Kumar, C Rama Krishna, and AK Solanki. A technique to resolve data integrity and confidentiality issues in a wireless sensor network. In 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pages 183–188. IEEE, 2018.

[60] Alya'a Abdulrazzak Msekh and Jamal Mohamed Kadhim. Security of wireless sensor nodes. Iraqi Journal of Science, 61(7):1773 – 1780, 2020.

[61] Alma E Guerrero-Sanchez, Edgar A Rivas-Araiza, Jose Luis Gonzalez-Cordoba, Manuel Toledano-Ayala, and Andras Takacs. Blockchain mechanism and symmetric encryption in a wireless sensor network. Sensors (Basel, Switzerland), 20(10), 2020.

[62] Opeyemi Osanaiye, Attahiru S Alfa, and Gerhard P Hancke. A statistical approach to detect jamming attacks in wireless sensor networks. Sensors, 18(6):1691, 2018.

[63] Bassam Hasan, Sameer Alani, and Mohammed Ayad Saad. Secured node detection technique based on artificial neural network for wireless sensor network. Int J Elec & Comp Eng ISSN, 2088(8708):8708.

[64] Sumit Pundir, Mohammad Wazid, Devesh Pratap Singh, Ashok Kumar Das, Joel JPC Rodrigues, and Youngho Park. Intrusion detection protocols in wireless sensor networks integrated to internet of things deployment: Survey and future challenges. IEEE Access, 8:3343–3363, 2019.

[65] Zixin Zhou, Lei Liu, and Guijie Han. Survival continuity on intrusion detection system of wireless sensor networks. In 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pages 775–779. IEEE, 2015.

Proposal of a Method to Measure Test Suite Quality Attributes for White-Box Testing

Mochamad Chandra Saputra[1], Tetsuro Katayama[2]
Dept. Materials and Informatics, Interdisciplinary Graduate School of Agriculture and Engineering,
University of Miyazaki, Miyazaki, Japan

***Abstract*—As an important asset in software testing, measuring quality attributes of the test suite is important to describe the quality of software. This research proposes a method to measure the test suite quality attributes for white-box testing. The attributes are usability, efficiency, reliability, functionality, portability, and maintainability that are selected from 28 attributes in software quality. By using the proposed method, the test suite quality attributes are calculated with various results of level of quality. The result of test suite quality attribute measurement then proves the validity of its result by the reliability analysis. It is used Cohen's kappa coefficient to validating the result of test suite quality attributes measurement based on the level of agreement between the result of measurement and expert assessment. Reliability analysis on test suite quality attribute finds the attribute that strongly related based on the minimum percentage of level of agreement value are usability, reliability and functionality. Hence, our proposed method is useful to measure test suite quality attributes.**

***Keywords*—Test case; test suite quality attributes; white-box testing; reliability analysis; software quality**

I. Introduction

The quality of software is confirmed by systematically exercising the software in carefully controlled circumstances especially in testing phase[1]. During the testing, test suite which contains several test cases play a very important role to check various aspects of the software such as actual program structure and the software functions as per the specification[2]. The test cases are usually developed by a set of inputs, execution preconditions, and expected outcomes for a specific objective. Testing in software development is one of the ways to ensure quality of the software.

The main activity of software testing is verification and validation[3]. In software development life cycle, verification and validation aim is to help the software development build software with good quality. Verification ensures the specific function of the software is correctly implemented. Validation ensures the software are suitable to customer requirement. One of the software testing approaches is white-box. The white-box testing approach aims to ensure that the program is successfully tested based on the internal structures of the software[4].

Software quality is defined as the whole of features and characteristics of a product or service that able to satisfy stated or implied user needs[1]. As an important asset in software testing, measuring the quality of a test suite is important to describe quality of the software. Software testing is one of the quality approaches to control the program before its delivery or installation at the user with an acceptable level of quality. Various software quality attributes have been used on software quality models to define the degree of quality. Software quality attributes are multipurpose attributes that mean any area of software development process can use the attributes. Examining code programs by using the test suite is one of the methods to assure their quality. Test suite quality measurement is necessary to gain information on the test suite performance.

The big problem with quality attributes is uncertainty attributes and their measurement for informing the degree of test suite quality. Currently, measuring the attributes of test suite quality is one of the interesting problems in software testing. The aim of test suite quality attributes measurement to provide useful information about the degree of test suite quality.

The objective of this research is to find and propose the test suite quality attributes measurement and then validate its measurement by the reliability analysis. The research concern with quality attributes for test suite in white-box testing. The research provides a questionnaire for the expert to assess the test suite quality attribute based on their experience. The reliability analysis uses Cohen's kappa coefficient approach. Cohen's kappa coefficient is used to analyze the reliability of test suite quality from test suite quality attributes measurement result and expert assessment. The level of agreement is presenting the reliability of the test suite quality attributes measurement with the expert assessment. This research uses only the test suite for white-box testing. The results of the reliability analysis are the test suite quality attributes that have strongly agreed to the quality of the test suite.

The rest of the paper is organized as follows. Section 2 describe the highlight work done by others that somehow ties in with this research. Section 3 describes the principle and formula to measure test suite quality attributes. Section 4 describes the reliability analysis of test suite quality attribute by using Cohen's kappa coefficient. Section 5 describes the research methodology to validate the test suite quality attributes by using the level of agreement between the result of the measurement and expert assessment. Section 6 describes for experimental activity and its result. Section 7 explains the result of the questionnaire to test suite quality attributes measurement. Section 8 describes the conclusion and future work of the research.

II. Related Works

Test suite consists of a set of test scripts or test procedures known as test case to be executed in a specific test run[2]. Test script in test suite is related to the test case that consists of expected results based on the inputs. The difficulties in software testing quality especially in white-box testing approach vary depending on the size and complexity of the program being tested[5]. It was a great idea to measuring the degree of test suite quality by using the attributes from software quality. Several studies have been reported in the scope of quality attributes of test case that focus on increasing the testing effectiveness consider to mutation testing[6]. The usability especially identification error with effective and efficient is important to enhance software quality[7][8]. Efficiency of test suite is related to number of redundant test cases in the test suite and reducing redundant test cases possible to improve the efficiency in testing[9]. Reliability is considered to number of mutants because the result on mutants coverage could be used to find the true reliability of a program[10][11]. Functionality in the testing approach is to ensure the method in the program satisfies functional requirements and assesses the quality itself[12]. The study on test suite reusability is related to portability has been reported that in the test suite reuse effective at discovering and repairing bugs inserted during pragmatic reuse[13]. Reducing number of test cases and ability of the test cases reused to examine another object should be considered to improve the maintainability[14][15].

With respect to previous work, this research analyzes the quality attributes for test suite to ensure the quality of the test suite. As we already introduced, this proposal a method to measure the test suite quality attributes is adopted the quality attributes from software quality which related to test suite, especially in white-box testing approach.

III. Test Suite Quality Attributes

Software quality defines as the degree of software, component, or process to establish the customer requirement under specific conditions[2]. Successful software testing activity is achieved by collaborative activity between testing activity and quality assurance activities[16]. One of the important assets in testing activity is test suite. Software quality has many approaches such as McCall's Model (1977), Boehm's Quality Model (1978), ISO 9126 Standard Quality Model (1986), FURPS (1987), FURPS+ (2000), Capability Maturity Model (CMM 1991), Ghezzi Model (1991), IEEE Model (1993), Dromey's Quality Model (1995), SATC's Quality Model (1996), Bansiya's QMOOD Model (2002), Aspect-Oriented Software Quality Model (2006), Component-based Software development Quality Model (2008), DEQUALITE Model (2009), Sehra S. K Model (2011) and SQuaRE's Model (2011)[17].

Software quality attributes are multipurpose attributes that mean any area of software development process can use the attributes. Examining code programs by using the test suite is one of the methods to assure their quality. The most used attributes on the software quality model are usability, efficiency, reliability, functionality, portability, and maintainability that selected from 28 attributes. The principle of test suite quality attributes on white-box testing in this research is related to the software quality principle. The research proposes the following formula and definition for test suite quality attributes. To simplify the formula, the research uses the following notation.

- SRTC : Successful Reused Test Cases
- DCC : Distinct Code Coverage
- OT : Objects Tested
- NOLOC : Number of Original Line of code
- NOTC : Number of Test Cases
- NOMut : Number of Mutants
- NOMutK : Number of Mutants Killed
- NOR : Number of Redundant Test Cases
- NOMet : Number of Method
- NOMetExec : Number of Method Executed

The parameters for measuring the test suite quality attributes are gathering from the test suite examination that enhances the accuracy of test suite quality measurement. The result of test suite quality attribute measurement is numerical which ranges from 0 to 1. The experiment assumes that the test suite quality attribute has three levels of quality such as low, medium, and high. The result of test suite quality attributes divided into those three levels which begin from 0 – 0.33 for low quality, 0.34 – 0.66 for medium quality, and 0.67 – 1 for high quality. One of the criteria of good test cases in the test suite related to white box testing is that the test cases can achieve 100% code coverage. The test suite quality attribute measurement additionally considers code coverage on its measurement.

1) Usability as test suite quality attribute: Usability defines as the degree of a program able to be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use[18]. The research assumes that the test suite usability should consider for the effectiveness and efficiency related to previous definition. In other words, test suite usability defines as the extent to test suite is successfully examine program by the software tester to guarantee that all statements have been exercised at least once and validate the internal data structure with effectiveness and efficiency. By using the notation, the formula for test suite usability as follows.

$$Test\ Suite\ Usability = \frac{(1-(\mathrm{NOR}/\mathrm{NOTC}))+(DCC/NOLOC)}{2} \tag{1}$$

2) Efficiency as test suite quality attribute: Efficiency defines as the capability of the software product to provide appropriate performance, relative to the number of resources used under stated conditions[2]. Number of resources in the case of test suite is related to number of test cases. Redundant test cases in the test suite are one of the problems that can reduce efficiency value. The previous research conducted the identification and elimination process of redundant test cases in the test suite[19]. The efficiency in this research is related to the degree of redundancy of test cases on the test suite. The test suite efficiency defines as the level of test suite redundancy to complete a certain task. The redundant test cases exist when both of the two test cases are executed in the same lines of code. By using the notation, the formula for test suite efficiency as follows.

$$Test\ Suite\ Efficiency = \left(1 - (NOR/NOTC)\right) \quad (2)$$

3) Reliability as test suite quality attribute: Reliability defines as the ability of the software to operating required functions in specific conditions and time, or number of operations[2]. One of the causes of the inability of the software product to perform a required function is mutant. Mutants define as changed/mutated statements of the source code. The capability of the test cases kills the mutants to ensure the quality of test cases in terms of reliability. Test suite reliability is defined as the probability of test cases in the test suite killed the mutants in testing that consider to its coverage. By using the notation, the formula for test suite reliability as follows.

$$Test\ Suite\ Reliability = \frac{NOMut}{NOMutK} \quad (3)$$

4) Functionality as test suite quality attribute: The functionality defines as the capability of the software to perform functions that are related to user requirements with specific conditions[2]. The research analyzes the terms function in Java program related to the white-box testing is a method that also considers the coverage of the test suite. The test suite functionality is defined as the capability of test cases on the test suite to performs the behavior of the program. The test suite has performed the behavior of the java program when a high number of the method examines by the test suite. The formula for test suite functionality measurement as follows.

$$Test\ Suite\ Functionality\ = \frac{NOMetExec}{NOMet} \quad (4)$$

5) Portability as Test Suite Quality Attribute: Portability defines as the capability of software that can be reused from one hardware or software environment to another[2]. Test suite portability is defined as the capability of the test cases in the test suite to run on a new program without change. Portability is related to the degree of reusability of test suite. The test suite reusability defines as the capability of test cases in the test suite to examine several or all paths of method that should be tested on diverse objects.

This research uses the clones of Banker's Algorithm with code clones type 1, 2, 3, and 4 [20][21]. The test suite portability measurement is applied on code clones because the portability of the test cases in the test suite needs to use the same characteristic of input for the program. Code clone type 1 (exact clones) are identical clones with no differences with original code. Code clone type 2 which the differences from the original code are renamed identifiers, literals, types, layout, and comments but the structurally and syntactically are similar. Code clones type 3 are modified the statement such as statement insertions/deletions in addition to changes in identifiers, literals, types, and layouts. Code clone type 4 has been modified on code fragments to perform the same objective but different syntactic variants. By using the notation, the formula for test suite reusability as follows.

$$Test\ Suite\ Portability = \frac{\sum SRTC + \sum DCC}{(\sum OT \times \sum TC) + \sum NOLOC} \quad (5)$$

6) Maintainability as test suite quality attribute: Maintainability defines as the capability of software to be modified for correct defects, meet new requirements, make future maintenance easier, or adapted to a changing environment with less effort to maintain[2]. Test suite maintainability is related to the capability of the test suite that suitable to test another program with less effort to maintain by avoiding redundant test cases. The maintainability considers to reusability and efficiency(non-redundant test cases) of the test suite. By using the notation, the formula for test suite maintainability as follows.

$$Test\ Suite\ Maintainability$$

$$= \frac{1 - (NOR/NOTC) + \left(\frac{\sum SRTC + \sum DCC}{(\sum OT \times \sum TC) + \sum NOLOC}\right)}{2} \quad (6)$$

IV. Reliability Analysis

Reliability analysis in this research is to validate the formulas of test suite quality attributes measurement. Validation of those formulas is to observe the result with the real condition based on the expert assessment. Reliability analysis objective is to validate the level of agreement from the result of measurement to the expert [22]. The validation refers more specifically to the consistency of measurement that involves the expert.

Cohen's kappa coefficient is generally for assessing agreement between raters. Cohen defined the coefficient as "the proportion of chance-expected disagreements which do not occur, or the proportion of agreement after chance agreement is removed from consideration"[23]. Cohen's Kappa has used a quantitative measurement of reliability for two raters that are rating the same thing, corrected for how often that the raters may agree by chance. The formula for Cohen's kappa coefficient as follows.

$$k\ = \frac{p_0 - p_c}{1 - p_c} \quad (7)$$

where,

p_0 = the proportion of units for which the judges agreed (relative observed agreement among raters)

p_c = the proportion of units for which agreement is expected by chance (chance-expected agreement)

TABLE I. COHEN'S KAPPA CONTINGENCY MATRIX

		Rater 1	
		Category 1	Category 2
Rater 2	Category 1	a: number of agreements on category 1 P(a) = a/N	b: number of disagreements (judge 1 and category 2, and judge 2 and category 1) P(b) = b/N
	Category 2	c: number of disagreements (judge 1 and category 1, and judge 2 and category 2) P(c) = c/N	d: number of agreements on category 2 P(d) = d/N

TABLE II. COHEN'S KAPPA INTERPRETATION

Kappa Statistic	**Strength of Agreement**
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.0	Almost Perfect

Distribution of the frequency for two raters on Cohen's kappa coefficient is represented by the contingency matrix as shown in Table I [23]. Based on Cohen's kappa contingency matrix, p_o and p_c are calculated as follows:

p_o=P(a)+P(d) $P_{category1}$=(P(a)+P(c))*(P(a)+P(b))

p_c=$P_{category1}$+$P_{category2}$ $P_{category2}$=(P(b)+P(d))*(P(c)+P(d))

The value of Cohen's kappa coefficient is positive when the value greater-than-chance agreement and negative when less-than-chance agreement. The maximum value for Cohen's kappa coefficient is +1.0. Its value related to the strength of agreement as shown in Table II.

V. Research Methodology

This section explains the research methodology to measure the test suite quality attributes and then validate the result by using the reliability analysis. The reliability analysis is used Cohen's kappa coefficient to validate the result of test suite quality measurement based on the level of agreement from the result of measurement and expert assessment. Fig. 1 shows the test suite quality attributes measurement and validation activity that consists of two main activities such as test suite profiling for test suite quality measurement and expert assessment for validating the result of measurement.

A. Proposed Test Suite Quality Measurement Activity

The objective of test suite profiling is to collect the important and useful information of the test cases in the test suite for the test suite quality attributes measurement in white-box testing. Test suite profiling uses the Java program and given test suites then running the test suite which has been implemented on Junit to test the Java program and the result is test suite information. The test suite information contains such as follow.

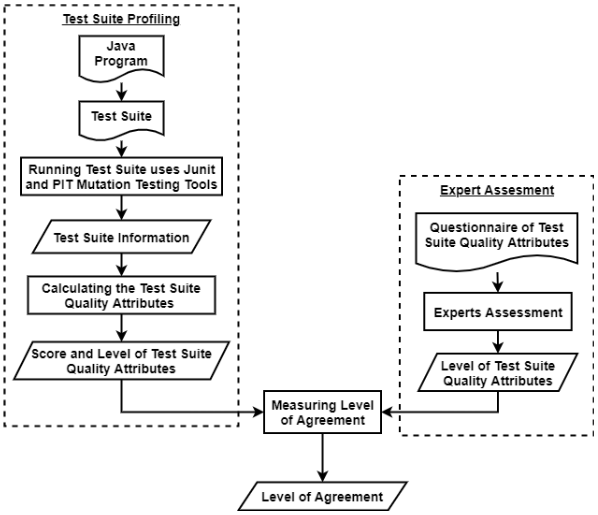

Test Suite Profiling
Java Program
Test Suite
Running Test Suite uses Junit and PIT Mutation Testing Tools
Test Suite Information
Calculating the Test Suite Quality Attributes
Score and Level of Test Suite Quality Attributes
Expert Assesment
Questionnaire of Test Suite Quality Attributes
Experts Assessment
Level of Test Suite Quality Attributes
Measuring Level of Agreement
Level of Agreement

Fig. 1. Test Suite Quality Attributes Measurement and Validation Activity.

a) Number lines of code,

b) Number lines code executed,

c) Number of test cases,

d) Distinct lines of code executed,

e) Number of mutants

f) Number of mutants killed by the test suite

g) Number of methods

h) Number of method executed.

Test suite information uses to calculate test suite quality attributes. Test suite quality attributes calculated in this research is usability, efficiency, reliability, functionality, portability, and maintainability. The result of the calculation is the score and level for test suite quality attributes.

B. Validation

Validation aims to prove the validity of test suite quality attributes measurement results with expert assessment. The expert assessment activity objective is to assess the quality of the test suite based on the experience from the expert. The questionnaire contains information and question that should answer by the expert such as how long the experience in software engineering and his work. The questionnaire provides information such as the explanation for test suite quality attributes, Java program and given test suites, and the result of test suite examination. The expert answers the question of test suite quality attributes assessment based on their experience, and knowledge. The expert assesses the quality of the test suite by choosing the level of quality such as low, medium, and high quality.

The level of test suite quality attribute from the test suite quality attributes measurement and expert assessment then analyze the reliability by using Cohen's kappa coefficient. The profiling and expert assessment for test suite quality attributes are used the same dataset. Cohen's Kappa coefficient and percentage of agreement are used to measure the level of agreement for test suite quality attributes. The result of measurement is the value of Cohen's kappa coefficient and percentage of agreement from test suite quality attributes measurement and expert assessment.

VI. The Experiment

The purpose of the experiment is to collect the test suite information for test suite quality attributes measurement for white-box testing. The result of measurement then validates the level of agreement from the result of measurement by using the result of expert assessment.

A. Dataset

The experiment uses Banker's Algorithm Java program as shown in Fig. 2 and given two test suites for the dataset as shown in Table III that contains the number of test case (TC), input data, and expected output. Banker's Algorithm is developed by Dijkstra for resource allocation and deadlock avoidance algorithm[11], [24]. The number of test cases is seven for each test suite. The test suites are implemented on Junit to examine the Banker's Algorithm. The test suite mutants information gains by using PIT mutation testing tool[25] and Junit for code coverage information.

B. Experiment Works

The Banker's Algorithm java program is examined by the given test suites in which test cases implemented in Junit. The test suite examination result is code coverage information and mutation coverage for profiling the test suite. The code coverage information for each test suite is collected by executing the Junit in Eclipse IDE that presents the information of lines of code executed by the test case in the test suite and percentage of code coverage.

The test suite mutants information is collected by using PIT mutation testing tool. PIT mutation testing tool are greatly manipulated bytecode to generates mutants and examine the test suite or test case to know the capability of the test suite to kill the mutants[25]. The mutants are killed by the test case in the test suite by showing different behaviour, and live when they are not. The ratio of mutants in the test suite is calculated in PIT mutation testing tool. The ratio of mutants is calculated by the mutants killed over the total number of mutants. The mutation score is used in test suite reliability measurement by combining with code coverage. Test suite quality attributes such as usability, efficiency, reliability, functionality, portability, and maintainability are calculated by using the formula that explains in Section 2.

The experiment uses the questionnaire to gathering the test suite quality attributes information from the expert assessment. The questionnaire contains the introduction and aims of the test suite quality attribute assessment, dataset, result of the test suites execution, and summary of the test suites examination. The question is focused on assessing an aspect of the level of test suite quality attributes from an expert view with seven questions. They are designed from the definition of test suite quality attributes and the result of all test suites examination. An example of the question is as follows.

```
public class Banker {
private final int need[][], allocate[][], max[][], avail[][], np, nr;
public Banker(int[][] need, int[][] allocate, int[][] max, int[][] avail, int np, int nr) {
 if (need.length != np || allocate.length != np || max.length != np || avail.length != 1|| need[0].length != nr ||
allocate[0].length != nr || max[0].length != nr || avail[0].length != nr) {
 throw new IllegalArgumentException("The matrices should have \"np\" rows and \"nr\" columns. \"avail\" should have only one
row.");}
 this.need = need;
 this.allocate = allocate;
 this.max = max;
 this.avail = avail;
 this.np = np;
 this.nr = nr;}
 private int[][] calc_need() {
  for (int i = 0; i < np; i++) {
   for (int j = 0; j < nr; j++) {
  need[i][j] = max[i][j] - allocate[i][j];}
 }
 return need;}
 private boolean check(int i) {
 for (int j = 0; j < nr; j++) {
  if (avail[0][j] < need[i][j]) {
  return false;}
 }
 return true;}
 public boolean isSafe() {
  calc_need();
        boolean done[] = new boolean[np];
        int j = 0;
        while (j < np) {
         boolean allocated = false;
         for (int i = 0; i < np; i++) {
         if (!done[i] && check(i)) {
          for (int k = 0; k < nr; k++) {
avail[0][k] = avail[0][k] - need[i][k] + max[i][k];}
 System.out.println("Allocated process : " + i);
 allocated = done[i] = true;
  j++;}
}
  if (!allocated) {
   break;}
}
return j == np;}
}
```

Fig. 2. Java Program for Banker's Algorithm.

TABLE III. TEST SUITES OF BANKER'S ALGORITHM

Test Suite-1	**TC-1**	Input data	int[][] intArray0 = new int[1][1];
			int[][] intArray1 = new int[1][1];
			int[] intArray2 = new int[1];
			intArray2[0] = 1;
			intArray1[0] = intArray2;
				Test Suite-2	**TC-1**	Input data	int[][] intArray0 = new int[1][1];
							int[][] intArray1 = new int[1][6];

		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")

	TC-2	Input data	int[][] intArray0 = new int[1][1];
			int[] intArray1 = new int[1];
			intArray1[0] = 1;
			intArray0[0] = intArray1;
					TC-2	Input data	int[][] intArray0 = new int[1][1];
							int[][] intArray1 = new int[1][1];
							int[] intArray2 = new int[9];
							intArray1[0] = intArray2;

		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")
	TC-3	Input data	int[][] intArray0 = new int[1][6];		**TC-3**	Input data	int[][] intArray0 = new int[0][6];
		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")

	TC-4	Input data	Banker banker0 = null;		**TC-4**	Input data	int[][] intArray0 = new int[1][6];
							int[][] intArray1 = new int[3][0];

		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")

	TC-5	Input data	int[][] intArray0 = new int[0][6];
			int[][] intArray1 = new int[1][0];
					TC-5	Input data	int[][] intArray0 = new int[0][6];
							int[][] intArray1 = new int[1][0];
							int[][] intArray1 = new int[1][0];

		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")

	TC-6	Input data	int[][] intArray0 = new int[1][1];
			int[][] intArray1 = new int[1][1];
			int[] intArray2 = new int[1];
			intArray2[0] = 1;
			intArray0[0] = intArray2;
					TC-6	Input data	int[][] intArray0 = new int[1][6];

		Expected Output	True or false for method isSafe()			Expected Output	fail("Expecting exception: IllegalArgumentException")

	TC-7	Input data	int[][] intArray0 = new int[1][1];
			int[][] intArray1 = new int[1][1];
			int[] intArray2 = new int[9];
			intArray1[0] = intArray2;
					TC-7	Input data	int[][] intArray0 = new int[11][6];

		Expected Output	fail("Expecting exception: IllegalArgumentException")			Expected Output	fail("Expecting exception: IllegalArgumentException")

"The efficiency of test suite usability in this research is related to the degree of redundancy of test cases on the test suite. The test suite efficiency defines as the level of test suite redundancy to complete a certain task. The redundant test cases exist when both of the two test cases are executed in the same lines of code.

Based on the definition and the summary of test suites examination, what is the degree of test suite efficiency for test suite 1 and test suite 2?".

The expert will choose one answer such as low, medium, high for each test suite. The questionnaire collects the personal information of the expert like name, kind of experience in software engineering. Result of the questionnaire is presenting the level of each test suite quality attributes of every test suite.

TABLE IV. RESULT OF TEST SUITE INFORMATION

Test Suite Information	Test Suite-1	Test Suite-2
Number lines of code	33	33
Distinct lines of code executed	33	3
Number of test cases	7	7
Number of mutants	41	41
Number of mutants killed	32	3
Number of methods	4	4
Number of methods executed	4	1
Number of object reused test suite(test cases)	4	4
Number of successful reused test cases for all object tested	28	28
Number of redundant test cases	3	6

TABLE V. RESULT OF TEST SUITE QUALITY ATTRIBUTES MEASUREMENT

Test Suite Quality Attributes	**Test Suite-1**		**Test Suite-2**	
	Score	Test Suite Quality Attribute Level	Score	Test Suite Quality Attributes Level
Usability	0.79	High	0.12	Low
Efficiency	0.57	Medium	0.14	Low
Reliability	0.78	High	0.07	Low
Functionality	1.00	High	0.25	Low
Portability	1.00	High	0.51	Medium
Maintainability	0.79	High	0.33	Low

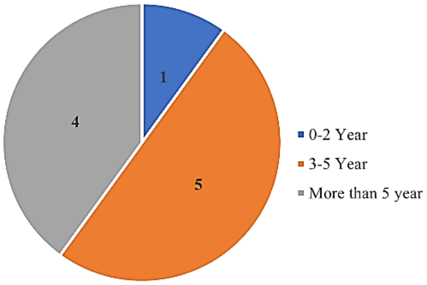


Fig. 3. Expert Time Experiences in Software Engineering.

The result of level test suite quality attributes and expert assessment is used to calculate the level of agreement by using Cohen's kappa coefficient and percentage of agreement. The contingency matrix for Cohen's kappa coefficient contains several conditions. True positive condition is the total number of instances that both raters said correct. False positive condition is the total number of instances that the result of test suite quality attribute measurement said incorrect, but experts said correct. False negative condition is the total number of instances that the result of test suite quality attribute measurement said correct, but the experts said incorrect. True negative condition is the total number of instances that both the result of test suite quality attribute measurement and the experts said incorrect. The value in the contingency matrix is used to calculate Cohen's kappa.

The percentage of agreement is calculated based on comparison data from the result of the level of test suite quality attributes measurement and expert assessment. Criteria of the match agreement if the result from the expert is the same as the measurement. The result from Cohen's kappa coefficient and percentage of agreement is to enrich the analysis of reliability in test suite quality attributes.

C. Experiment Result

The dataset from Banker's Algorithm and given test suites then examines by using Junit and PIT mutation testing tool, the result of test suite information is shown in Table IV. The test suite information consists of the number of lines of code, number of lines executed by the test case, distinct number of lines executed, number of methods, and number of method executed. Table V shows the result of test suite quality attributes measurement for usability, efficiency, reliability, functionality, portability, and maintainability by using the formula in Section 2 and their quality level.

The level of agreement is analyzing by Cohen's kappa coefficient and percentage of agreement. The Cohen's kappa coefficient in this research is to present the degree of agreement. The percentage of agreement is to enhance detailed information about the percentage of test suite quality attributes agreement. The number of experts is ten with different year experiences as shown in Fig. 3. Most of the experts have experience 3-5 years in software engineering. Time experience in software engineering from the expert helps to answer the question.

Calculation of Cohen's kappa coefficient uses the value of true positive, false positive, false negative, and true negative on contingency matrix from the result of the questionnaire as shown in Table VI. The percentage of agreement is measured with the proportion of experts who respond that identical or similar, lower or higher than the result of test suite quality attributes measurement as shown in Table VII. Similar terms in Table VII means that the result from the expert is the same as the measurement result, lower means that the result from the expert is lower than the measurement result, and higher means that the result from the expert is higher than the measurement result.

TABLE VI. CONTINGENCY MATRIX FOR ALL TEST SUITE AND EACH TEST SUITE

Cohen's Kappa Coefficient											
All Test Suite				Test Suite -1				Test Suite-2			
	Expert				Expert				Expert		
Measurement	Yes	No	Total	**Measurement**	Yes	No	Total	**Measurement**	Yes	No	Total
Yes	31	40	71	**Yes**	28	4	32	**Yes**	4	37	41
No	29	20	49	**No**	20	6	26	**No**	1	19	20
Total	60	60	120	**Total**	48	10	58	**Total**	5	56	61
Po	0.425			**Po**	0.586			**Po**	0.377		
Pc	0.5			**Pc**	0.533			**Pc**	0.356		
Kappa	-0.15	Less than Chance Agreement		**Kappa**	0.112	Slight Agreement		**Kappa**	0.032	Slight Agreement	

TABLE VII. RESULT OF PERCENTAGE OF AGREEMENT FOR ALL TEST SUITE AND EACH TEST SUITE

Quality Attributes	Similar		Lower		Higher	
	Total	Percentage	Total	Percentage	Total	Percentage
Usability	11	55%	6	30%	3	15%
Efficiency	6	30%	1	5%	7	35%
Reliability	10	50%	3	15%	7	35%
Functionality	13	65%	2	10%	5	25%
Portability	8	40%	7	35%	9	45%
Maintainability	4	20%	7	35%	9	45%
All Attributes	52	43%	26	22%	40	33%

VII. DISCUSSION

This section provides implications from the result of the questionnaire to test suite quality attributes measurement. The Cohen's kappa coefficient result for all test suites is – 0.15 which means less than chance agreement or no agreement as shown in Table VI. Negative value of Cohen's kappa coefficient is represented great disagreement between measurement and experts. Disagreement means that the observed agreement is less than chance agreement. There is no strict lower value for the kappa coefficient and its meaning. The weakness for a negative value of kappa has no fixed threshold for a lower value that difficult to have suitable interpretation from the assessment especially on the level of agreement.

The kappa coefficient is measured for each test suite. Table VI shows the result of kappa coefficient is 0.112 for test suite 1 and 0.032 for test suite 2 which has similar meaning is slight agreement. The slight agreement means that condition is needed to consider the result from the experts and measurements. Interpretation of the kappa coefficient for indicated the good agreement is not easy because in the terms of accuracy from a single kappa analysis itself.

The research analyzes that level of agreement by kappa coefficient is incompleted to represent the level of agreement between test suite quality attributes measurement and expert assessment. This research uses the percentage of agreement to complete the analysis. The percentage of agreement has improved the result of positive value of kappa coefficient. Table VII shows the percentage of agreement for all and each test suite quality attribute.

The result from all quality attributes measurement and the expert shows that 46% are similar to the test suite quality attributes measurement which means that the expert 46% agree with the principle of test suite quality attributes measurement. The highest attribute which similar is reliability with a percentage of 70%. The lowest attribute which similar is maintainability with a percentage of 20%.

The result of test suite quality attributes measurement for each test suite and expert assessment are confirmed for slight agreement. This result is approved by the result of the percentage of agreement. The research assumes that the level of agreement is strongly agreed when the percentage of the agreement greater than or equal to 50%. The result shows that usability with 55%, reliability with 50%, and functionality 65% which means that strongly agreed.

VIII. CONCLUSION AND FUTURE WORK

This research investigated the quality attributes for test suite based on the attributes of software quality. The attributes are usability, efficiency, reliability, functionality, portability, and maintainability that are selected from 28 attributes in software quality. The test suite quality attributes measurement uses the result of test suite examination as parameters and input. The experiment uses the Banker's Algorithm by using given two test suites. The result of test suite quality attributes measurement is presented by score and level of quality as the degree of test suite quality.

The experiment uses reliability analysis to prove the validity of test suite quality attributes measurement. The reliability analysis uses Cohen's kappa coefficient and percentage of agreement. Cohen's kappa coefficient analyzes the reliability of test suite quality based on test suite quality attributes measurement result and expert assessment. The result of Cohen's kappa coefficient measurement is – 0.15 for all test suites, 0.112 for test suite 1, and 0.032 for test suite 2. The result of test suite quality attributes measurement for each test suite is confirmed for slight agreement. This result is approved by using the result of the percentage of agreement. The research assumes that the level of agreement is strongly agreed when the percentage of the agreement greater than or equal to 50%. The result shows that usability with 55%, reliability with 50%, and functionality 65% which means that strongly agreed. Hence, our proposed method is useful to measure test suite quality attributes.

Our approach is a method to measure test suite quality attributes for white box testing which is specifically contained usability, efficiency, reliability, functionality, portability, and maintainability as attributes. In the future, it's important to consider the weight of each test suite quality attribute to define a formula that can measure test suite quality more accurately.

REFERENCES

[1] Lovely Professional University, Software Testing and Quality Assurance. New Delhi: Excel Books Private Limited, 2012.

[2] International Software Testing Qualifications Board (ISTQB), "Standard Glossary of Terms used in Software Testing Version 3.5," 2020.

[3] P. D. Roger S. Pressman, Software Engineering: A Practitioner's Approach, 7th ed. New York: McGraw-Hill, 2009.

[4] I. Sommerville, "Software Engineering 9th Edition., USA: Addison-Wesley Publishing Company, 133–170. 2010.

[5] Z. Nayyar, N. Rafique, N. Hashmi, N. Rashid, and S. Awan, "Analyzing Test Case Quality with Mutation Testing Approach," Proc. 2015 Sci. Inf. Conf. SAI 2015, pp. 902–905, 2015.

[6] G. Grano, “A new dimension of test quality: Assessing and Generating Higher Quality Unit Test Cases,” Proc. 28th ACM SIGSOFT Int. Symp. Softw. Test. Anal., pp. 419–423, 2019.

[7] M. M. Ali-Shahid and S. Sulaiman, “A Case Study on Reliability and Usability Testing Of A Web Portal,” in 2015 9th Malaysian Software Engineering Conference (MySEC), 2015, no. June, pp. 31–36.

[8] F. A. Muqtadiroh, H. M. Astuti, E. W. T. Darmaningrat, and F. R. Aprilian, “Usability Evaluation to Enhance Software Quality of Cultural Conservation System Based on Nielsen Model (WikiBudaya),” Procedia Comput. Sci., vol. 124, pp. 513–521, 2017.

[9] R. Ibrahim, M. Ahmed, R. Nayak, and S. Jamel, “Reducing Redundancy of Test Cases Generation Using Code Smell Detection and Refactoring,” J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 3, pp. 367–374, 2020.

[10] A. Dimov, S. K. Chandran, S. Punnekkat, A. Nasir, and N. Azam, “Mutation Testing Framework for Software Reliability Model Analysis and Reliability Estimation,” 6th Central and Eastern European Software Engineering Conference (CEE-SECR), pp. 163–169. 2010.

[11] G. Guizzo, F. Sarro, and M. Harman, “Cost Measures Matter for Mutation Testing Study Validity,” Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 1127–1139. 2020.

[12] Z. Q. Zhou, S. Xiang, and T. Y. Chen, “Metamorphic Testing for Software Quality Assessment: A Study of Search Engines,” IEEE Trans. Softw. Eng., vol. 42, no. 3, pp. 264–284. 2016.

[13] S. Makady and R. J. Walker, “Debugging and Maintaining Pragmatically Reused Test Suites,” Inf. Softw. Technol., vol. 102, no. March 2017, pp. 6–29, 2018.

[14] H. Ghandorh, A. Noorwali, A. B. Nassif, L. F. Capretz, and R. Eagleson, “A Systematic Literature Review for Software Portability Measurement,” Proceedings of the 2020 9th International Conference on Software and Computer Applications, pp. 152–157. 2020.

[15] D. Pfluger et al., “The Scalability-Efficiency/Maintainability-Portability Trade-Off in Simulation Software Engineering: Examples and a Preliminary Systematic Literature Review,” in 2016 Fourth International Workshop on Software Engineering for High Performance Computing in Computational Science and Engineering (SE-HPCCSE), pp. 26–34. 2016.

[16] I. C. Society, Guide to the Software Engineering Body of Knowledge Version 3.0 (SWEBOK Guide V3.0).

[17] Suman and W. Manoj, “A Comparative Study of Software Quality Models,” Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 4, pp. 1-8. 2014.

[18] ISO-Comitte, International Standard - ISO 9241-210. 2010.

[19] M. C. Saputra, T. Katayama, Y. Kita, H. Yamaba, K. Aburada, and N. Okazaki, “Test Cases Redundant Elimination on Code Coverage Uses Distance and Correlation Measurement Method,” Proc. Int. Conf. Artif. Life Robot., vol. 25, pp. 755–758. 2020.

[20] S. Bellon, R. Koschke, G. Antoniol, J. Krinke, and E. Merlo, “Comparison and Evaluation of Clone Detection Tools,” IEEE Trans. Softw. Eng., vol. 33, no. 9, pp. 577–591, Sep. 2007.

[21] C. K. Roy and J. R. Cordy, “Survey on Software Clone Detection Research,” in Technical Report No. 2007-541, 2007.

[22] R. T. Lange, “Inter-rater Reliability,” in Encyclopedia of Clinical Neuropsychology, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. New York, NY: Springer New York, pp. 1348–1348. 2011.

[23] J. Pérez, J. Díaz, J. Garcia-Martin, and B. Tabuenca, “Systematic Literature Reviews in Software Engineering Enhancement of The Study Selection Process Using Cohen’s Kappa Statistic,” J. Syst. Softw., vol. 168, p. 110657. 2020.

[24] “GitHub - iguit0/BankersAlgorithm: Dijkstra’s famous algorithm.” [Online]. Available: https://github.com/iguit0/BankersAlgorithm. [Accessed: 08-Apr-2021].

[25] H. Coles, T. Laurent, C. Henard, M. Papadakis, and A. Ventresque, “Demo: PIT A Practical Mutation Testing Tool for Java (demo),” in Proceedings of the 25th International Symposium on Software Testing and Analysis, 2016, pp. 449–452.

Intelligent Scroll Order Generator Software from View Movements in People with Disabilities

Juan Ríos-Kavadoy[1], Harold Guerrero-Bello[2], Michael Cabanillas-Carbonell[3]
Facultad de Ingeniería
Universidad Privada del Norte
Lima, Perú

***Abstract*—People with motor disabilities face problems such as being able to move around independently, as well as having difficulties to take advantage of the technological tools developed for their rehabilitation. This research is based on computer vision and robotics, and was carried out with the objective of generating displacement orders using smoothed and binarized algorithms to assist the displacement of disabled people. The intelligent software for people with disabilities or motor deficiencies generates movement commands for an acceptable time through visual commands made by the person, by means of communication between the camera and the software, constantly capturing images. The results obtained for the scene (image) is cropped according to the face scene, then the view scene is cropped according to the face scene, and by means of necessary algorithms the pupil must be found; The complexity lies not only in locating the pupil, but also in identifying when a command is being sent and when it is not. Finally, the unit processes the movement command (left and right) to turn the LEDs on and off.**

Keywords—Displacement; motor disability; pupil; computer vision of images

I. Introduction

According to the UN and WHO [1]–[3], around the world, more than 1 billion people have disabilities, about 15% of the world's population, of whom more than 200 million suffer from motor difficulties, being 3.8% considered to be severely disabled. According to Hawking [3], who has suffered from motor neuropathy since adulthood, he states that this motor disability has not been an obstacle to achieve his goals, being technology his support to lead a normal life.

In Peru [4], there are different types of diseases that prevent people with disabilities from carrying out their activities normally, representing 59.2% of the total population with disabilities, preventing them from moving from one place to another, climbing stairs, going down stairs, etc. In everyday life, orders or actions are performed by different means, such as voice, for example, to activate the cell phone, open/close the door of the house, etc., and all that the world of home automation has to offer. But little in eye tracking is to give the same orders as voice, hand, etc.

Motor difficulty (movements in the hands, feet, etc.) also called as coordination disorder which prevents the generation of movement commands to the body due to chronic neurological impairment [5], [6]. There are different options to be able to capture the brain's orders, one of which is by means of nerve impulses, brain waves where physiological orders can be carried out by means of these [7], [8], another option is to use eye movements by means of visual perception.

The interaction between the brain and the computer allows the patient with motor disabilities to increase their cognitive abilities and self-sufficiency; however, accuracy problems are often encountered [7]. To solve these problems we propose the use of eye movements translated by the computer into physiological movements by training algorithms for greater accuracy and for the efficient operation of external devices such as wheelchairs.

Taking the last mentioned, the present project is intended to cover 02 areas of Artificial Intelligence (digital image processing [9] and robotics [10]), two problems were encountered to carry out the orders by means of the view, the first one is to be able to discriminate the orders of those that are not, considering that at a certain moment it can be visualizing a landscape, but not generating an order, therefore it is desired to find the recognition of orders, and second problem encountered is that the orders of displacement carried out are correct, carrying out the action in the indicated direction.

As a background to the study, the following studies have been considered, an important contribution presented in Panama, is the SOLCA software [11], which supports people with disabilities in the use of hands, for this purpose is agency of sight to be able to manipulate a computer. Another relevant contribution presented in USA [12], is the design of a virtual environment used as a simulator contributing to be a support tool in the rehabilitation of people with some degree of motor disability, based on virtual reality. It was obtained that increased confidence in patients.

In research [13], a drawing interface for human-computer interaction was realized where image processing was used, interface control was developed by tracking the patient's nose. As a result, it was found that young patients found the system easy to use, but in adult patients there was a lack of familiarity with the software.

Although there have been several previous investigations on the case, carried out with a variety of methods with the aim of helping the patient in his rehabilitation, it has been identified that most of the patients show some difficulty in adaptation, this is due to the lack of familiarity of the patient, to the unfriendly environments of the interfaces that cause the disabled person to make some physical effort.

The present research aims to provide a trained algorithm from which a digital image processing was performed, typical of artificial intelligence, where information has to be extracted from a scene, in order to be implemented in the future to a wheelchair. It also has an applicative justification since it works with hardware that is used for people with motor disabilities, as valuation we have that it can be scalable and can be used by a greater number of people and also has a social justification, since this research is born from a problematic reality for the support of people with motor deficiency providing a non-invasive and easy to use support for the patient.

II. Methodology

The main functions to be performed are to locate the eyes and analyze the position of the pupil within the central position of the eye by listening to the left (i) and right (d) commands, to subsequently switch on the LED defined for the left and right position and start processing the scene and improving the captured image, finally identifying the commands.

The research conducted is of a pre-experimental type, which allows to optimally administer the treatment to a group [14], applying a stimulus (intelligent software) for the resolution of the problem.

A. *Population and Sample (Materials, Instruments and Methods)*

The population is made up of 100 realized visual orders, as trials the orders are left and right. Therefore, the object of study is given by the visual orders. To obtain these visual orders, 5 attempts of order of each side are considered, performed to 10 persons, therefore 10 attempts x 2 sides x 10 persons, in total 200 orders of displacement, which will be submitted to evaluation of the certainty.

B. *Techniques and Instruments for Data Collection and Analysis*

1) *Data collection*

a) Transfer technique: It consists of recording the data that are obtained in the instruments called files, which are duly prepared and organized and contain most of the information collected in a research, and therefore constitute a valuable auxiliary instrument in this task [15].

b) Record card: Evaluation card of visual displacement orders generated and evaluated according to their result.

c) Validity: The degree to which an instrument actually measures the variable it intends to measure.

d) Data analysis: Applied research refers to scientific study that seeks to solve practical problems [16]. It is used to find solutions to everyday problems, cure diseases and develop innovative technologies.

C. *Procedure*

For the present investigation, two techniques will be taken into account: Observation and Record Card, considered for the acquisition of data, as shown in the data table, where the data obtained from the attempts to generate visual orders are observed, identifying which of them have really been visual orders.

1) *Data Collection Before using the Intelligent Software*

a) Indicador 1: Rate of recognition of travel orders. Data collection before using smart software for the travel order recognition rate indicator (see Table I). Correctly Generated Visual Commands (CGCOs).

b) Indicator 2: Orders performed correctly. Data collection before using smart software for indicator correct orders made (see Table II).

TABLE I. Pre-Test Data of the Travel Order Recognition Index Indicator

Test case	**Attempts**	**Total Attempts**	**CGCO**	**Recognition Rate**	**Recognition Rate (%)**
Person 1	Attempts left	10	6	0.60	60%
	Attempts right	10	4	0.40	40%
Person 2	Attempts left	10	6	0.60	60%
	Attempts right	10	7	0.70	70%
Person 3	Attempts left	10	6	0.60	60%
	Attempts right	10	4	0.40	40%
Person 4	Attempts left	10	7	0.70	70%
	Attempts right	10	7	0.70	70%
Person 5	Attempts left	10	8	0.80	80%
	Attempts right	10	6	0.60	60%
Person 6	Attempts left	10	7	0.70	70%
	Attempts right	10	4	0.40	40%
Person 7	Attempts left	10	7	0.70	70%
	Attempts right	10	4	0.40	40%
Person 8	Attempts left	10	7	0.70	70%
	Attempts right	10	8	0.85	85%
Person 9	Attempts left	10	8	0.80	80%
	Attempts right	10	4	0.40	40%
Person 10	Attempts left	10	4	0.40	40%
	Attempts right	10	6	0.60	60%
TOTAL(TO)		200	48		

TABLE II. PRE-TEST DATA FOR THE NUMBER OF CORRECT ORDERS RECOGNIZED INDICATOR

Measurement	Attempts	Total Attempts	CO	Visual Orders	Correct Visual Commands (%)
Person 1	Attempts left	6	2	0.3	30%
	Attempts right	4	1	0.3	30%
Person 2	Attempts left	6	3	0.5	50%
	Attempts right	7	2	0.3	30%
Person 3	Attempts left	6	5	0.8	80%
	Attempts right	4	1	0.3	30%
Person 4	Attempts left	7	2	0.3	30%
	Attempts right	7	3	0.4	40%
Person 5	Attempts left	8	3	0.4	40%
	Attempts right	6	3	0.5	50%
Person 6	Attempts left	7	4	0.6	60%
	Attempts right	4	0	0.0	0%
Person 7	Attempts left	7	2	0.3	30%
	Attempts right	4	0	0.0	0%
Person 8	Attempts left	7	0	0.0	0%
	Attempts right	8	3	0.4	40%
Person 9	Attempts left	8	2	0.3	30%
	Attempts right	4	0	0.0	0%
Person 10	Attempts left	4	0	0.0	0%
	Attempts right	6	2	0.3	30%
TOTAL(OTR)		120	38		

2) Development of the intelligent software methodology

This section describes the phases to develop a visual command generator system, using computer vision and digital image processing with Python software and Arduino hardware making use of algorithms and functions from the OpenCV library. This research project is developed under the cascading life cycle [17], through the following stages:

a) Analysis: To identify the visual commands, an analysis of pupil movement was performed, using digital image processing to locate the eyes on the face and fix a pupil position within the central position of the eye, as well as to generate visual commands.

b) Design: In this phase the methods and techniques are determined in the different phases of image processing for the generation of visual orders.

3) Scene capture: An 18 Megapixel camera was used to capture the scene. The process is shown in Fig. 1.

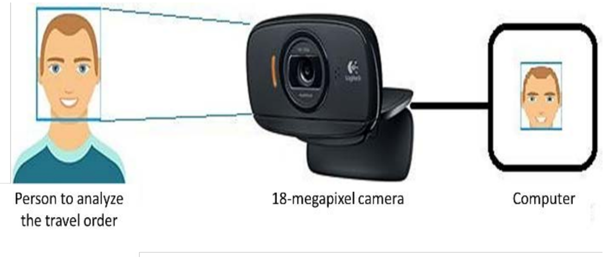


Fig. 1. Process of Capturing the Scene of the Face.

- Interface for software operation: Interface shows the capture of the scene by the camera, to process the image, to find the face, the eye and the pupil. Fig. 2 shows the precise moment when the command is given to the left side. Note that the left side is with respect to the screen, but not to the person.

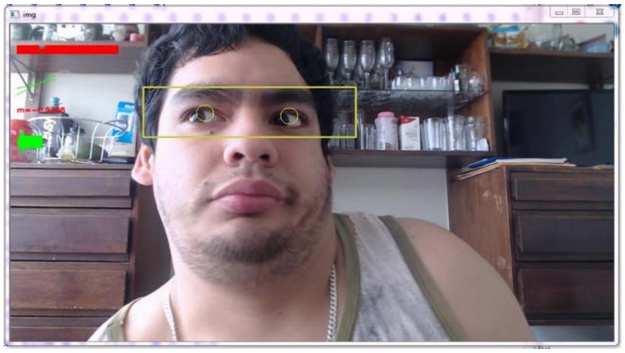


Fig. 2. Scene Interface as it Generates Visual Order.

- Fig. 3 shows the diagram of the hardware device made for this project.

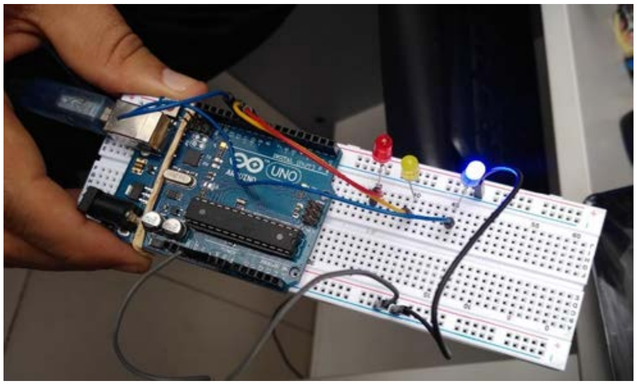


Fig. 3. Arduino Hardware Developed for the Project.

4) Processing: A flow chart was made to explain the process to send an order captured by the camera. (see Fig. 4)

5) Segmentation: It allows to highlight the scene in gray tone, necessary to be able to analyze the position of the pupil, in Fig. 6, it can be seen at the end of the process as the RGB graph changed to gray tone. The following Python source code was used for this purpose:

```
gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
```

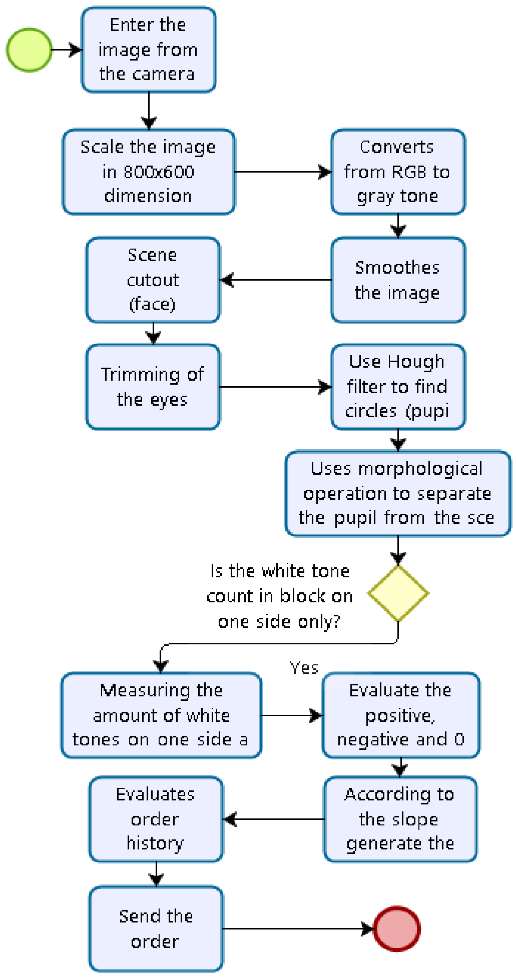

Enter the image from the camera
Scale the image in 800x600 dimension
Converts from RGB to gray tone
Smoothes the image
Scene cutout (face)
Trimming of the eyes
Use Hough filter to find circles (pupi
Uses morphological operation to separate the pupil from the sce
Is the white tone count in block on one side only?
Yes
Measuring the amount of white tones on one side a
Evaluate the positive, negative and 0
According to the slope generate the
Evaluates order history
Send the order

Fig. 4. Flow Chart of the Order Generation Process.

6) Description: Feature extraction technique. This is necessary to identify which is the necessary technique for the extraction of particular characteristics in a scene, and by means of it to locate the object in question helping to make a decision, for it is supported by the hough model and thus to be able to find the pupil.

First the image was obtained in gray tone, then the search was performed by the circular shape of the pupil (Hough filter) for this a morphological operation was used to extract the pupil from the environment, the amount of white tones at the ends of the pupil was counted. Finally, a slope line was generated to give order (+,-,0) = (d,i,e), this process is shown graphically in Fig. 10 and 11.

7) Visual order recognition: For the identification of movement orders (see process in Fig. 5), the number of pixels obtained as a gray tone (previously binarized) must be taken into account, therefore, it can be counted by means of the number of gray tones 1 (or 255) and 0 (or 0), for each column of the image, which in turn can be represented by the statistical method of dispersion, finding the line that represents that set of statistical values and thus finding the slope of that line created.

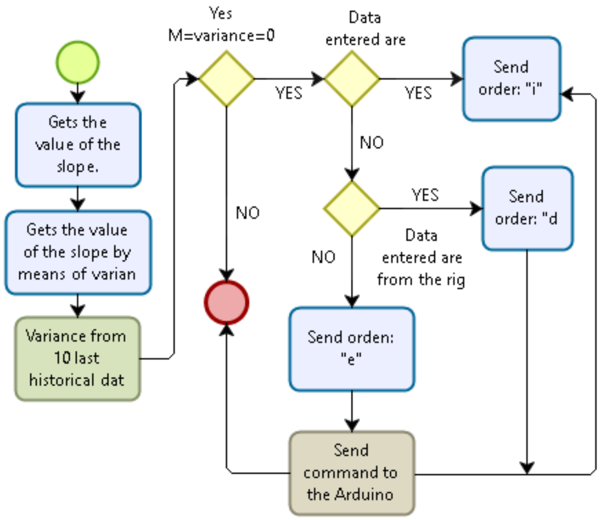

Yes
M=variance=0
Data
entered are
YES
YES
Send order: "i"
NO
Gets the value of the slope.
Gets the value of the slope by means of varian
Variance from 10 last historical dat
NO
YES
Send order: "d
Data entered are from the rig
NO
Send orden: "e"
Send command to the Arduino

Fig. 5. Diagram of Visual Order Identification.

Considering the values of the orders of the slopes generated by the line of the white shades, to consider an order it is necessary to have a statistic of them.

- Order classification using digital image processing: To recognize the visual order it is necessary to consider the amount of gray tone data that are blank in order to identify them, then use the variance method for the amount of orders.

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{N-1}$$

Being X1, X2,...,Xn, the set of N data, and x is the mean of these data. It is analyzed if S is 0, therefore, there is no variance, this implies that there is an intention to give an order, but if the data are different from 0 it implies that it has different types of orders, therefore, it does not generate an established order; this may be because it is looking at one side and then the other, etc.

Fig. 6 shows that the first step is to analyze whether there is a "Face" in the scene, for this it is necessary to locate and cut the scene in its position and size. Then it is necessary to reduce the scene to locate the eyes and detect the order that is performing, we proceed to the segmentation of the face, from the middle of the face upwards. In the third step of the process, the eyes are extracted in their position and size. Then the cut is made to perform the analysis. Finally, the pupil is identified to see its position and size, then it is changed to gray tone and the order is processed.

8) Implementation: For the implementation of the system based on digital image processing for the generation of visual orders, the following was considered:

a) Applied Technologies: Python 3.7.5 and OpenCV

b) Tests: The tests have been performed by means of the black box to measure the results of the application, detailing below with an example of test performed specified in Fig. 10 and 11.

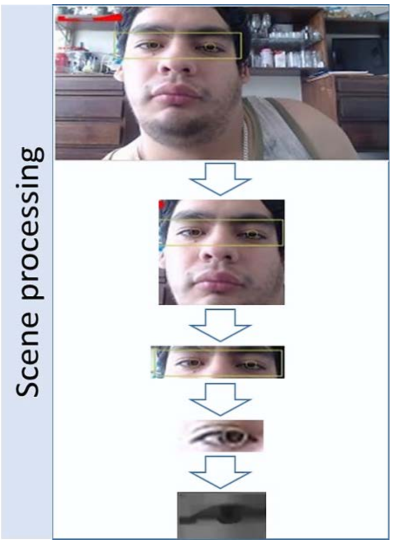


Fig. 6. Architecture for the Identification of the Visual Order.

Fig. 7, 8 and 9 detail the process for obtaining the evidence after the order is sent.

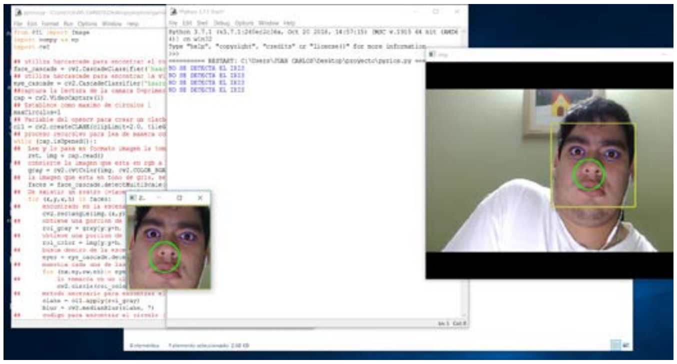
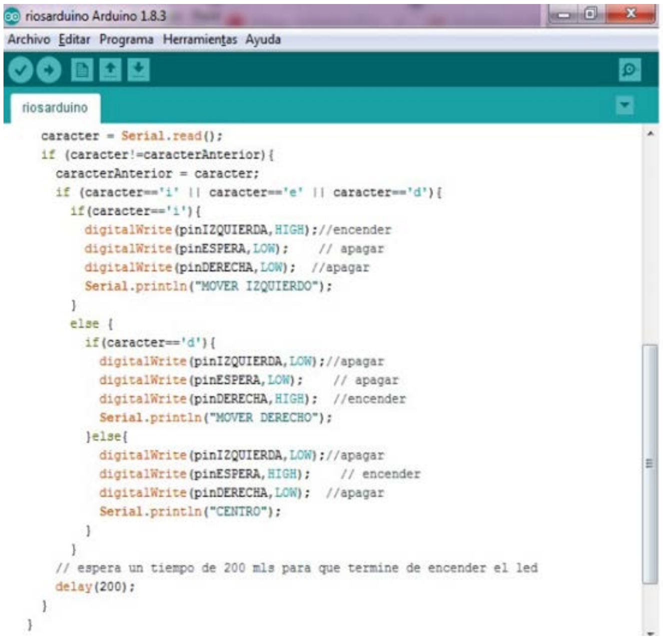


Fig. 7. Capturing the Scene of the Face - Initial Part.


riosarduino Arduino 1.8.3
Archivo Editar Programa Herramientas Ayuda
riosarduino

```
    caracter = Serial.read();
    if (caracter!=caracterAnterior){
      caracterAnterior = caracter;
      if (caracter=='i' || caracter=='e' || caracter=='d'){
        if(caracter=='i'){
          digitalWrite(pinIZQUIERDA,HIGH);//encender
          digitalWrite(pinESPERA,LOW);    // apagar
          digitalWrite(pinDERECHA,LOW);  //apagar
          Serial.println("MOVER IZQUIERDO");
        }
        else {
          if(caracter=='d'){
            digitalWrite(pinIZQUIERDA,LOW);//apagar
            digitalWrite(pinESPERA,LOW);    // apagar
            digitalWrite(pinDERECHA,HIGH);  //encender
            Serial.println("MOVER DERECHO");
          }else{
            digitalWrite(pinIZQUIERDA,LOW);//apagar
            digitalWrite(pinESPERA,HIGH);    // encender
            digitalWrite(pinDERECHA,LOW);  //apagar
            Serial.println("CENTRO");
          }
        }
      // espera un tiempo de 200 mls para que termine de encender el led
      delay(200);
    }
  }
```

Fig. 8. Switching on the LED in Order.

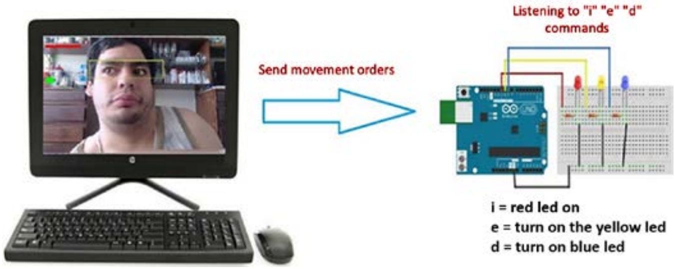
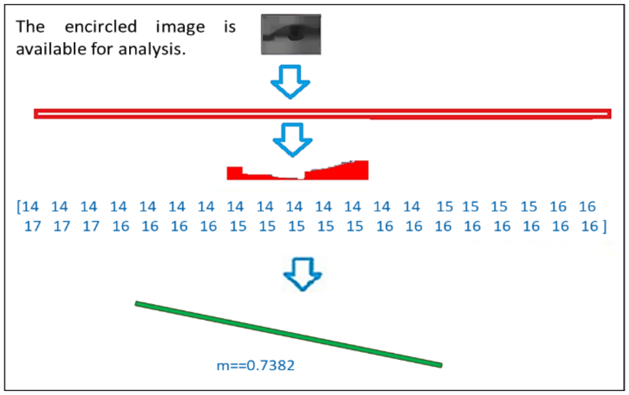


Fig. 9. Listening to Orders.

- Case assessed to person 1:

TABLE III. EVALUATION OF THE GENERATION OF LEFT-HAND MOVEMENTS

Measurement	Attempts made	Correct action
Person 1	Attempt left	Attempt left



The encircled image is available for analysis.

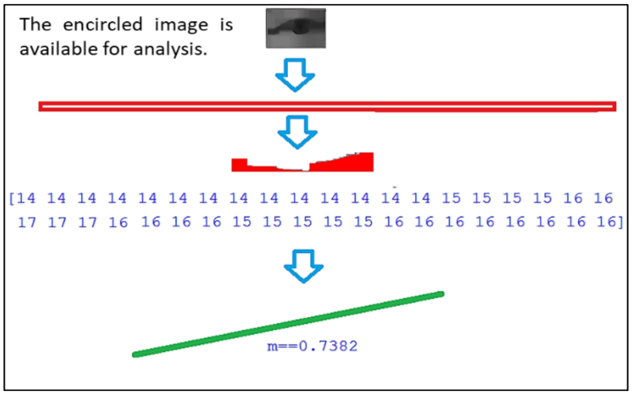
[14 14 14 14 14 14 14 14 14 14 14 14 14 14 15 15 15 15 16 16 17 17 17 16 16 16 16 15 15 15 15 15 16 16 16 16 16 16 16 16]

Fig. 10. Result Obtained by Performing the Test to Generate Visual Order to the Left.

- Case assessed to person 2:

TABLE IV. EVALUATION OF THE GENERATION OF RIGHT-HAND MOVEMENTS

Measurement	Attempts made	Correct action
Person 2	Attempt right	Attempt right



The encircled image is available for analysis.

[14 14 14 14 14 14 14 14 14 14 14 14 14 14 15 15 15 15 16 16 17 17 17 16 16 16 16 15 15 15 15 15 16 16 16 16 16 16 16 16]

Fig. 11. Result obtained by Performing the Test of Generating Visual Order to the Right.

Regarding the slope (m), its inclination is evaluated, and for the different experiments it is considered that the value of -0.2 or less indicates that it will perform the order to go to the right, values greater than 0.1 indicate that it will perform orders to the left, for any other case, it indicates that it will not perform any order.

9) Data collection after using the intelligent software:

a) Indicator 1: Index of recognition of travel orders. Data collection after using the intelligent software for the indicator travel order recognition index specified in Table V.

TABLE V. POST-TEST DATA FOR THE TRAVEL ORDER RECOGNITION INDEX INDICATOR

Test case	Attempts	Total Attempts	CGCO	Recognition Rate	Recognition Rate (%)
Person 1	Attempts left	10	9	0.9	90%
	Attempts right	10	8	0.8	80%
Person 2	Attempts left	10	9	0.9	90%
	Attempts right	10	9	0.9	90%
Person 3	Attempts left	10	9	0.9	90%
	Attempts right	10	8	0.8	80%
Person 4	Attempts left	10	9	0.9	90%
	Attempts right	10	8	0.8	80%
Person 5	Attempts left	10	10	1	100%
	Attempts right	10	8	0.8	80%
Person 6	Attempts left	10	8	0.8	80%
	Attempts right	10	6	0.6	60%
Person 7	Attempts left	10	8	0.8	80%
	Attempts right	10	6	0.6	60%
Person 8	Attempts left	10	8	0.8	80%
	Attempts right	10	10	1	100%
Person 9	Attempts left	10	10	1	100%
	Attempts right	10	8	0.8	80%
Person 10	Attempts left	10	9	0.9	90%
	Attempts right	10	9	0.9	90%
TOTAL(TO)		200	169		

b) Indicator 2: Orders placed correctly. Data collection after using the intelligent software for the indicator of correct orders placed specified in Table VI.

Finally, calculations were made based on the formulas of the indicators.

- For indicator: Travel Order Recognition Rate

$$IRO = \frac{CGCO}{TO} x100 = \frac{169}{200} = 0.85$$

IRO: Index of Recognition of Travel Orders

CGCO: Correctly Generated Visual Commands.

TO: Total orders

- For the indicator: Correct visual orders carried out.

$$CVC = \frac{CO}{TOP} = \frac{145}{169} = 0.85$$

CVC: Correct visual orders.

TABLE VI. POST-TEST DATA FOR CORRECT VISUAL ORDERS INDICATOR

Test case	Attempts	Total Attempts	CGCO	Recognition Rate	Recognition Rate (%)
Person 1	Attempts left	9	8	0.9	90%
	Attempts right	8	8	1.0	100%
Person 2	Attempts left	9	8	0.9	90%
	Attempts right	9	5	0.6	60%
Person 3	Attempts left	9	9	1.0	100%
	Attempts right	8	6	0.8	80%
Person 4	Attempts left	9	8	0.9	90%
	Attempts right	8	6	0.8	80%
Person 5	Attempts left	10	9	0.9	90%
	Attempts right	8	6	0.8	80%
Person 6	Attempts left	8	8	1.0	100%
	Attempts right	6	4	0.7	70%
Person 7	Attempts left	8	8	1.0	100%
	Attempts right	6	5	0.8	80%
Person 8	Attempts left	8	8	1.0	100%
	Attempts right	10	8	0.8	80%
Person 9	Attempts left	10	6	0.6	60%
	Attempts right	8	8	1.0	100%
Person 10	Attempts left	9	8	0.9	90%
	Attempts right	9	9	1.0	100%
TOTAL(TO)		169	145		

TOP: Total orders placed.

CO: Correct Orders

III. Results

A. Descriptive Analysis

In the study, an intelligent software (Independent Variable) was applied to evaluate the order recognition index and the correct orders carried out for the displacement order (Dependent Variable); for this purpose, a Pre-Test was applied to know the initial conditions of the indicator; subsequently, the intelligent software was implemented and the order recognition index and the correct orders carried out for the displacement order were recorded again.

1) Order recognition index indicator: The descriptive results of the order recognition index. In the pre-test a value of 60.25% was obtained, while in the post-test it was 84.5% as shown in Fig. 14; this indicates a great difference before and after the implementation of the intelligent software (see Fig. 12).

Regarding the dispersion of the order recognition index, in the pre-test there was a variability of 0.15; however, in the post-test there was a value of 0.11.

2) Indicator - Number of correct orders placed: The descriptive results of the number of correct realized orders. In the case of the number of correct orders to generate displacement orders, in the pre-test a value of 29% was obtained, while in the post-test it was 86% as shown in Fig. 15; this indicates a great difference before and after the implementation of the intelligent software (see Fig. 13).

Regarding the dispersion of the number of correct orders, in the pre-test there was a variability of 0.21; however, in the post-test there was a value of 0.13.

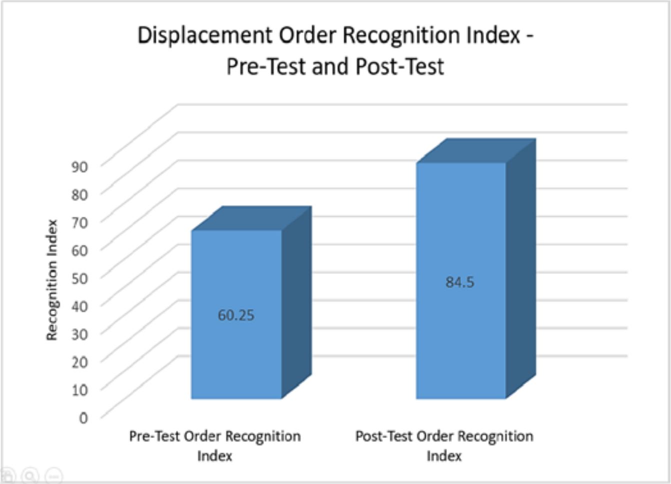


Fig. 12. Displacement Order Recognition Index - Pretest and Posttest.

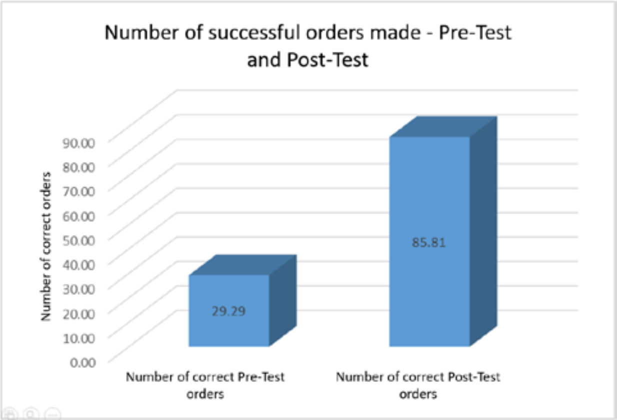


Fig. 13. Number of Correct Orders Placed, Pretest and Posttest.

B. Order Recognition Index

The order recognition index according to the data in Table V and Table III is shown in Fig. 14.

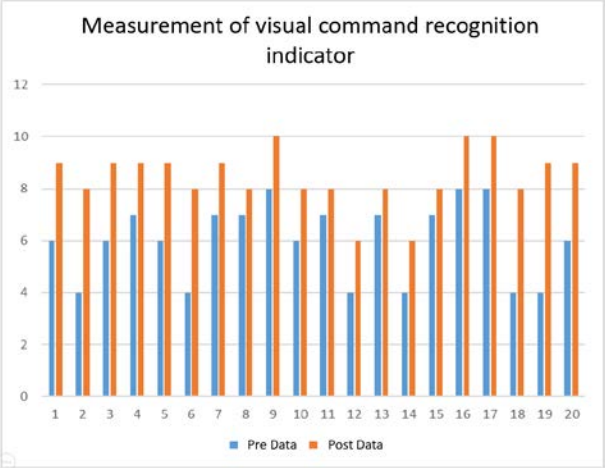


Fig. 14. Measurement of the Visual Command Recognition Indicator.

1) Indicator of correct visual orders: The number of orders recognized according to the data in Tables VI and IV is shown in Fig. 15.

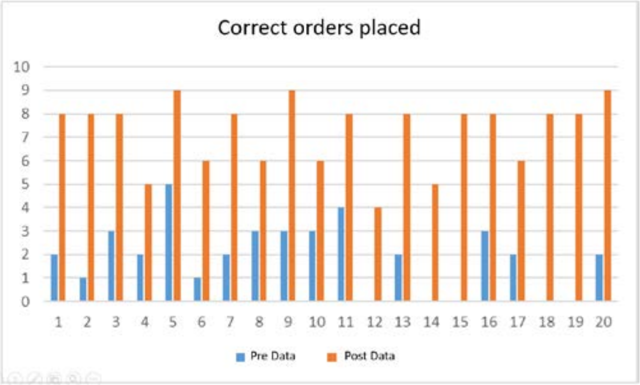


Fig. 15. Measurement of Order Recognition Rate.

IV. Discussion and Conclusion

The influence of intelligent software on the intelligence degree dimension of visual command recognition for people with disabilities according to the visual command recognition rate of 169 correct recognitions out of 200; therefore, using the appropriate model and algorithms results in a visual command recognition rate of 0.85. The influence of intelligent software on the visual order dimension of recognizing the correct orders was demonstrated for people with disabilities according to the rate of recognition of correct visual commands obtained in the tests, which were 145 correct recognitions out of 169; therefore, it is obtained with respect to the rate of generation of correct visual commands of 0.86.

Based on the results obtained, we can infer that the solution can be applied to another person with a disability. Based on the results obtained, we can infer that the solution can be applied to another person with a disability. In addition, the intelligent software had a positive influence on the recognition of movement commands from movements.

This prototype serves as a basis for new knowledge concerning digital image processing and external devices such as sensors.

Recommendations

The present research can be used as a background to find the order recognition rate taking as a reference an effectiveness rate of 0.845 and a rate of correct orders placed of 0.858 plotted.

It is recommended to use a general model to be applied in software for anyone who wants to use it, regardless of whether they have a motor disability, as well as using devices such as Raspberry Pi, to try to become independent with a portable hardware.

References

[1] UN, “Disability and Development Report: Realizing the Sustainable Development Goals by, for and with persons with disabilities.,” New York, United States of America, 2019. doi: 10.4337/9781847202864.00035.

[2] UN, “Disability and Development Report: Realizing the SDGs by, for and with persons with disabilities. Chapter 2.,” New York, United States of America, 2019.

[3] Organización Mundial de la Salud and Banco Mundial, “Disability and Development Report,” 2011. [Online]. Available: http://www.who.int/disabilities/world_report/2011/report/en/.

[4] INEI, “Perú: Características de la Población con Discapacidad,” Lima - Peru, 2015. [Online]. Available: https://www.inei.gob.pe/.

[5] R. Li et al., “Automated fine motor evaluation for developmental coordination disorder,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 5, pp. 963–973, 2019, doi: 10.1109/TNSRE.2019.2911303.

[6] M. Skarsgard, S. C. D. Dobri, D. Samdup, S. H. Scott, and T. C. Davies, “Toward Robot-Assisted Diagnosis of Developmental Coordination Disorder,” IEEE Robot. Autom. Lett., vol. 4, no. 2, pp. 346–350, 2019, doi: 10.1109/LRA.2018.2885197Y.

[7] K. J. Wang et al., “Brain-computer interface combining eye saccade two-electrode EEG signals and voice cues to improve the maneuverability of wheelchair,” IEEE Int. Conf. Rehabil. Robot., pp. 1073–1078, 2017, doi: 10.1109/ICORR.2017.8009392.

[8] J. L. Collinger et al., “High-performance neuroprosthetic control by an individual with tetraplegia,” Lancet, vol. 381, no. 9866, pp. 557–564, 2013, doi: 10.1016/S0140-6736(12)61816-9.

[9] M. Tang, “Image Segmentation Technology and Its Application in Digital Image Processing,” IEEE, 2020, doi: 10.1109/ICAACI50733.2020.00040.

[10] G. Quere et al., “Shared Control Templates for Assistive Robotics,” Proc. - IEEE Int. Conf. Robot. Autom., pp. 1956–1962, 2020, doi: 10.1109/ICRA40945.2020.9197041.

[11] A. Broce, “ARQUITECTURAS COMPARTIDAS,” 2018. http://solca.innovacion.gob.pa/wp-content/uploads/2019/04/1-ARQUITECTURAS-COMPARTIDAS.pdf (accessed Mar. 29, 2021).

[12] T. I. Chowdhury, “Towards Reverse Disability Simulation in a Virtual Environment,” 25th IEEE Conf. Virtual Real. 3D User Interfaces, VR 2018 - Proc., pp. 803–804, 2018, doi: 10.1109/VR.2018.8446146.

[13] S. S. Khan, M. S. H. Sunny, M. S. Hossain, E. Hossain, and M. Ahmad, “Nose tracking cursor control for the people with disabilities: An improved HCI,” 3rd Int. Conf. Electr. Inf. Commun. Technol. EICT 2017, vol. 2018-January, no. December, pp. 1–5, 2018, doi: 10.1109/EICT.2017.8275178.

[14] C. Manterola, G. Quiroz, P. Salazar, and N. García, “Metodología de los tipos y diseños de estudio más frecuentemente utilizados en investigación clínica,” Rev. Med. Clin. Condes, vol. 30, no. 1, pp. 36–49, 2019, [Online]. Available: https://doi.org/10.1016/j.rmclc.2018.11.005.

[15] F. García, “Ingeniería en Software,” pp. 277–388, 2018, [Online]. Available: https://repositorio.grial.eu/bitstream/grial/1228/1/07-rep.pdf.

[16] C. Escudero and L. Cortez, “Técnicas y Métodos cualitativos para la ivestigación científica. Cap. 1,” in Técnicas y métodos cualitativos para la investigación científica., First edit., vol. 6, no. 11, Editorial UTMACH, 2018, pp. 951–952.

[17] I. SOMMERVILLE, “Ingenieria del Software 7ma. Ed. - Ian Sommerville.pdf.” p. 691, 2004.

Onion Crop Monitoring with Multispectral Imagery using Deep Neural Network

Naseer U Din[1*], Bushra Naz[2], Samer Zai[3], Bakhtawer[4], Waqar Ahmed[5]
Department of Computer System Engineering, Mehran University of Engineering & Technology, Pakistan[1, 2, 3, 4]
Department of Computer & Information Engineering, NED University of Engineering & Technology, Karachi, Pakistan[5]

***Abstract*—The world's growing population leads the government of Pakistan to increase the supply of food for the coming years in a well-organized manner. Feasible agriculture plays a vital role for sustain food production and preserves the environment from any unnecessary chemicals by the use of technology for good management. This research presents the design and development of a multi-spectral imaging system for precision agriculture tasks. This imaging system includes an RGB camera and Pi NoIR camera controlled by a raspberry pi in a drone. The images are captured by Unmanned Aerial Vehicle (UAV) and then send images to the Java application. Images are processed to sharp, resize by application. The Normalized Difference Vegetation Index (NDVI) is calculated to determine the crop health status based on real-time data. The Deep Learning (DL) technique is used to recognize the onion crop growth stage using the captured dataset. We express how to implement a progressive model for the deep neural network to recognize the onion crop growth stage. The performance accuracy of the system for batch size 16 is 96.10% and for batch size 32 is 93.80%.**

***Keywords*—UAV; deep neural networks; onion crop; NDVI; crop monitoring; VGG16**

I. Introduction

In all homes around all the year, Onion is one of the essential condiments broadly used. Recent research has recommended that in the diet, onions may play a part in avoiding heart disease and other illnesses. Onion is one of the essential crops in all continents with worldwide production of about 25 million tons. In Pakistan, there has been a progressive growth in the area and production of onion and is commercially grown with a production of 1.8 million tons on an area of 131.4 thousand hectares. The major onion growing districts are Hyderabad, Badin, Mirpurkhas, Nawabshah, Sanghar, Shikarpur, Nosharo Feroze, Ghotki and Dadu in Sindh; Kasur and Vehari, in Punjab; Chaghi, Kalat, Killa Saifullah, Nasirabad, Mastung, Khuzdar, Kharan, Turbat, and Jaffarabad in Balochistan and Swat and Dir in Khyber Pakhtun Khuwa [1, 2]. The general objective of the Agriculture Division is to grow farm yield and guarantee increased incomes for the farmers, particularly smallholders. With the predictable growth in the country's population and per capita income, marketing is usual to play an important role in guaranteeing that clients obtain food at reasonable costs and what is the food quality.

Remote sensing can be described as the obtaining of information about an object through sensors without approaching in to direct interaction with it [3]. Since the last part of the 1980s, incredible advancement has been made in remote sensing in precision agriculture [4]. Precision agriculture can be measured as a management methodology of temporal and spatial inconsistency in fields utilizing communication and information technologies [5] with the means to improve sustainability agro-environmental services. In this respect, expanding the accessibility of technological solutions are obtainable to remotely fetch and transfer environmental parameters [6, 7]. Unmanned Aerial Vehicle (UAV), usually identified as a drone, is an aircraft without a human pilot on panel. UAVs are a module of an unmanned aircraft system (UAS), which consists of a UAV, a ground-based controller, and a method of communications between both. The airlift of UAVs may function with several grades of autonomy: either control by a human with help of remote control or automatically with help of onboard computers denoted as an autopilot [8]. In previous years, a variety of UAV models running on military and civilian applications [9]. The essential fact related to UAV technology is, it has been probably 75% of UAVs like drones from 2016 to 2024 will be used in applications related to the precision of agriculture. Reliable and fast information of crop fields by capturing Images through UAV. The "Fig. 1" shows the drone which is used in this research to capture images with help of RGB and Raspberry Pi NoIR Camera and both cameras are embedded with Raspberry Pi 4.

Now a day, Farming is done in the traditional ways in Pakistan. The reality is that our farmers are mostly non-technical persons or poor persons and they have lack of appropriate information that makes it more unreliable. A big part of farming activities is depending on the guesses and these guesses mostly fail on time. Therefore, we come with an idea of crop monitoring via UAV. We consider that this idea becomes standard in the agri-business cause of its reliability and remote monitoring. With the help of this idea, we try to digitalize agricultural activities and farming, and farmers can test on necessities of the crop and perfectly calculate the growth of the crop. We believe that our idea will confidently speed up their business to reach new heights and it will be more profitable. The implementation is mostly based on awareness among farmers, which will be produced due to its various benefits. In precision agriculture applications, the use of UAV [10] has been increased in the last three years. This is primarily due to UAV ability is to deliver the farmers essential information regard crop condition (health) for good input management.

*Corresponding Author

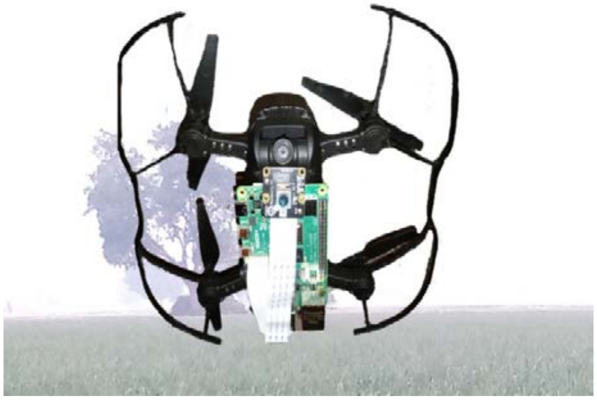


Fig. 1. UAV used in this Research.

VGG16 is a variant of Visual Geometry Group (VGG) [11] with 16 convolution layers and is very attractive because of its very unvarying architecture. It is presently the most preferred choice in public for extracting features from pictures. The weight formation of the VGGNet is widely accessible and has been utilized in several other applications and challenges as a baseline feature extractor. Instead of having a big number of hyper-parameter, they concentrated on having convolution layers of 3 X 3 filter with a stride 1 and every time utilized comparable padding & maxpool layer of 2 X 2 filter of stride 2 and this is the most inimitable thing about VGG16. It follows this procedure of convolution & maxpool layers regularly through the entire model. At last, it has 2 Fully Controlled (FC) layers tailed by a softmax for results.

Considerable damage is done to onion crops by using high or low quantities of fertilizers and water. Farmers are facing difficulties in the crop monitoring field at a bigger scale, therefore the use of UAV will allow farmers to constantly monitor their crop. This research aims to design UAV-based application to improve the health and production of onion crops. One of the major goals of this research was to achieve a state of the art classifying results using widely helpful data and models, with transfer learning to speed up training processes & balances the limited sample size so that the latest hardware can deliver sensible outcomes. This paper shows monitoring of onion crop, through the use of Near-Infrared (NIR) & RGB imagery captured by Unmanned Aerial Vehicle.

II. Related Work

A brief overview of a few significant contributions from the existing literature is provided in this section.

A. Montes, et al. [8] proposed a Low-cost multispectral imaging system for crop monitoring. The author used digital & NIR cameras to calculate orthomosaics images from NIR & RGB images using built-in software FIJI. The limitation of this system is the use of JPEG format images which makes final NDVI representation lower as compared to the TIFF format.

Gaetano Messina et al. [12] proposed Monitoring Onion Crops Using Multispectral Imagery from UAV. The author used a fixed-wing UAV, fitted out with Multispectral Camera Sequoia Parrot (R-G-RedEdge-NIR). The results of the analysis of the three datasets showed a high correlation of Green Normalized Difference Vegetation Index (GNDVI) and NDVI with Soil-Adjusted Vegetation Index (SAVI). The use of a higher resolution sensor might probably have partly solved these problems.

Kim et al. [13] proposed Machine vision-based automatic disease symptom detection of onion downy mildew aim to train the Deep Neural Network (DNN) model for detection of disease symptoms using Pan, Tilt, Zoom (PTZ) camera, wireless transceiver, a motor system, and image logging module. They achieved an accuracy of 90.7%.

Chauhan, S. et al. [14] proposed Wheat lodging assessment. This study’s main object was to analyze spectral changeability of the features derived for UAV data and their capability to differentiate between dissimilar grades of lodging severity. Using an object-based segmentation result, the nearest neighborhood classification was executed with an overall accuracy of 90%.

Raeva, Paulina Lyubenova, Jaroslav Šedina, and Adam Dlesk [15] proposed monitoring of Corn, Barley crop fields. The author used multispectral and thermal cameras to monitor these crop fields. 3 vegetation indices were determined. Furthermore, 2 thermal maps are existed to show the relation between vegetation and soil temperature. The conclusion is that the values vegetation indices obey with definite growth of crops as known from certainty. It means that thermal & multispectral imagery has great potential in agriculture.

Biao Jia et al. [16] proposed “Use of Digital Camera to Monitor the Growth and Nitrogen Status of Cotton”. The objective of this study was to develop a non-destructive technique for monitoring the growth and nitrogen status of cotton crops using an RGB camera. The use of digital cameras as a tool for near-ground remote sensing in precision agriculture is a new field of research.

C Y N Norasma et al. [17] proposed Rice crop monitoring using multi-rotor UAV and RGB digital camera at the early stage of growth. The paper's objective is to monitor rice crops by using UAV and RGB digital camera in Kelantan.

Shan et al. [18] purpose to predict Covid-19 in Computed Tomography (CT) scan by utilizing the deep learning model named VB-Net. They used 250 images for training, and 300 pictures for validation. A precision of 91.6% they achieved.

III. Materials and Method

In this section, we have discussed in detail our proposed methodology. The “Fig. 2” shows the Overall System Architecture based on the two phases. In the 1st phase, a UAV is used to capture images of the onion crop using both Raspberry Pi NoIR and RGB camera from the field and send these images to the application for further processing. The 2nd phase is to develop an application that monitors onion crop health and recognizes the onion crop growth stage. The dataset is also collected from different fields of onion crops with different growth stages. Images captured by the UAV are further processed in the application where an image sharpen filter is applied and resize the image to get more accurate results. After that application provides two features, NDVI & recognizes the onion crop growth stage.

To calculate NDVI (Normalized Difference Vegetation Indices) values of the captured images of onion crop using the application. The NDVI is computed as following expressions:

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (1)$$

Where NIR resembles to the Near-Infrared's reflectance values and RED correspond to Red radiations imitated by plants. Images containing NIR and RED radiations from vegetation are mandatory to obtain NDVI.

To recognize the onion crop growth stage, we use the deep neural networks method (VGG16) consisting of a convolution layer, maxpool layer, fully connected (hidden layer), and an output layer. The "Fig. 3" gives details of the VGG16 model.

A. Input Layer

The input layer is responsible for fetching a pre-handled onion crop picture dataset. It carries primary data into the system for additional processing by following layers of deep learning. The input layer is the very starting layer of the workflow of Deep Learning (DL) [19].

B. Convolution Layer

This layer will perform most computations that why this is the critical layer in our proposed DL model. Convolution layers are the significant building blocks utilized in deep learning. Convolution is the basic application of a filter to an input that outcomes in an activation. This layer's main purpose is to recuperate features from the dataset of images and to preserve the spatial relationship between pixels.

C. Max Pool Layer

Max pooling is a pooling process that chooses the most extreme component from the locale of the feature map covered by the filter. Accordingly, the outcome after the max-pooling layer would be a feature map having the most noticeable features of the past feature map.

D. ReLU Layer

This layer generates the non-linearity plan of the DNN features. The aim is to put the negative values of a pixel with 0 in the convolved features.

E. Fully Connected / Hidden Layer

All activation functions of the previous layer are associated with the neurons of this layer. The principal assignment of this layer is to order the gathered features in the predefined classes from the image data sets.

F. Softmax Layer

This layer is purely used to deliberate the conceivable values of the preceding layer activation function. The values can be taken in four sets of 'Initial Stage', 'Middle Stage', 'More than Middle Stage', and 'Final Stage' in the recognition case.

G. Output Layer

The output / final layer of the DL model can be categorized with outcomes of previous layers. It results in the recognition of the onion crop growth stage, for example, Initial Stage, Middle Stage, More than Middle Stage, and Final Stage.

The "Fig. 3" shows the VGG-16 architecture of the RGB images. Convolutions are used to expand the 3rd dimension but 1st and 2nd dimensions are regularly remains untouched. Pooling layers are used to reduce (dividing with 2) 1st and 2nd dimensions (Height X Width) & left 3rd dimension unaffected. VGG-16 has convolution kernels of uniform size 3 X 3 with stride as 1 just applies them several times. Using equation (2), it's not difficult to obtain comparable values.

$$ConvMat = \frac{OriMat+2*P-F}{S} \quad (2)$$

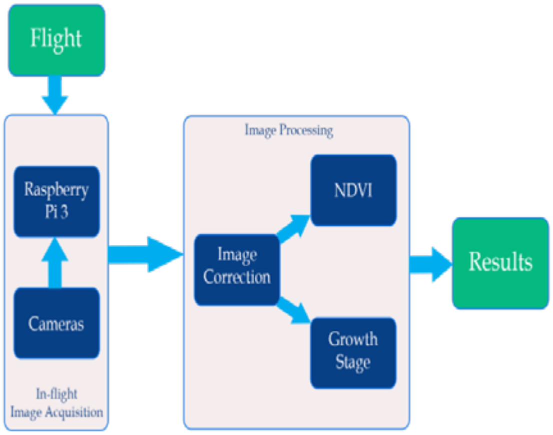


Fig. 2. Overall System Architecture.

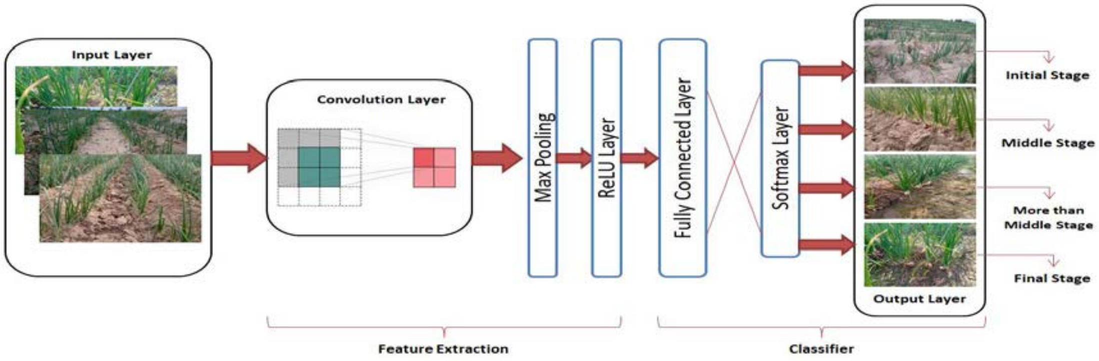


Fig. 3. VGG16 Model to Recognize Onion Crop Growth Stage.

Where ConvMat is convolutional matrix dimensions, OriMat is the original matrix, P is the padding number, S is the striding number of positions and F is the filter dimensions.

IV. Results and Discussion

We demonstrate and discuss the outcomes of our system developed and tested under preset conditions in this section. This study was conducted in an onion field located in Hala New, Sindh, Pakistan. The total field acreage was 6 acres with a well-organized plot.

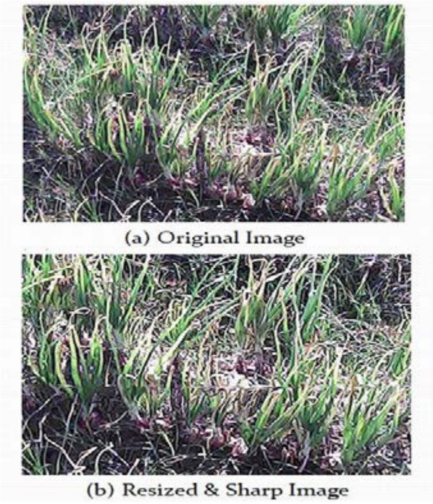
A. Flight Plan and Image Pre-processing

The UAV takes a flight of 2 minutes with an altitude of 7 feet and captures the images with Raspberry Pi NoIR camera for NDVI calculation with a 10-second interval in an acre of field and after that, it reduces altitude to 3 feet and again starts capturing the images from 3 feet altitude with RGB camera for Crop Growth Stage and send these images to the application. After sending images to the application the image sharpens filter is applied, resize the image and then process the images for NDVI calculation & Onion Stage recognition.

The “Fig. 4(a)” is the original image captured by the RGB camera and then the RGB image is sharped and resized to process and recognize the crop growth stage by application. We have the sharper image, as shown in “Fig. 4(b)”, the edges are visible to recognize the image correctly and results will be accurate. We have also sharper the NIR images to compute NDVI results accurately because the images are taken from 7 feet height with UAV. The “Fig. 5(a)” is the original image capture by the Raspberry Pi NoIR camera and “Fig. 5(b)” shows the sharper image by using the image sharpening technique through the application.

B. Compute NDVI

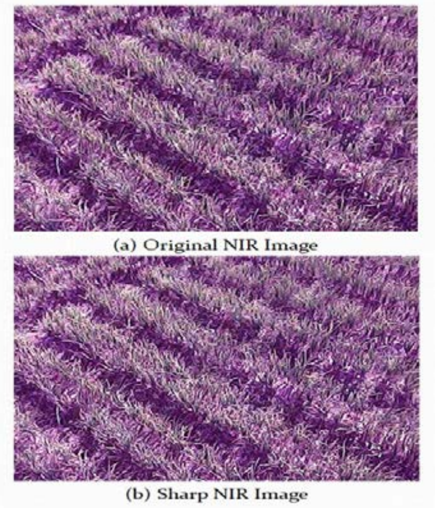
The NDVI computation outcomes range from -1.0 to 1.0. Negative values relate to regions with rocks, manmade structures, snow, water surfaces; simple soil typically comes in between 0.1 to 0.2 ranges, and plants will constantly have positive values 0.2 - 1.0. Healthy plants should be above 0.5, and damages and unhealthy plants will most likely come in between 0.2 to 0.5.



(a) Original Image

(b) Resized & Sharp Image

Fig. 4. RGB Image Processing.



(a) Original NIR Image

(b) Sharp NIR Image

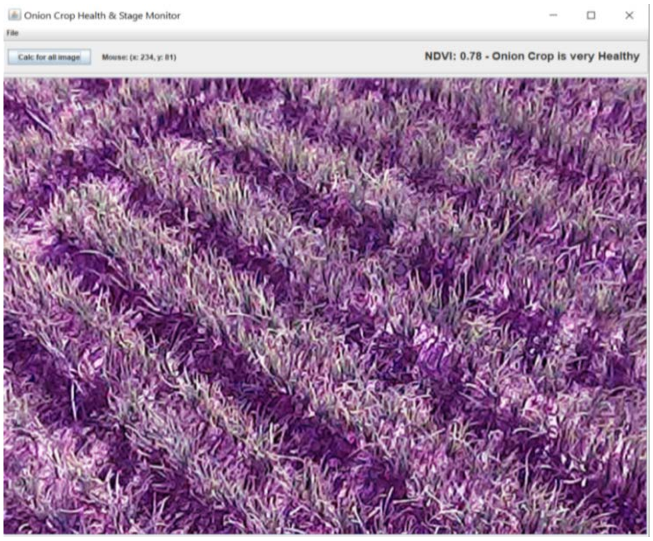
Fig. 5. NIR Image Processing.

The “Fig. 6” shows that we take the near-infrared image, process it using the NDVI calculation algorithm, and then we calculate the NDVI of the processed image. The system shows NDVI of 0.78 – Onion Crop is very healthy.

C. Onion Crop Growth Stage Recognition

In our dataset, the total number of 4000 images of the onion crop stored. 798 are classified as Initial Stage, 1022 are classified as Middle Stage, 1209 are classified as more than Middle Stage and 971 are classified as Final Stage. The images stored in the dataset have the same size of 1024x768. Brightness, Contrast, and subject finding are all very inconstant in the dataset. The dataset used in this research for recognition of the onion crop growth stage is shown in “Fig. 7”.

The RGB images are processed to recognize the onion crop growth stage. Computer Vision problems tend to be difficult to resolve with an incredible improvement of deep learning. Neural Networks (NN) are attempting to learn profoundly complex functions as much as possible, like image recognition or image object detection.



Mouse: (x: 234, y: 81)

NDVI: 0.78 - Onion Crop is very Healthy

Fig. 6. Computed NDVI.

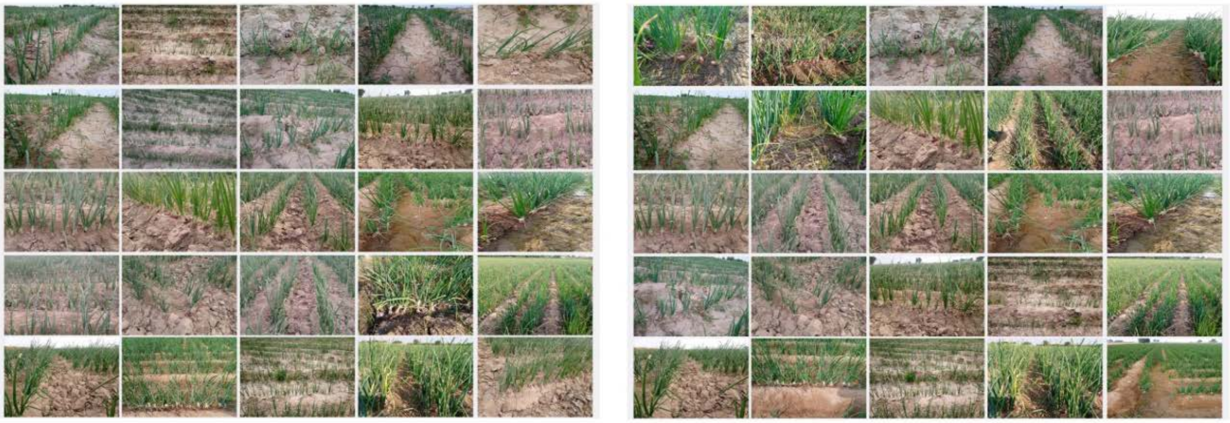


Fig. 7. Dataset of Different Growth Stages of Onion Crop.

It is one of the motives of the NN. We have a cluster of pixel values, and from that point; we might want to sort out what is in the image, so this is an unpredictable issue. We are reusing previously successful architecture by decreasing the required time for selecting alternate neural hidden layers, convolution layers, and other configuration parameters, for example, learning rate. We used java deep learning to train our system that system is capable to recognize the onion crop growth stage. For use on distributed GPUs & CPUs, Deep Learning for Java (DL4J) fetches Artificial Intelligence (AI) to business environments. By creating a java application, the methodology was established. The execution was GPU defined. All trials were done on an HP Elite Book AMD PRO A10-8700B R6, 10 Compute Cores 4 CPU + 6 GPU of 1.8 GHz, 12 GB of RAM. All research trials were carried out with 80% of the data-set for training while 15% of the dataset for testing the leftover 5% is for validation.

Now we train our system with java DNN because DNNs capability is to achieve record-breaking accuracy. The “Fig. 8” shows the feature extraction with help of the convolution layer. VGG-16 is capable to predict 1000 classes of images, while we require only 4 classes that are Initial Stage of Crop, Middle Stage of Crop, More than Middle Stage of Crop, and Final Stage of Crop. We have slightly modified this model to results in only four classes instead of 1000. The “Fig. 8” shows the feature extraction from images by inserting some convolutions and pooling layers that are utilized to train a big neural network with FC / hidden and Softmax layer to recognize the output from the image. A batch size of 16 & 3 epochs we are using. 1 epoch is a complete traversal throughout data and one loop is one forward and one backward propagation on batch size (16 images). This model trains with small strides of 16 images and each time get smarter and more intelligent. As we know every Machine Learning (ML) [20, 21] issue begins with the data. For the performance of the system, the quantity and nature of the data are very crucial and most of the time; it needs excessive deal exertions and resources.

Testing accuracy is a prediction that demonstrates the precision and accuracy of any selected deep learning model. The outcome of the testing and training model is shown below. “Fig. 9(a)” shows the accuracy with batch size 16 & “Fig. 9(b)” shows the accuracy with batch size 32. Using equation 3, the accuracy of the model computed. The model provides an accuracy of 96.10% for batch size 16 and 93.80% for batch size 32.

1) Accuracy: The important metric for the results of Deep Learning classifiers, as stated in equation (3). It’s the sum of true positive and true negative divided by the total values.

$$Accuracy = \frac{TPos+TNeg}{TPos+FPos+TNeg+FNeg} \quad (3)$$

2) Precision: To provide the relationship between the TPos estimated values and total positive estimated values as denoted in equation (4).

$$Precision = \frac{TPos}{TPos+FPos} \quad (4)$$

3) Recall: is the segment between the true positive values of prediction and the sum of probable TPos & FNeg values as shown in equation (5).

$$Recall = \frac{TPos}{TPos+FNeg} \quad (5)$$

4) F1-score: As denoted in equation (6), F1-score is a total degree of correctness that chains the precision and recall. F1-score is the twice of the fraction between the multiplication to the sum of recall and precision metrics.

$$F1 - Score = \ 2\,X \frac{Precision\ X\ Recall}{Precision+\ Recall} \quad (6)$$

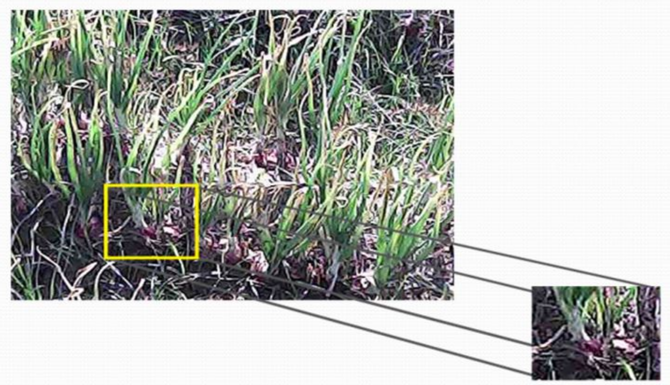
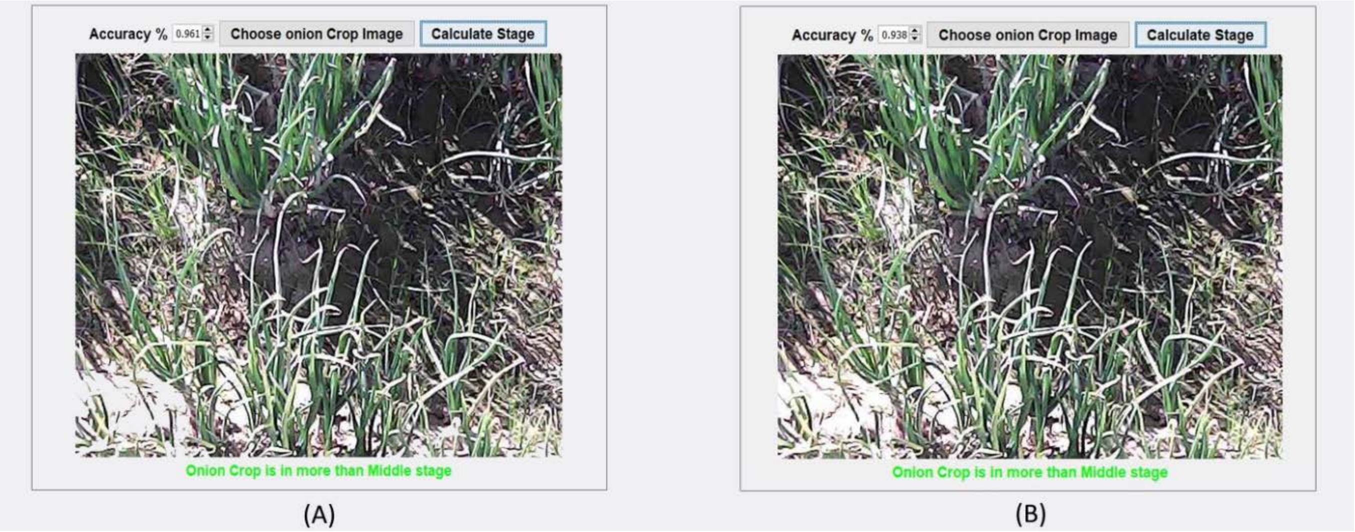


Fig. 8. Feature Extraction.



(A) Accuracy % 0.961

Onion Crop is in more than Middle stage

(B) Accuracy % 0.938

Onion Crop is in more than Middle stage

Fig. 9. Results of Onion Stage Recognition.

Table I gives the information about the performance of the model in terms of accuracy, precision, recall, and F1-scores respectively.

TABLE I. PERFORMANCE OF THE MODEL

Batch Size	**Accuracy**	**Precision**	**Recall**	**F1-Score**
16	96.10%	98.31%	98.05%	96.88%
32	93.80%	97.64%	96.93%	98.06%

VGG-16 has a total number of 138 million parameters. The significant point is that all convolutional kernels are of fixed size 3 X 3 and max-pool kernels are of size 2 X 2 with a stride of two. VGG-16 [22] has a fixed size of convolutions that’s why it takes less training time & predicts high accuracy.

V. CONCLUSION

A full imaging system has been established to get NDVI & Growth Stage information of the onion crop using an Unmanned Aerial Vehicle. A pair of cameras is used, one of both has been modified to capture NIR images and the other captures the RGB images. If images taken are not corrected, then so image sharpening techniques have been used to clarify the image and we have got accurate results. In upcoming years, it is probably expected to use a fixed-wing aircraft in place of a UAV to cover the large areas of the crop fields. Also, another camera could be used to capture red edge radiation and calculate the Normalized Difference Red Edge (NDRE). Different outcomes are achieved as shown in Table I, due to variation of some parameters. From the present research, the interpretations accomplished verified that the recommended method is capable and can be more applied to the multi-step on the estimate of a various group of disease factors of the onion crop.

ACKNOWLEDGMENT

We would like to thank the Mehran University of Engineering and Technology, Jamshoro, Pakistan, for providing us the necessary assets to conduct this research.

REFERENCES

[1] Khokhar, Khalid Mahmood. (2018). Growing onion in Pakistan. “Bulb development and seed formation in onion”.

[2] Shah, Syed Tanveer, Muhammad Sajid, Riaz Alam, Abdur Rab, Abdul Mateen, Ibadullah Jan, Asad Ali, and F. Wahid. "Comparative study of onion cultivars at Mardan, Khyber Pakhtunkhwa-Pakistan." Sarhad J. Agric 28, no. 3 (2012): 399-402.

[3] Chuvieco, Emilio. Fundamentals of satellite remote sensing. CRC press, 2019.

[4] Mulla, D.J.: Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosyst. Eng. 114(4), 358–371 (2013).

[5] Blackmore, S., Godwin, R.J., Fountas, S.: The analysis of spatial and temporal trends in yield map data over six years. Biosyst. Eng. 84(4), 455–466 (2003).

[6] 5. Merenda, M., Felini, C., Della Corte, F.G.: A monolithic multisensor microchip with complete on-chip RF front-end. Sensors (Switzerland) 18(1), 110 (2018).

[7] Merenda, M., Iero, D., Pangallo, G., Falduto, P., Adinolfi, G., Merola, A., et al.: OpenSource hardware platforms for smart converters with cloud connectivity. Electronics 8(3), 367 (2019).

[8] de Oca, A. Montes, et al. "Low-cost multispectral imaging system for crop monitoring." 2018 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2018.

[9] Mogili, UM Rao, and B. B. V. L. Deepak. "Review on application of drone systems in precision agriculture." Procedia computer science 133 (2018): 502-509.

[10] Ballesteros, Rocio, Jose Fernando Ortega, David Hernandez, and Miguel Angel Moreno. "Onion biomass monitoring using UAV-based RGB imaging." Precision agriculture 19, no. 5 (2018): 840-857.

[11] Yu, Wei, Kuiyuan Yang, Yalong Bai, Tianjun Xiao, Hongxun Yao, and Yong Rui. "Visualizing and comparing AlexNet and VGG using deconvolutional layers." In Proceedings of the 33 rd International Conference on Machine Learning. 2016.

[12] Messina, Gaetano, Vincenzo Fiozzo, Salvatore Praticò, Biagio Siciliani, Antonio Curcio, Salvatore Di Fazio, and Giuseppe Modica. "Monitoring Onion Crops Using Multispectral Imagery from Unmanned Aerial Vehicle (UAV)." In INTERNATIONAL SYMPOSIUM: New Metropolitan Perspectives, pp. 1640-1649. Springer, Cham, 2020.

[13] Kim, Wan-Soo, Dae-Hyun Lee, and Yong-Joo Kim. "Machine vision-based automatic disease symptom detection of onion downy mildew." Computers and Electronics in Agriculture 168 (2020): 105099.

[14] Chauhan, S., et al. "Wheat lodging assessment using multispectral uav data." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42.2/W13 (2019).

[15] Raeva, Paulina Lyubenova, Jaroslav Šedina, and Adam Dlesk. "Monitoring of crop fields using multispectral and thermal imagery from UAV" European Journal of Remote Sensing 52.sup1 (2019): 192-201.

[16] Jia, Biao, Haibing He, Fuyu Ma, Ming Diao, Guiying Jiang, Zhong Zheng, Jin Cui, and Hua Fan. "Use of a digital camera to monitor the growth and nitrogen status of cotton." The scientific world journal 2014 (2014).

[17] Norasma, C. Y. N., MY Abu Sari, M. A. Fadzilah, M. R. Ismail, M. H. Omar, B. Zulkarami, Y. M. M. Hassim, and Z. Tarmidi. "Rice crop monitoring using multirotor UAV and RGB digital camera at early stage of growth." In IOP Conference Series: Earth and Environmental Science, vol. 169, no. 1, p. 012095. IOP Publishing, 2018.

[18] F. Shan+ et al., "Lung infection quantification of covid-19 in ct images with deep learning," arXiv Prepr. arXiv2003.04655, 2020.

[19] Larochelle, Hugo, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. "Exploring strategies for training deep neural networks." Journal of machine learning research 10, no. 1 (2009).

[20] M. A. ZAKI, S. ZAI, M. AHSAN, and U. ZAKI, "Development of An Android App for Text Detection," J. Theor. Appl. Inf. Technol., vol. 97, no. 20, pp. 2485–2496, 2019.

[21] M. A. Zaki, S. Narejo, S. Zai, U. Zaki, Z. Altaf, and N. U Din, "Detection of nCoV-19 from Hybrid Dataset of CXR Images using Deep Convolutional Neural Network," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 699–707, 2020, doi: 10.14569/ijacsa.2020.0111281.

[22] Qassim, Hussam, Abhishek Verma, and David Feinzimer. "Compressed residual-VGG16 CNN model for big data places image recognition." In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 169-175. IEEE, 2018.

Combined Non-parametric and Parametric Classification Method Depending on Normality of PDF of Training Samples

Kohei Arai[1]
Faculty of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—Classification method with combined nonparametric and parametric classifications which depends on the normality of Probability Density Function of training samples is proposed. The proposed classification method is also based on spatial information for high spatial resolution of satellite based optical sensor images is proposed. Also, a classification method which takes into account not only spectral but also spatial features for LANDSAT-4 and 5 Thematic Mapper (TM) data is proposed. Treatment of the spatial-spectral variability existing within a region is more important for such high spatial resolution of satellite imagery data. Standard deviations in small cells, such as 2x2, 3x3 and 4x4 pixels, were used as measures to represent the spatial-spectral variabilities. This information can be used together with conventional spectral features in a unified way, for the traditional classifier such as the pixelwise Maximum Likelihood Decision Rule (MLHDR). The classification performance of new clear cuts and alpine meadows which are very close in spectral space characteristics and difficult to distinguish them by conventional methods are focused. Through experiments, it is found that there is a substantial improvement in overall classification accuracy for TM forestry data. The Probability of Correct Classification (PCC) for the new clear cuts and the alpine meadows classes rose by 7% to 97% correct. The confusion between alpine meadows and new clear cuts was reduced from 9% to 3%.**

Keywords—Spectral information; spatial information; maximum likelihood decision rule; satellite image; image classification; classification performance; instantaneous field of view

I. Introduction

Maximum Likelihood Decision Rule-based classification method: MLHDR is widely used for satellite imagery data classification. There are some assumptions of the MLHDR, such as (1) there is no correlation among the pixels, (2) Probability Density Function: PDF of the training samples is normal distribution, and so on. These assumptions, however, are not always true. There are pixel-to-pixel correlations and spatial information which can be extracted from the satellite imagery data and can be used for image classification are followed by non-normal distribution. The proposed image classification method uses not only spectral but also spatial information and allows classification with non-normal distribution of PDF of the training samples.

Spatial resolution of spaceborne based optical sensors is improved remarkably. Classification performance is getting down because the variances of the specified class categories are getting large in accordance with spatial resolution which results in increasing overlapped areas among the class categories in the feature space for classification. It might be possible to improve classification performance by including spatial information into classification (Spatial information is getting large in accordance with spatial resolution).

Due to the increased spatial resolution of TM (30m compared with 80m for Multi Spectral Scanner: MSS), the number of ground cover spectral classes which are included in the Instantaneous Field of View (IFOV), decreases comparatively. This implies that spatial spectral variability for TM data increases in comparison to MSS. With the increase of the number of spectral bands and quantization bits of Landsat Thematic Mapper: TM (30 m of Instantaneous Field of View: IFOV (spatial resolution) of optical sensor), the discrimination ability or classification accuracy of the surface observation object improved [1], [2], [3]. However, improvement in spatial resolution does not always have a favorable effect on classification accuracy. The spatial frequency components of the object to be observed are widely distributed. The spatial frequency components are integrated with an area corresponding to the spatial resolution and sampled to obtain image data. At this time, the variance of the data increases as the spatial resolution increases.

Therefore, with the improvement of spatial resolution, the class variance in the feature space increases, and the classification accuracy tends to deteriorate [4]. Each institution has proposed a method that can cope with this [5]-[14], but here the author proposes a method that also uses the spatial information that has been increased together with the spectral information as the spatial resolution improves.

Spatial information can be roughly divided into (1) Context and (2) Texture, both of which use feature values defined within a relatively large window area. This is because the size of the feature quantity depends on the size of the window area, and a large window is required for the spatial resolution of the MSS. However, in the TM image, the spatial information is increased, so that even a small window is effective, and the physical meaning of the variation or difference in the spectral information between the surrounding pixels and the immediate pixel is clear. Furthermore, the definition of spatial information that has already been proposed and its application to classification methods are often complex, and require a lot of processing time.

Based on the above, this paper defines the standard deviation of pixel values in a small window (cell consisting of several pixels) as spatial information, adds it to the dimension of the spectral information in the feature space, and uses the maximum likelihood classification. The author proposes a method to do it. Although the standard deviation within this cell was proposed by Strahler et al [15], the basis for using the standard deviation was not clear [16], and the consideration of the optimal cell size, the probability density function, its normality, etc., was lacking [17]. In this paper, the author clarifies these and mention the effect and practicality of this method based on statistical features [18].

In the following section, related research works and research background including motivation of the research are described. Then, the proposed context classification method is described followed by experimental method together with experimental results. After that concluding remarks and some discussions are described.

II. Related Research Works

Classification by re-estimating statistical parameters based on auto-regressive model is proposed for purification of training samples [19]. Meanwhile, multi-temporal texture analysis in TM classification is proposed for high spatial resolution of optical sensor images [20]. On the other hand, Maximum Likelihood (MLH) TM classification taking into account pixel-to-pixel correlation is proposed [21].

Supervised TM classification with a purification of training samples is proposed [22] together with TM classification using local spectral variability is proposed [23]. A classification method with spatial spectral variability is also proposed [24] together with TM classification using local spectral variability [25].

Application of inversion theory for image analysis and classification is proposed [26]. Meanwhile, polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the polarization signature is proposed [27]. On the other hand, a hybrid supervised classification method for multi-dimensional images using color and textural features is proposed [28].

Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA is proposed [29]. Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space is conducted and well reported [30].

Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods is conducted [31]. On the other hand, multi spectral image classification method with selection of independent spectral features through correlation analysis is proposed [32]. Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor is proposed [33].

Image classification considering probability density function based on Simplified beta distribution is proposed and evaluated its performance [34]. Meanwhile, maximum likelihood classification based on classified result of boundary mixed pixels for high spatial resolution of satellite images is proposed [35]. Also, context classification based on mixing ratio estimation by means of inversion theory is proposed [36]. Optimum spatial resolution of satellite-based optical sensors for Maximizing Classification: MLH performance is discussed [37].

III. Research Background

A. Problems in Classifying High Spatial Decomposition Images

When the class defined for the Multi Spectral Scanner: MSS (80 m of IFOV of optical sensor onboard the same satellite of Landsat) image is directly applied to the TM image, for example, if the same training area is set for both data obtained by observing the same object at the same time, the TM image has a larger class variance. This is due to the high radial and spatial resolution of TM. Both data can be obtained by sampling and quantizing the observation object that is continuous both spatially and radiometrically. However, when the sampling frequency and the number of quantization bits increase, the variance of the observation value increases.

However, the class variance also increases, leading to a decrease in the degree of separation between classes, and as a result, the classification accuracy decreases. On the other hand, the spatial information included in the image increases as the sampling frequency and the number of quantization bits increase. Therefore, an improvement in accuracy can be expected by considering spatial information in the classification of high-resolution images.

B. Classification Method in Concern

Image classification methods can be divided into the following four categories.

1) Supervised / Unsupervised.

2) Use / do not use spatial information.

3) Assuming probability density function (parametric) / not assuming (non-parametric).

4) Image unit (pixel-wise) / cell unit (cell-wise).

In this paper, the author examined the supervised maximum likelihood method per plane element which was effective for Landsat MSS data for the following reasons.

1) Unsupervised classification is difficult to interpret the physical meaning of the class after classification, and is suitable as a preliminary classification. However, correspondence with classes according to purpose of use such as wheat fields, paddy fields, soybean fields in agricultural areas is difficult.

2) Since the radiometric resolution of TM is improved compared to MSS, the spectrally homogeneous spatial region is narrowed, and cell-based classification is not effective.

3) Nonparametric methods generally require more computation time and storage capacity than parametric methods.

On the other hand, when applying the maximum likelihood method using only the spectrum information per plane element to the TM image, the following points need to be considered.

1) Definition of spatial information and its application to classification methods.

2) Multidimensional normality of probability density function of observation vector.

3) Furthermore, surface element correlation in spectrally homogeneous regions.

In this paper, especially for (1), the author proposes a method that defines the standard deviation in the cell as spatial information and uses it as one dimension of the feature space.

IV. Proposed Method

A. Process Flow of the Proposed Classification Method

The proposed method is based on MLHDR which is based on multi variate normal distribution of probability density function [16] which is expressed with equation (1) and (2).

$$D_{ij} = \int_{-\infty}^{\infty} log\left\{\frac{p(x|w_j)}{p(x|w_i)}\right\}\{p(x|w_i) - p(x|w_j)\}dx \qquad (1)$$

$$p(x|w_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^t|\Sigma_i|^{-1}(x-\mu_i)\right\} \qquad (2)$$

where *Dij* denote separability between class *i* and *j* and *p* is the number of dimensions, *Σi* is the covariance matrix of the class, x is the observation vector, *μi* is the mean vector of class *i*, and *t*, -1 are the transposed matrix and the inverse matrix, respectively.

The degree of separation between classes is tested using only spectral information, and the following method is applied only to classes that do not show a satisfactory value (for example, divergence of 500 or more). That is, by selecting the spectral band exhibiting the largest variance and examining the spatial frequency components of the class, the optimal cell size is determined, and the standard deviation within the cell is calculated by moving the cell by one pixel. The leverage result is added to the spectrum space as a new dimension (this is called an integrated feature space). After that, the multidimensional normality in the integrated feature space is tested again to confirm the degree of separation, and then the result of classification is applied by applying the maximum likelihood classification to this result, and the final classification image is obtained.

Fig. 1 shows process flow of the proposed classification method.

B. Standard Deviation in the Cell in Concern as a Texture Information

The variance of pixel values in a cell is essentially equal to "contrast" *c*, one of the well-known texture measures, with a constant multiple.

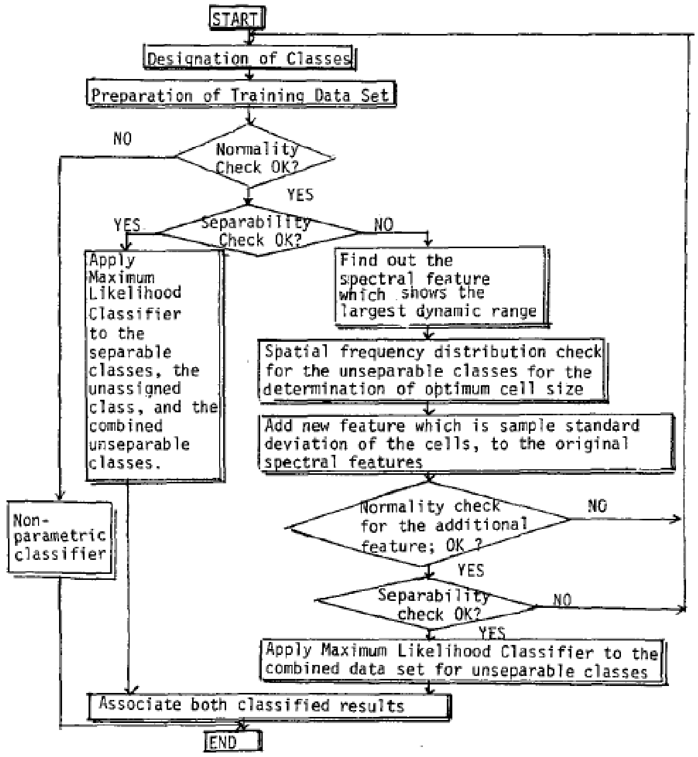

START
Designation of Classes
Preparation of Training Data Set
NO
Normality Check OK?
YES
YES
Separability Check OK?
NO
Apply Maximum Likelihood Classifier to the separable classes, the unassigned class, and the combined unseparable classes.
Find out the spectral feature which shows the largest dynamic range
Spatial frequency distribution check for the unseparable classes for the determination of optimum cell size
Add new feature which is sample standard deviation of the cells, to the original spectral features
Non-parametric classifier
Normality check for the additional feature; OK ?
NO
YES
Separability check OK?
NO
YES
Apply Maximum Likelihood Classifier to the combined data set for unseparable classes
Associate both classified results
END

Fig. 1. Flowchart for the Proposed Classifier.

$$\sigma_c^2 = \sum_i \sum_j (x_{ij} - \bar{x})^2/N \qquad (3)$$

$$c = \sum_{i,j} \sum_{k,l} (x_{ij} - x_{kl})^2/N_p \qquad (4)$$

where *i* and *j* are the pixels and line numbers, *N* is the number of pixels in the cell, *Nρ* is the number of pixel pairs in the cell, x_{ij} is the value at pixel position *ij*, and} {is the average value in the cell. is there.

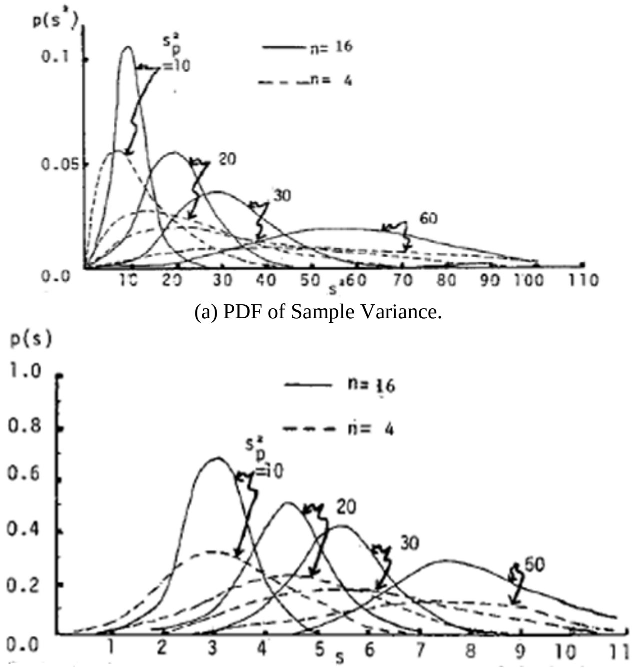
According to Percival's theorem, the following equation holds between the power spectrum and $\sigma_c{}^2$.

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |F(s)|^2 ds = \sigma_c^2 + \bar{\bar{x}}^2 \qquad (5)$$

where, *f (x)* is a value at a pixel position *x*, and *F(s)* is a function obtained by performing Fourier transform in a frequency domain. Eq. (5) shows that the variance of the pixel values in the cell is closely related to the power spectrum and the contrast as texture information.

C. Probability Density Function of Standard Deviation and Variance in the Cell in Concern

Assuming now that the observed value population follows the multidimensional normal distribution, the probability density function of the sample variance and sample standard deviation in the cell is obtained, and each normality is tested. Fig. 2 shows an example of calculating each probability density function (see Reference [17]) for a population variance S_p^2 of 10 to 60, using the cell size *n* as a parameter. Here, assuming a square cell, *n* = 4 (2 × 2) and 16 (4 × 4) were selected as parameters.



(a) PDF of Sample Variance.

(b) PDF of Sample Standard Deviation.

Fig. 2. Comparison between Probability Density Functions of Sample Standard Deviation s and Variance s^2 where n: Sample Size (Cell Size), $s_p{}^2$: the Population Variance.

The normality improves as this approach infinity. From the viewpoint of normality, it is known that the standard deviation (square root transformation) is superior to the variance, and that the cubic root transformation is optimal. (On the other hand, since the tendency tends to be the opposite from the viewpoint of spatial information, if normality is satisfied, it is better to use the flat-root transformation and the variance itself than the cubic root transformation. Standard deviation was used, and the normality of a single variable was judged to be satisfied if the χ^2 value was less than 5 in the χ^2 test.

V. Experiment

A. Data used

The data used for analysis is TM data acquired by Landsat 5 on August 15, 1984, and an image processing system maintained at the Canada Center for Remote Sensing: CCRS; radiometric and geometric corrected, Geocoded data by MOSAICS system. Geocoded Data is proposed for the first time in Canada and has pixel intervals determined so that it can be easily combined with not only heterogeneous sensor data (aircraft, satellites, etc.) but also topographic maps and administrative information. This is the corrected map reference data.

The analysis area is a forested area at Spruce-Balsam in Cranbrook, British Columbia, Canada. These include logging areas, alpine meadows, rocks, rivers, small lakes, etc., in addition to forests. Fig. 3 shows a forest polygon overlay (dark blue) on the Landsat-5 TM data. Spruce-Balsam polygons have been extracted and used as a mask to cluster the TM data. The inhomogeneity of the polygons is demonstrated by the multiple classes (red, yellow, white, blue) found inside the polygons (arrow).

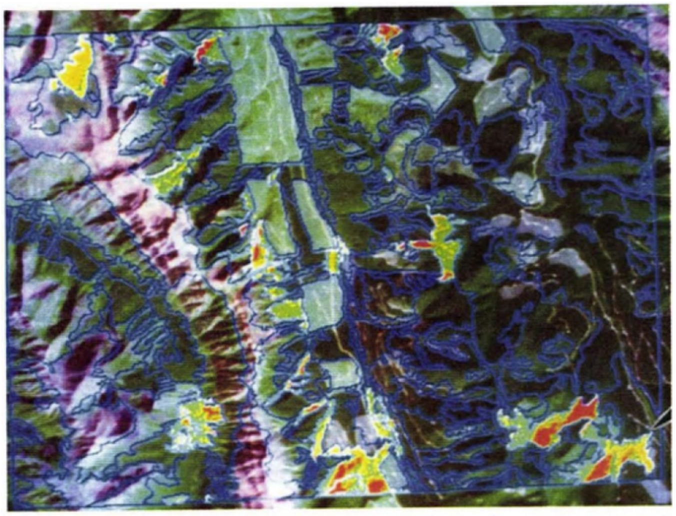


Fig. 3. Forest Polygon Overlay (Dark Blue) on the Landsat-5 TM Data. Spruce-Balsam Polygons have been Extracted and used as a Mask to Cluster the TM Data.

Logging areas are further divided into two types: those that are less than 5 years after logging and those that are 5 to 40 years later. The former is called the new logging area, and the latter is called the old logging area. The former consists of topsoil, grassland, stumps, clumps of cut down trees, roads, young trees with low height, etc. The latter consists of young trees with relatively high heights, grassland, roads, etc. Therefore, four classes were set up here: (1) new logging area, (2) old logging area, (3) alpine grassland and (4) forest.

B. Separability between Classes

From the components and their ratios, the spectral characteristics of the new cutting area and the alpine meadows are expected to be very close. Fig. 4(a) shows the class distribution in the feature space of TM bands 4 and 5 (TM-9, 5), confirming the above prediction. Here, the size of the training field of each class is 956, 270, 435, and 1377 pixels for the new logging area, old logging area, alpine meadows, and forests, respectively.

The contour in the figure corresponds to twice the standard deviation of the two-dimensional normal distribution. From this figure, it can be seen that the freshly cut area and the alpine meadows are short in distance and the dispersion is relatively large, so the degree of separation is low.

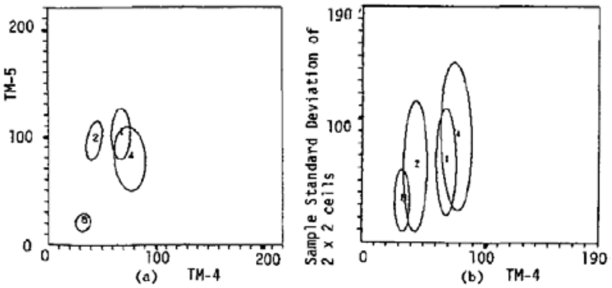


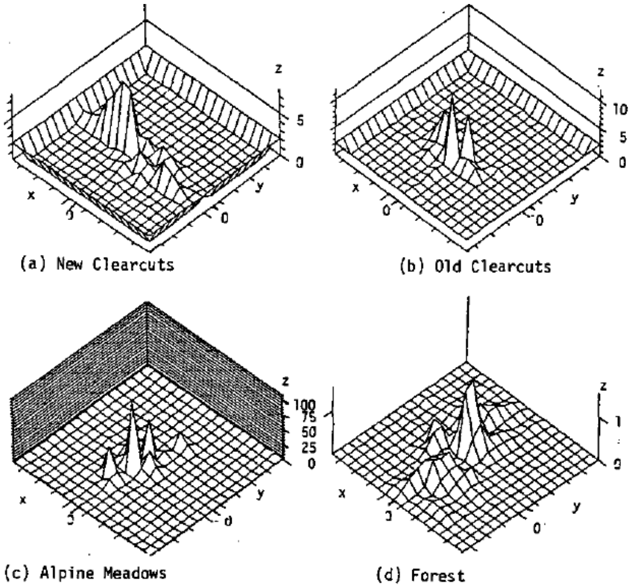
Fig. 4. An Example of the Scatterogram showing the difficulty of Distinction between New Clear Cuts and Alpine Meadows Classes. (a) is for the Original Data Set, (b) is the Scatterogram between TM-4 and Sample Standard Deviation of 2 x 2 Cells for TM-4. The Classes are: 1) New Clear cuts, 2) Old Clear Cuts, 3) Alpine Meadows and 4) Forest.

Fig. 4(b) shows the distribution of each class in the space between the TM-4 and the standard deviation of the pixel values in the 2x2 cell of the TM-4 image. It suggests that the degree could be improved.

Fig. 5 shows a typical 16 × 16 pixels window in the training field of each class, and the power spectrum is displayed by removing the DC component in the TM-4 window from each pixel value is there.

The display method makes it easy to see the difference in spectrum between each class by subtracting the D.C. component of the dominant power spectrum. Table I shows the average and standard deviation of TM-4 and power spectrum.

From the above, it can be seen that it is difficult to separate classes between the newly harvested area and the alpine meadows using only spectral information, but it is possible to improve the degree of separation by adding spatial information to this information. Here, the standard deviation is obtained by normalizing the minimum / maximum value in an image composed of 512 × 512 pixels to data of 0 to 255 (8 bits).



(a) New Clearcuts

(b) Old Clearcuts

(c) Alpine Meadows

(d) Forest

Fig. 5. Power Spectrum of the Typical Sample Area for Each Class, Subtracted Mean Value from Original Image (TM-4 of B.C. forest). Spatial Variability for the Class of Alpine Meadows is Larger than those for the Other Classes. The x and y Axes Correspond to the Wave Numbers for the Horizontal and Vertical Axes in the Frequency Domain and the Z Axis Corresponds to the Power Spectrum of the Image with the Mean Value Subtracted from the Original Image.

TABLE I. MEAN AND STANDARD DEVIATION (S.D.) OF THE EXTRACTED SUB-IMAGE FOR EACH CLASS, MEAN AND STANDARD DEVIATION OF THE POWER SPECTRUM (P.S.) OF THE IMAGE WITH MEAN VALUE SUBTRACTED FOR TM-4 IN THE 16X16 WINDOW

Class	TM-4		PS	
	Mean	SD	Mean	SD
New Clear-cut	66.15	22.2	0.26	2.37
Old Clear-cut	39.8	22.42	0.26	2.95
Alpine Meadow	60.52	58.7	1.79	25.04
Forest	27.48	12.06	0.08	0.48

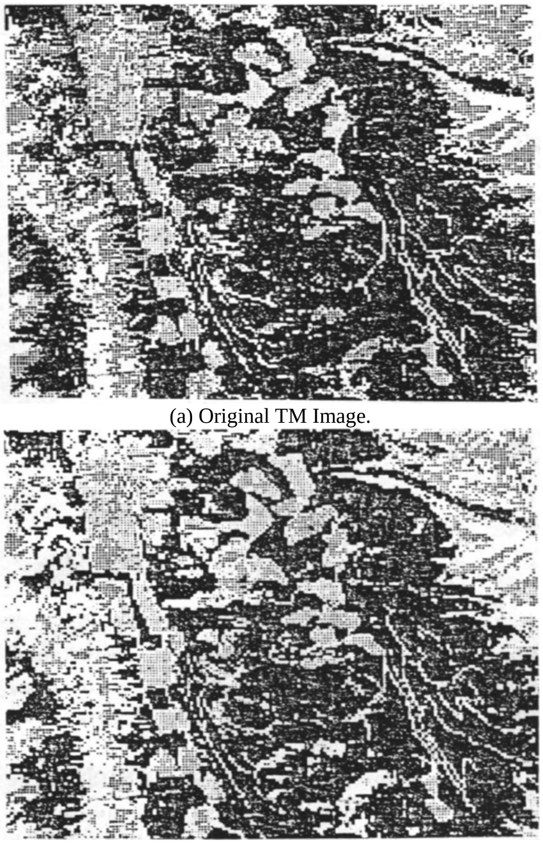
C. Improvement of Percent Correct Classification: PCC

Fig. 6 shows the case where only the spectral information of TM-1 to TM-5 and 7 is used, and the case where the information indicating the standardized standard deviation within 2 × 2 cells for TM-4 is added 3 shows a maximum likelihood classification image.

Table II shows the discrimination efficiency matrices for both data sets. The average discrimination efficiency can be improved by about 3.7% from 94.81% to 9828% by adding the information of the standard deviation, and the misclassification rate between the new cutting area and alpine meadows can be reduced from 9% to 3% understood.

When spectral information was limited to TM-4 and 5, the discrimination efficiency * for newly felled areas and alpine meadows was 65.3% and 87.1%. When the standard deviation was added, it became 83.2% and 91.4%, and it was found that the discrimination efficiency of 27.4% and 4.9% was improved.

The discrimination efficiency is an element of the discrimination efficiency matrix (confusion matrix), and is a percentage ratio of the pixels in the set training region and the region set for the discrimination efficiency evaluation in each class.



(a) Original TM Image.

(b) Combined Image of Original TM and of Standard Deviation of 4x4 Pixels Windows of TM.

Fig. 6. Examples of the Classified Images with Original and Combined Data Sets. Clear Cuts Areas which are Located at Upper Left around the Center of this Image Tend to Classify to the Alpine Meadows Class for the Original Data Set, but the Classified Image with the Combined Data Set Shows the Improvement on the Distinction between Alpine Meadows and New Clear Cuts.

TABLE II. CONFUSION MATRICES FOR THE ORIGINAL AND THE COMBINED DATA SETS

(A) CONFUSION MATRIX OF ORIGINAL TM IMAGE (PCC=94.81%)

Class	New Clear-cut	Old Clear-cut	Alpine Meadow	Forest
New Clear-cut	92.98	2.22	8.82	0
Old Clear-cut	0	95.56	0	0
Alpine Meadow	4.97	2.22	90.69	0
Forest	0	0	0	100
Unassigned	2.05	0	0.49	0

(B) COMBINED DATA OF ORIGINAL TM AND STANDARD DEVIATION OF 4X4 PIXELS OF TM IMAGE (PCC=98.28%)

Class	New Clear-cut	Old Clear-cut	Alpine Meadow	Forest
New Clear-cut	96.06	0	2.94	0
Old Clear-cut	0	100	0	0
Alpine Meadow	3.42	0	97.06	0
Forest	0	0	0	100
Unassigned	0.51	0	0	0

D. Effect of Cell Size

The standard deviation within a cell increases with cell size. Therefore, if the region is limited to a region having the same spectrum, the discrimination efficiency of each class monotonically increases as shown by the solid line in Fig. 7. However, if a cell exists at the boundary between classes, the spatial information is not class specific. Therefore, the discrimination efficiency of the area within the range of the cell size from the boundary is affected by this.

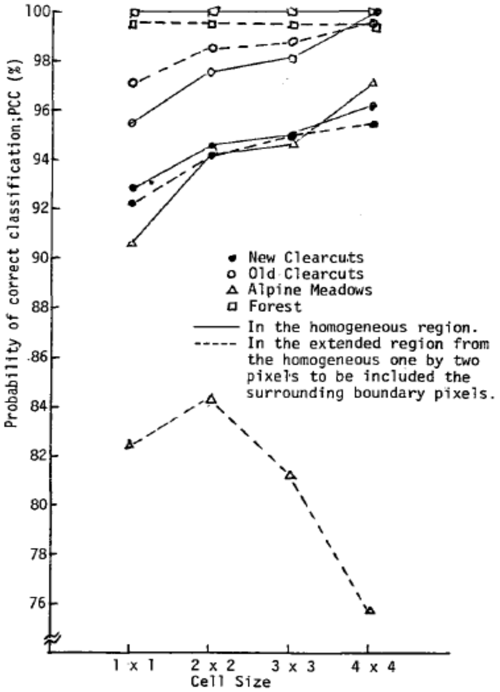


Fig. 7. Cell Size Effect on the Probability of Correct Classification.

This effect depends on the difference between the spatial information of adjacent class questions, the boundary shape, and the like, and can be reduced by reducing the cell size. If the area is set to include the boundary and the relationship between the cell size and the discrimination efficiency is determined, the result is as shown by the broken line in Fig. 6. By size, discrimination efficiency shows a peak. For this reason, a 2 × 2 cell size is set as the optimum value here, but the optimum cell size is generally determined based on the spatial frequency components of the class to be classified.

VI. CONCLUSION

The proposed method using the spectral information and the standard deviation of the pixel values in a small cell has the following features:

1) The spectral characteristics are very similar, and are effective for classes with different spatial information, respectively.

2) Unlike the other texture measures, the standard deviation of the pixel value in the cell is relatively close to a normal distribution, so that it can be simply used together with spectral information.

3) Therefore, compared with the classification method using other spatial information, it does not require complicated processing and is practical.

Also, the proposed image classification method allows classify images whose PDF is not followed by normal distribution because it supports nonparametric classification.

VII. FUTURE RESEARCH WORKS

The proposed method is adopted in the real earth observation satellite imagery data, and it is a future subject to realize a more usable classification method.

ACKNOWLEDGMENT

The author conducted this research at the Canada Center for Remote Sensing from 1985 to 1987 as a Canadian Government Scholarship Student (Postdoctoral Fellow) of the NSRC (Natural Science Research Council of Canada). The author would like to thank Dr. D. G. Goodenough, Dr. J. Iisaka, Dr. B. Guinden and their staff who participated in the discussion for distorting valuable suggestions. The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

[1] Whelms, D.L., et al., A Statistical evaluation of the advantages of Landsat Thematic Mapper Data in Comparison to multispectral scanner data, IEEE Trans. on Geosci. Remote Sensing, 22, 294, 1984.

[2] Irons, J.R., et al., The Effects of Spatial Resolution on the classification of Thematic Mapper Data, Int. J. Remote Sensing, 6, 8 1385 – 1403, 1985.

[3] Rosenfield, G.H., Analysis of Variance of Thematic Mapping Experiment Data, Photo. Eng. and Remote Sensing, 47, 12, 1685 – 1692, 1981.

[4] Kohei Arai, A Consideration on Optimum Spatial Resolution for the Multispectral Classification, 36th IAF, IAF 85-104, 1985.

[5] Cushnie, J.L. et al, Effect of Spatial Filtering on Scene Noise and Boundary Detail in Thematic Mapper Imagery, Photo. Eng. And Remote Sensing, 51, 9, 1483 – 1493, 1985.

[6] Toll, D.L., Landsat-4 Thematic Mapper Scene Characteristics of a Suburban and Rural Area, ibid, 51, 9, 1471 – 1482, 1985.

[7] Latty, R.S. et al, Performance Comparisons Between Information Extraction Techniques Using Variable Spatial Resolution Data, Photo.Eng. and Remote Sensing, 51, 9, 1459 – 1470, 1985.

[8] .Kohei Arai, et al, Multitemporal Analysis of Texture Measures for TM Classification, Proceeolings of IGARSS'87, Vol. 1, 117-118, 1987..

[9] Haralick, R.M., et al, Contextual Classifier, Digest Int. Geosci. and Remote Sensing, 1, 247 – 254, 1985.

[10] .Richards, J.A. et al, On the Accuracy of Pixel Relaxation Labeling, Tech. Report of LARS, SR-PO-00455, 31, 1980..

[11] Kast, J.L. et al, ECHO User's Guide, LARS Publication, 083077, 72 1977.

[12] Strahler, A.H., The Use of Prior Probabilities in Maximum Likelihood Classification of Remotly Sensed Data, Remote Sensing Environment, 10, 135 -163, 1980..

[13] Goodenough, D.G. et al, On the Classification for TM Data, 10th Canadian Symp. on Remote Sensing, Session C3, 164, 1986..

[14] Haralick, R.M., Statiatical and Structural Approaches to Texture, Proc. of the LEEE, 67, 5, 786 – 804, 1979.

[15] Strahler, A.H., Forest Classification and Inventory System Using Landsat, Digital Terrain, and Ground Sample Data, Proc. of the 13th Int. Symp. on Remote Sensing of Environment, vol-Ill, 1541 – 1557, 1979.

[16] Lesaffre, E., Normality tests and transformations, Pattern Recognition Letters Vol.1, pp.187 – 199, 1983.

[17] Johnson, N.L. and S.Kotz, Distributions in Statistics : Continuous Univariate Distributions-1, John Wiley and Sons, 1970.

[18] Kei Takeuchi, Probability distribution and statistical analysis, Japan Standards Association, 1979.

[19] Kohei Arai, Classification by Re-Estimating Statistical Parameters Based on Auto-Regressive Model, Canadian Journal of Remote Sensing, Vol.16, No.3, pp.42-47, Jul.1990.

[20] Kohei Arai, Multi-Temporal Texture Analysis in TM Classification, Canadian Journal of Remote Sensing, Vol.17, No.3, pp.263-270, Jul.1991.

[21] Kohei Arai, Maximum Likelihood TM Classification Taking into account Pixel-to-Pixel Correlation, Journal of International GEOCARTO, Vol.7, pp.33-39, Jun.1992.

[22] Kohei Arai, A Supervised TM Classification with a Purification of Training Samples, International Journal of Remote Sensing, Vol.13, No.11, pp.2039-2049, Aug.1992.

[23] Kohei Arai, TM Classification Using Local Spectral Variability, Journal of International GEOCARTO, Vol.7, No.4, pp.1-9, Oct.1992.

[24] Kohei Arai, A Classification Method with Spatial Spectral Variability, International Journal of Remote Sensing, Vol.13, No.12, pp.699-709, Oct.1992.

[25] Kohei Arai, TM Classification Using Local Spectral Variability, International Journal of Remote Sensing, Vol.14, No.4, pp.699-709, 1993.

[26] Kohei Arai, Application of Inversion Theory for Image Analysis and Classification, Advances in Space Research, Vol.21, 3, 429-432, 1998.

[27] Kohei Arai and J.Wang, Polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the polarization signature, Advances in Space Research, 39, 1, 149-154, 2007.

[28] Hiroshi Okumura, Makoto Yamaura and Kohei Arai, A hybrid supervised classification method for multi-dimensional images using color and textural features, Journal of the Japanese Society of Image Electronics Engineering, 38, 6, 872-882, 2009.

[29] Kohei Arai, Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA, International Journal of Advanced Computer Science and Applications, 2, 9, 37-42, 2011

[30] Kohei Arai Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space, International Journal of Research and Reviews on Computer Science, 2, 4, 1005-1009, 2011.

[31] Kohei Arai, Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods, International Journal of Advanced Computer Science and Applications, 4, 1, 95-102, 2013.

[32] Kohei Arai, Multi spectral image classification method with selection of independent spectral features through correlation analysis, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 21-27, 2013.

[33] Kohei Arai, Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor, International Journal of Advanced Research in Artificial Intelligence, 2, 10, 19-25, 2013.

[34] Kohei Arai, Image classification considering probability density function based on Simplified beta distribution, International Journal of Advanced Computer Science and Applications IJACSA, 11, 4, 481-486, 2020.

[35] Kohei Arai, Maximum Likelihood Classification based on Classified Result of Boundary Mixed Pixels for High Spatial Resolution of Satellite Images, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9, 24-30, 2020.

[36] Kohei Arai, Context Classification based on Mixing Ratio Estimation by Means of Inversion Theory, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 44-50, 2020.

[37] Kohei Arai, Optimum Spatial Resolution of Satellite-based Optical Sensors for Maximizing Classification Performance, Journal of Advanced Computer Science and Applications, Vol. 12, No. 2, 363-369, 2021.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

Online Training and Serious Games in Clinical Training in Nursing and Midwife Education

Galya Georgieva-Tsaneva[1]
Institute of Robotics
Bulgarian Academy of Sciences
Sofia, Bulgaria

Ivanichka Serbezova[2]
Faculty of Public Health and Healthcare
Angel Kanchev University
Ruse, Bulgaria

***Abstract*—The article examines the application, methods, and trends in online training of health care and medical professionals in Bulgaria. Attention is paid to modern methods for the effective application of online training and the extent to which online training can replace traditional training. The article presents the results of a survey for online training conducted in April 2021 at universities in Bulgaria Health Care professional field, specialty Nurse and Midwife. The results of the survey can serve to improve online education in Bulgaria by including in it educational resources recommended by respondents. The creation of new web-based educational resources (video materials, serious games, virtual simulations, video presentations, webinars, etc.) can complement the traditional methods of training in the students in the Health Care professional field, specialties Nurse and Midwives in Bulgaria.**

***Keywords*—Health care; medical education; serious educational games; nurse; midwife; online training**

I. Introduction

In today's technological age, new diverse methods of online learning are increasingly being integrated into higher health care and medical education. the questions about which are the most effective methods for online training, which technologies would be most suitable for students of the specialties Midwife and Nurse (Health Care professional field) are to be studied. Online learning of students today is influenced by the development and application of modern communication information technologies in all spheres of life, by advances in medical sciences in the detection and treatment of diseases, and by the improvement of the quality of health services.

An important issue in online learning is the implementation of learning strategies based on classical real learning, modern online learning, and practical training, which in the field of training of medical professionals can not be completely replaced by online innovation. The assessment of online learning can be made based on the degree of satisfaction of the learner (from the learning objectives and content, format, and effectiveness of learning resources), the learning outcomes (assessment of knowledge and skills of learners), improving the effectiveness of learning (improvement of practical skills) and the real health results when working with patients (changes in the behavior of patients and improvement of their health indicators as a result of the practical skills of the learners).

Online learning [1, 2] is learning mediated by the World Wide Web, intranet, and multimedia computer applications. Central to online learning is the online learning resources used by students: illustrations, videos, medical images, virtual patients, and others. Medical and health care universities are looking for effective approaches to optimize online learning, which will become more widespread.

The purpose of this report is to give an idea of the online training of health care students in the academic year 2019/2020 and 2020/2021 (until April 2021) in higher universities in Bulgaria, professional field of Health Care, specialty Midwives, and Nurse. The results shown were obtained by conducting an online survey of students in these specialties at the following medical universities: Varna, Plovdiv, Sofia, Pleven, Stara Zagora in Bulgaria; University “Angel Kanchev” in Ruse, Southwestern University "Neofit Rilski" - Blagoevgrad, Burgas University “Asen Zlatarov” (these three universities are not medical, but are accredited to teach in Health Care professional field).

The results of the study provide an opportunity for practical planning of future online training of health care students at universities. The differences between the theoretical views on online learning and its practical results can be correctly assessed by surveying the opinions of the students involved in online learning.

The rest of the paper is summarized as follows: Literature survey is shown in Section II. The use of serious games in medical and health care education is presented in Section III. The survey results are shown in Section IV. The conclusion is presented in Section V.

II. Trends in the Entry of Online Training in Medical Education

Online medical and health care education uses the advances of information and communication technologies to transmit knowledge at a distance. In the field of medicine, teaching and acquiring practical professional skills and habits cannot be completely replaced by online learning. It can be used as complementary training, such as a lecture or demonstration complemented by an online lesson. The use of serious games in the form of quizzes for testing students' progress, the use of serious simulation games in order to consolidate theoretical knowledge and improve students' practical skills through training of real manipulations in an artificial environment is applicable. Digitized learning resources, digital libraries, and databases with learning resources are available on the web.

In Bulgaria from March 2020 to April 2021 (when the survey was conducted), in connection with the emerging epidemiological situation, much has been done to enter and improve the quality of online education in health care education. Online learning is integrated into the curricula and a blended learning strategy is created. Many digital repositories of e-learning materials are being set up around the world, some of which are peer-reviewed, where instructors or developers can submit materials for widespread use.

Study [3] of the Liaison Committee on Medical Education in the U.S. shows an increase in the use of educational software products in medical education, as well as an increase in the number of websites that use integrated online learning. There is an increase in the effectiveness of training [4] using web-based knowledge and an increase in consumer satisfaction with online learning.

Initiatives are being created [1] to create a digital repository for peer-reviewed electronic resources for public distribution.

It is proposed to create a mechanism for sharing quality resources for online learning between individual educational institutions and stimulating the work of teachers through peer review processes. Providing learning materials in various formats - using video, audio, other technologies, or software in the form of serious educational games [5] – these are tools that can make learning more interesting and also allow students to improve the quality of their learning.

As a professional, the health care and medical specialist must contribute to [6]:

- promoting healthy habits;
- carrying out disease prevention;
- treatment of patients;
- to work for the rehabilitation of the recovered patients;
- to conduct a constant process of learning everything new.

The main principles of medical and health care education are the acquisition of education leading to the first professional degree; clinical education, which is preparation for a general medical and health care practice or specialty and obligatory continuous, lifelong training of every higher specialist. According to [7] continuing medical education is a very important requirement for the development of medical professionals.

Internet technologies enable the distribution and use of digital content among many users simultaneously at any time and from anywhere in the world. For employed medical professionals [8], the ability to learn when they can or need information is highly desirable. The educational needs in the field of medicine are diverse: providing easy, fast, and timely access to textual content or photos; inclusion of serious games in the form of virtual reality systems, virtual learning techniques, and virtual patient care in health care and medical education.

Research reported in [9, 10] shows that effective learning can be achieved by combining traditional learning with distance learning.

There is a continuous growth in knowledge of the structure and function of the human body. In many areas of application, there is a simultaneous rapid growth of various innovative methods. Today, many factors in higher education make teaching and learning difficult. This includes a shorter hospital stay for critically ill patients and limited time for medical professionals to work with each patient. There is a shift from traditional practical medical actions to computer-based manipulations or images due to the modernization of technology.

According to [11] free access to medical education sites, online medical journals, textbooks, and the latest information on medical innovation, the use of serious games also encourages learning and research.

The development of clinical informatics [12] aims to improve patient care through the use of modern technologies and increase the efficiency of both patient care and safety.

E-learning began to be used about 20 years ago. From then until today, it has undergone significant development - from its original idea to the modern experience in an electronic or web-based application. E-learning enters all stages of education and then continues in the form of lifelong learning. At the beginning of the creation of the web, e-learning included a virtual learning environment (Virtual Learning Environment), expensive software, work plan, and tested. At this earlier stage of development, e-learning is oriented to the interests and goals of the educational institution.

The development of e-learning in the coming years is already oriented to the needs of the learners themselves, who become active participants in this bilateral process. Learners share ideas, offer solutions to problems, collectively use different sources, and can participate in the creation of new knowledge. In this way a learning community is already created, the aim is to support the interaction not only between the teacher and the students but also between the students themselves. According to [13] upgrading on the already acquired knowledge of the learners, stimulating creativity, creative thinking, spontaneous reaction, and working on breaking the traditional education by introducing an element of fun. The role of the teacher and the trainer is expanded and enriched, the qualitative characteristics of the education are put in the foreground, new methods for increasing the motivation of the students are sought.

The study, conducted in [14], aims to examine the extent to which online resources are used by medical graduates and ways to increase interest in them. The authors point out as a conclusion the need to integrate information technology into traditional education, change the curriculum, transition to structured computer training of medical students.

III. Serious Games as a Method for Effective Learning

Computer games, which, unlike games created for entertainment, serve to achieve certain useful goals, supporting the formation of professional knowledge, skills, and competencies, and are called serious games [15, 16]. Serious educational games differ from entertainment games mainly in the presence of pre-set pedagogical goals. According to [17], the serious game is “an interactive computer application, with or without a significant hardware component, that has a challenging goal, is fun to play with, incorporates some concept of scoring, and imparts in the user a skill, knowledge or attitude which can be applied in the real world”. Serious games are gradually becoming a popular modern development [18], due to the support they provide to educational processes, attracting and retaining attention to the learning material and provoking useful behavior in students. Along with the acquisition of knowledge, learners develop problem-solving skills, cooperation, communication, and interaction in a competitive environment, lead to improved creativity and strengthen their participation in the learning process.

Traditional lecture-based teaching [19] emphasizes the transmission of information and the use of memory to achieve educational effects. Serious learning games are attractive to their users because they are not based on the traditional teaching methods that are boring for learners today. Traditional learning is increasingly perceived by learners as ineffective and cannot hold their attention. The games confront students with a problem that engages their attention and interest and offer possible ways to explore the problem situation. In this way, students have the opportunity to upgrade their knowledge. Game design elements are designed to attract students' attention [20] and challenge them to solve problems.

Looking to the near future we can note the implementation of Gamification in education, which can be defined as a learning platform [21] that aims to give additional motivation to learners, provide feedback and use as incentive rewards (points, badges, boards for leaders, progress lanes, etc.). In gamification [22], the whole learning process becomes a game in order to motivate and interact students through game mechanics and game elements. Unlike gamification, game-based learning uses games only as part of the learning process. To the goals of gamification can be added the achievement of an attractive learning process, provided by performing fun activities, gaining a modern learning experience.

The authors of [19] in 2019 conducted a study of the use and impact of serious games scientific education, using mixed qualitative and quantitative data. The opinions of students and parents about the influence of serious educational games were studied. The obtained results show positive results on the motivation of students in the learning process.

Today, web technologies are influential factors in students' lives and learning. Interaction in online and offline media, which offer a modern educational environment, creates significant changes in students [23]. Online educational platforms play a key role in changing students' values, beliefs and behavior.

Modernization of technologies leads to mandatory integration of technology with education and educational systems need to be constantly changed and improved. This leads [24] to the emergence of new approaches to learning and teaching.

According to [25], any serious game must have the following characteristics: a method of communication between the person and the game; assessment by the number of correct answers; conflict (challenge); the ability of users to change the game (controllability); environment; history or fiction; interaction between users; immersion in the game; rules and goals of the game provided to the user.

In the training of health care and medical professionals, it is essential that students be able to enhance their practical skills through the use of games involving virtual simulations of real activities [26] performed on patients. Thus, the trial and error method has no negative consequences on the patients themselves.

Despite the entry of serious games into training [27], game developers do not pay enough attention to validating the effect of using serious games.

Works [28, 29] examined the barriers in web space and the accessibility of serious training games for students with visual impairments.

IV. Results of the Study on the Online Training and Serious Games in Clinical Training

A survey on the issues of online training education and serious games in clinical training in Bulgaria was conducted in April 2021 with students learning in the specialties nurses and midwives at health care and medical Universities in Bulgaria. Teachers' and students' views on virtual methods were surveyed using a survey created using the Google Forms application. The survey includes 486 respondents, health care students in Bulgaria. The online survey was conducted on a voluntary and anonymous basis. The survey was conducted among respondents from the Faculties of Public Health and Health Care of Universities in the Republic of Bulgaria, that train specialists in the professional field of Health Care, such as Ruse University “Angel Kanchev”, Southwestern University "Neofit Rilski" - Blagoevgrad, Burgas University “Asen Zlatarov”, Varna Medical University (and its branches in the country), Universities in Plovdiv, Sofia, Pleven, Stara Zagora. The questions in the survey are 15 in number.

Table I presents the characteristics of the respondents according to their training course. The survey was conducted among students who in April 2021 are studying in the 1st, 2nd, 3rd, and 4th year at higher universities in Bulgaria. The 1st and 3rd year students were the most active in filling in the questionnaire.

Regarding the degree of satisfaction of students with the online training conducted in the academic year 2019/2020 and 2020/2021, 51.4% believe that they like online training completely. 24.3% think that they are not satisfied enough and expected more, and 22.2% cannot give a clear assessment. According to other students who indicated an open answer, online learning cannot replace face-to-face learning (0.8%), 0.6% say they are satisfied with online learning but prefer to learn face-to-face. Several students point out the importance of attending the internship and exercises during the online training. Overall, 67.5% (238) of the students indicated that they preferred the present form of education to online learning.

TABLE I. COURSE OF TRAINING

Course	1	2	3	4
% (Number of participants)	32,9 % (160)	23% (112)	32,5% (158)	11,5% (56)

The study of what students liked the most in online learning is presented in Fig. 1. The question allows you to choose several answers. According to 294 (60.5%) respondents, the biggest advantage of online learning is the created opportunity for Accessibility to learning materials (platforms, online resources, presentations, video materials, etc.). This shows the desire of students to use web educational resources. This opportunity can be for them provided and through real training and lead to an increase in its quality. The opportunity to learn from home is essential for 246 participants, which is more than half of the surveyed students. Respondents point to the advantages of greater independence and flexibility in learning and the ability to use modern technologies.

The study on the shortcomings of online learning according to students is presented in Fig. 2. The question allows you to choose several answers. According to the respondents, the biggest shortcomings are: the lack of direct contact with their teachers (according to 61.7%); the lack of a real hospital environment in which to learn skills for manipulations (according to 56.8%); in training in higher health care specialties, real training is necessary and very important (according to 53.5%); the lack of daily direct contact with fellow students is essential for 51% of respondents.

Fig. 3 shows the degree of response of the respective university in its transition to online learning. According to 48.6% of the respondents, their university immediately switched to online education, while 44% of the students stated that the transition was gradual. 7% of students say that in their higher university there was a distance form of education even before the pandemic.

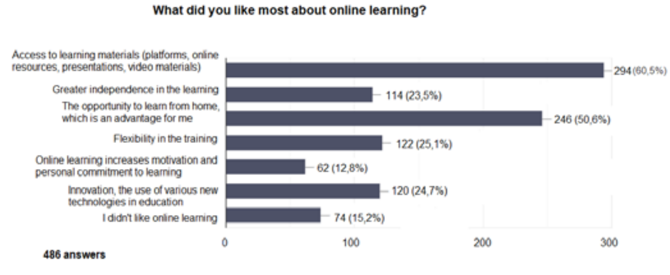


Fig. 1. Exploring the Benefits of Online Learning.

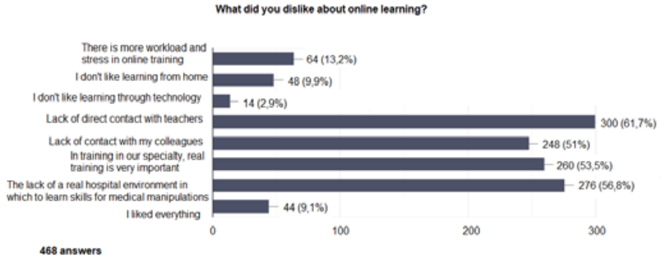


Fig. 2. The Studying the Shortcomings of Online Learning.

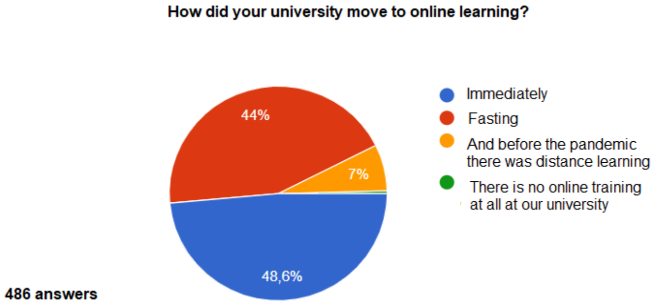


Fig. 3. Studying the Speed of the Transition to Online Learning.

Fig. 4 presents the means of communication through which students connected with their teachers during online learning. According to respondents, the most commonly used means of exchanging information is e-mail (63%), followed by Google Classroom (37%), Viber (74%), and Phone (35.8%).

Fig. 5 shows the degree to which students coped during the online training in mastering the material and practical health care manipulations. 41.6% of the respondents state that they have not encountered problems with online learning, 36.6% declared that they had difficulties at the beginning of online training, which they then coped with, according to 11.1% of students online learning creates difficulties in mastering the health care procedures that require real clinical practice.

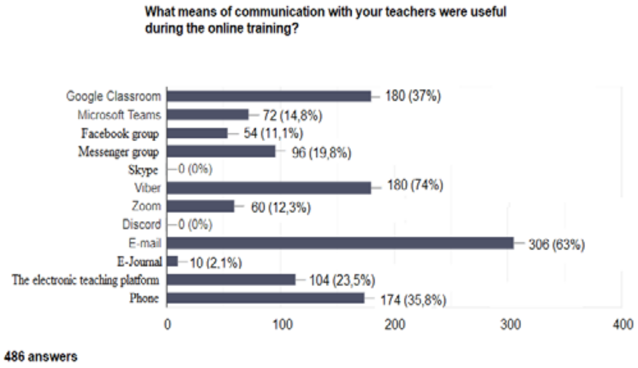


Fig. 4. Means of Communication.

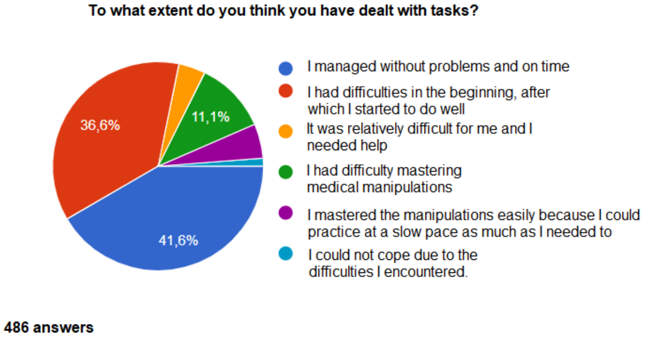


Fig. 5. Degree of Coping in Online Learning.

Fig. 6 presents the useful learning tools that can be used in future online learning according to the respondents. The largest part (68.7%) of the surveyed students believe that the video materials on the respective medical disciplines will be of great help to them in the online training. According to 56.4% of the respondents, video presentations prepared on the lecture material will be of great benefit in their training. 44.4% of the respondents indicate the websites with educational resources uploaded on them as a useful tool in mastering the study material. 24.28% of respondents point to serious games as a modern learning tool.

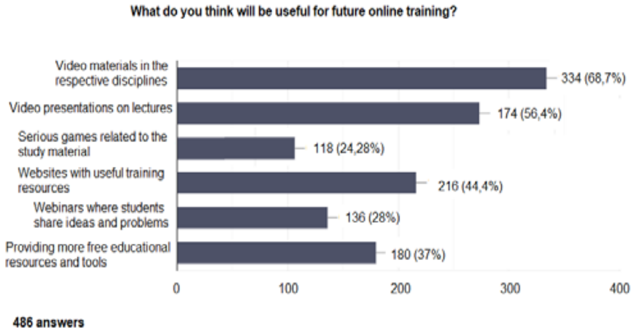


Fig. 6. Useful Tools for Future Online Training.

Regarding the device with which they conducted their training, 65% (316 participants) indicated a telephone, 90.1% (438 participants) used a laptop, 9.9% (48 participants) used a computer, 9.1% (44 participants) used a tablet. The questionnaire allows you to choose more than 1 answer to this question.

According to the survey, 17.3% of students report that they have had problems with the Internet connection, 63% believe that sometimes the connection has been interrupted, and 19.8% of the respondents, the connection has always been good.

V. Conclusion

In today's technological society, the inclusion of innovative interactive digital tools in the process of training in medical specialties is imperative and undoubtedly contributes to improving the quality of teaching. In periods requiring online learning as the only method, it is necessary to use the Internet effectively with all its advantages: access to digitized online learning resources, creation and use of appropriate videos, presentations, webinars, serious educational games, video simulations, and much more.

The results of the survey conducted among students trained in specialties Midwife and Nurse show that higher universities in Bulgaria have done well in the periods of compulsory online training in 2020 and 2021 (until April, when the survey was conducted) and have offered effective, quality training to their students.

The results of the survey can be a tool for assessing and improving the quality of online learning and serve as a corrective in the next periods of online learning in higher universities in Bulgaria.

Future plans and recommendations for the future development of the education of health care and medical students. One of the most important conclusions made as a result of the research is that the real education of health care students cannot be replaced by sufficient quality online education. This shows that the training program for health care students must be carefully considered and include all opportunities for real involvement of students in clinical practice, where they can gain real experience in working with patients.

The results of the study also indicate the positive aspects of the introduction of technology in the education of Health Care students and the application of online learning. The creation of new web-based educational resources (video materials, serious games, virtual simulations, video presentations, webinars, etc.) can complement traditional teaching methods and make higher education in medicine in Bulgaria high quality, modern, effective and attractive for future health care and medical professionals.

Acknowledgment

This work was supported by the Bulgarian Ministry of Education and Science under the National Research Programme “Young scientists and postdoctoral students” approved by DCM # 577/17.08.2018.

References

[1] T. Anastasiadis, G. Lampropoulos, K. V. Siakas. “Digital Game-based Learning and Serious Games in Education”. International Journal of Advances in Scientific Research and Engineering, Vol. 4, Iss. 12, 2018, pp. 139-144, DOI: 10.31695/IJASRE.2018.33016.

[2] N. Hamzah, A. Ariffin, H. Hamid. “Web-Based Learning Environment Based on Students’ Needs”. International Research and Innovation Summit (IRIS2017) IOP Publishing, IOP Conf. Series: Materials Science and Engineering 226, 2017, 012196, doi:10.1088/1757-899X/226/1/012196.

[3] M. Al-Balas, H. I. Al-Balas, H. M. Jaber, K. Obeidat, H. Al-Balas, E.A. Aborajooh, R. Al-Taher, and B. Al-Balas, “Distance learning in clinical medical education amid COVID-19 pandemic in Jordan: current situation, challenges, and perspectives. BMC Medical Education, 2020, 20:341, https://doi.org/10.1186/s12909-020-02257-4.

[4] K. Soussi, “Web-based Learning: Characteristics, Practices, Challenges and Recommendations”. International Journal of Science and Research, 2020, 9(3), pp. 936-943, DOI: 10.21275/SR20312135240.

[5] G. Georgieva-Tsaneva, “Serious Games and Innovative Technologies in Medical Education in Bulgaria”. TEM Journal. Vol. 8, Iss. 4, pp. 1398-1403, ISSN 2217-8309, DOI: 10.18421/TEM84-42, 2019, pp. 1398-1403.

[6] A. Ghanizadeh, S. Mosallaei, M. S. Dorche, A. Sahraian, P. Yazdanshenas. „Use of E-learning in education: attitude of medical students of shiraz“, Iran. Int Med Medical Investigation J. 2018, 3(3), pp.108–11.

[7] M. Steinman, S. Landefeld, R. Baron, “Industry Support of CME — Are We at the Tipping Point?” The New England Journal of Medicine, 2012, vol. 366, pp. 1069-1071.

[8] D. Pullen, “Doctors online: Learning using an internet based content management system. International Journal of Education and Development using Information and Communication Technology”, 2013, Vol. 9, Iss. 1, pp. 50-63.

[9] I. Gorbanev, S. Agudelo-Londoño, R.A. González, A. Cortes, A. Pomares, V. Delgadillo, et al. “A systematic review of serious games in medical education: quality of evidence and pedagogical strategy”. Med Educ Online, 2018, 23(1):1438718, doi: 10.1080/10872981.2018.1438718.

[10] H. Banda, R. Franco, D. Simpson, K. Brennan, J. McKanry, & D. Bragg, “Assessing the learning outcomes and cost effectiveness of an online sleep curriculum for medical students”. Journal of Clinical Sleep Medicine, 2012, vol. 8, No. 4, pp. 439-443.

[11] M. Azeem. „Ten ways to improve information technology in the NHS“. BMJ. 2003 Jan 25; 326(7382): 202–206.

[12] S.V. Gentry, A. Gauthier, B.L. Ehrstrom, D. Wortley, A. Lilienthal, L.T. Car, et al. „Serious gaming and gamification education in health professions: Systematic review“. J Med Internet Res, 2019, Mar 28;21(3):e12994, doi: 10.2196/12994.

[13] J. Dašić, P. Dašić, V. Šerifi, “Evolution of E-Learning”. 7th International Conference ICQME 2012. 19th-21st Sept. 2012, Tivat. pp. 311-316.

[14] P. Kasat, S. Gupta, R. Jadhav, G. Muthiyan, “The Study of Usage of Online Learning Resources in Medical Courses”. International Journal of Research in Electronics and Computer Engineering Vol. 6 (1). 2018. Pp. 1-6.

[15] D. Paneva-Marinova, M. Rousseva, M. Dimova, L. Pavlova. “Tell the Story of Ancient Thracians through Serious Game”. Ioannides M. et al. (eds) Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. EuroMed 2018. October 29th – November 3rd, 2018. Cyprus, 11196 LNCS, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 509-517, https://doi.org/10.1007/978-3-030-01762-0_44.

[16] M. Dimova, D. Paneva-Marinova, L. Pavlova. “Towards Better Understanding of Ancient Civilizations by Storytelling and Gaming”. TEM Journal, 2018, 7 (3), UIKTEN - Association for Information Communication Technology Education and Science, Serbia, ISSN:2217-8309.

[17] B. Bergeron, “Appendix A: glossary In Developing serious games”. Hingham: Charles River Media; 2006. p. 398.

[18] C. Hursen and C. Bas, “Use of gamification applications in science education,” Int. J. Emerg. Technol. Learn., vol. 14, no.1, pp. 4–23, 2019.

[19] W. Bedwell, D. Pavlas, K. Heyne, et al. “Toward a taxonomy linking game attributes to learning: an empirical study”. Simul Gaming. 2012; 43(6):729–760.

[20] C.-H. Su, C.-H. Cheng, “A mobile gamification learning system for improving the learning motivation and achievements”. Journal of Computer Assisted Learning. Vol. 31 (3), 2014, pp. 268-286. https://doi.org/10.1111/jcal.12088.

[21] F. Gokkaya, “Peer Bullying in Schools: A Cognitive Behavioral Intervention Program”. In book: Child and Adolescent Mental Health. 2017. DOI: 10.5772/66701.

[22] R. Al-Azawi, F. Al-Faliti, and M. Al-Blushi, “Educational Gamification Vs. Game Based Learning: Comparative Study”. International Journal of Innovation, Management, and Technology, Vol. 7, No. 4, 2016, pp.132-136.

[23] S. Nemorin, “Affective capture in digital school spaces and the modulation of student subjectivities”. Emotion, Space and Society, 2017, 24, 11-18. https://doi.org/10.1016/j.emospa.2017.05.007.

[24] M. Gencer, T. Tok, & A. Ordu, „Power Base Games That School Principles Use Scale: Its Development, Validity, and Reliability“. International Journal of Assessment Tools in Education, 2018, 5(2), 274-288. DOI: 10.21449/ijate.407218.

[25] I. Gorbanev, S. Agudelo-Londoño, R. Gonzales, A. Cortes, A. Pomares V. Delgadilo, F. Vepes, and M. Óscar, “A systematic review of serious games in medical education: quality of evidence and pedagogical strategy.” Medical Education Online. 2018; 23 (1); https://doi.org/10.1080/10872981.2018.1438718.

[26] B. Spellberg, D. Harrington, S. Black, et al. “Capturing the diagnosis: an internal medicine education program to improve documentation.” Am J Med. 2013;126(8):739–743.

[27] M. Graafland, J. Schraagen, M. Schijven, “Systematic review of serious games for medical education and surgical skills training.” Br J Surg. 2012, 99(10):1322–1330.

[28] N. Noev, G. Bogdanova, T. Todorov, N. Sabev. „Innovative approach to the presentation of cultural heritage in the game module of serious game for blinded people“. Digital Presentation and Preservation of Cultural and Scientific Heritage, 9, 213-218, 2019, ISSN 1314-4006 (Print) ISSN 2535-0366 (Online).

[29] G. Bogdanova, N. Sabev, N. Noev. „Accessibility and some educational barriers for visually impaired users“. Proc. of the 11th International Technology, Education and Development Conference, Mar 2019, doi: 10.21125/inted.2019.2333, 2019, ISBN: 978-84-09-08619-1, ISSN: 2340-1079.

An Implementation of Hybrid Enhanced Sentiment Analysis System using Spark ML Pipeline: A Big Data Analytics Framework

Raviya K[1]
Research Scholar, PG and Research Department of
Computer Science, Presidency College
Chennai, India

Dr. Mary Vennila S[2]
Associate Professor and Research Supervisor,
PG & Research Department of Computer Science
Presidency College, Chennai, India

***Abstract*—Today, we live in the Big Data age. Social networks, online shopping, mobile data are main sources generating huge text data by users. This "text data" will provide companies with useful insight on how customers view their brand and encourage them to make business strategies actively in order to maintain their trade. Hence, it is essential for the enterprises to analyse the sentiments of social media big data to make predictions. Because of the variety and existence of data, the study of sentiment on broad data has become difficult. However, it includes open-source Big Data platforms and machine learning techniques to process large text information in real-time. The advancement in fields including Big Data and Deep Learning technology has influenced and overcome the traditional restrictions of distributed computing. The primary aim is to perform sentiment analysis on the pipelined architecture of Apache Spark ML to speed upward the computations and improve machine efficiency in different environments. Therefore, the Hybrid CNN-SVM model is designed and developed. Here, CNN is pipeline with SVM for sentiment feature extraction and classification in ML to improve the accuracy. It is more flexible, fast and scalable. In addition, Naive Bayes, Support Vector Machines (SVM), Random Forest, Logistic Regression classifiers have been used to measure the efficiency of the proposed system on multi-node environment. The experimental results demonstrate that in terms of different evaluation metrics, the hybrid sentiment analysis model outperforms the conventional models. The proposed method makes it convenient for effective handling of big sentiment datasets. It would be more beneficial for corporations, government and individuals to improve their great value.**

Keywords—Big data; sentiment analysis; machine learning; apache spark; ML pipeline

I. Introduction

The success of Smart devices' makes people's daily lives more focused to mobile services. People use mobile devices to collect information about firms, products, deals and recommendations. Online consumer reviews for a wide variety of goods and services are widely accessible and evaluating the sentiment in customer feedback has become greatly useful for business, where companies can monitor positive and negative feedback about the brand which allow themselves to assess its over-all success and can also perform a major role in evaluating sales and optimizing business marketing approaches. Reviews from customers are one of the massive amounts of information. Since it includes millions of reviews from different websites, and the number of reviews is rising every day. This vast amount of data that increases every moment is known as big data, which involves modern technologies and architectures to capture and evaluate process to derive value from it [1]. Big data analytics is essential for the purposes of business and society. Big data requires strong machine learning methods, and environments to accurately analyses the data. A large amount of data cannot be processed by conventional methods, so to handle the huge amount of information, a new computing platform for big data, such as Apache Hadoop and Apache Spark, are intended to incorporate machine learning systems to attain high performance [2], [3]. Apache Spark, established in 2009 at University of California, is an open-source processing system. It has been one of the main frameworks in the world for large-scale data processing and analytics, achieving high efficiency for both batch and stream data. It is an API that is simple to use and run-on large datasets. For large-scale data processing, Spark is 100 times faster than Hadoop by utilizing memory computing and other optimizations. Sentiment Analysis of huge volume of data has become more and more significant and drawn many researchers. Sentiment Analysis, also referred like opinion mining, is characterized as a task to identify the views of authors on specific entities [4]. Sentimental analysis is used in various places, for example: To analyse the reviews of a product whether they are positive or negative, If a political party strategy has been successful or not, evaluate the ratings of a film and analyse information of tweets or another social media data [5]. Sentiment analysis is all about having people's real voice of a particular product, programs, organization, movies, news, events, problems and their characteristics. Social media monitoring apps in businesses rely primarily on sentiment analysis using machine learning to help them gain insights into mentions, brands and goods [6]. The machine learning is a subset of AI [7]. It trains the computers to learn and behave like human beings with the assistance of algorithms and data [8]. Machine learning is the science of preparing a system to learn and act from data [9]. Machine learning is being used by wide variety of applications, and the trend is rising every day and often denoted to as fixing the model with knowledge is the method of training the model. Fig. 1 depicts two sub-sections of machine learning algorithms.

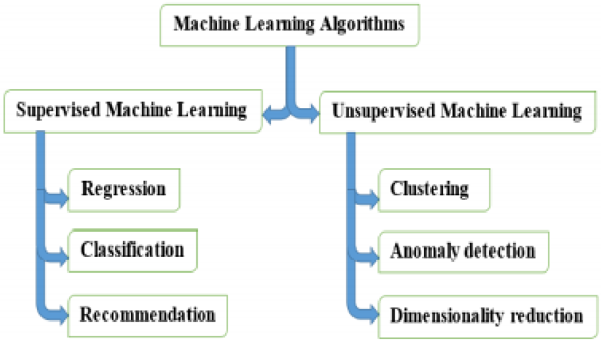


Fig. 1. Machine Learning Algorithms.

Supervised machine learning refers to working with a set of labelled training data to learn [7]. Every observation has a collection of features and label in the training dataset. Algorithms for supervised machine learning can be classified into regression, classification and recommendation engines. Unsupervised machine learning has been used when a dataset is un-labelled, means when a model does not require labelled data that is referred to unsupervised learning. These types of models try to learn or discovering hidden structures in un-labelled data or reduce the data down to its most important features [9]. With unsupervised learning there is no right or wrong answer [7]. They are commonly used during clustering, detection of anomalies and reduction of dimensionality. Sentiment analysis is recognized as a problem of classification and it can also be solved by the method of machine learning techniques [10]. In this work, MLlib by Spark, that is the machine learning libraries developed on top most layer of Spark to provide great-quality and high-speed machines. MLlib utilized Java, Scala, and Python, so that this may incorporate it into full workflows [8] that can be used for data analysis of large scale and since Spark's MLlib is a recent library developed in 2014. According the limited awareness of researchers, A small number of researches have been carried out to measure the sentiment of large-volume data using Spark's MLlib, so more analytical work is necessary for this field. The purpose of this study is to have new sentiment classification experiments on large volume of data by the Spark ML and DL with TensorFlow by implementing deep-learning models and evaluating their performance with existing algorithm. The rest of this paper is laid out as follows. In Section 2, we begin with a related work by Apache Spark ML. The core components of Apache Spark architecture are then introduced in Section 3. Section 4 introduces Apache Spark's machine learning library and computation. The ML pipelines for machine learning in Spark are discussed in Section 5. Then, for the suggested solution in Section 6, we move on to custom DL pipelines. Following that, Section 7 discusses some of the ML classifiers. The proposed methodology for sentiment analysis through big data analytics using hybrid CNN-SVM with spark DL is implemented in Section 8. Dataset, pro-processing, and word embedding are all discussed in Sections 9, 10, and 11. Section 12 and 13 describes the experimental setup, results, and discussions. Finally, summary and conclusions of this paper is presented.

II. Related Work

There is also a large amount of research on distributed systems for solving big data problems, particularly using Apache Spark. Sentiment analysis is the most ongoing research field in recent years that researchers have concentrated on, and several researchers have used various methods to perform sentiment analysis. The enormous interest in the field of sentiment analysis is largely dependent on availability of information and the developments in the internet. Advances in new techniques and algorithms have shown that combining sentiment analysis with machine learning can provide greater potential to predict the success of newly released products. Baltas et al. [8] implemented a Twitter data sentiment analysis system that used Spark MLlib for classification. On real Twitter info, three algorithms were used: decision tree, Naïve Bayes and logistic regression with binary and ternary classifications were evaluated. The Pre-processing of data is handled to maximize performance. The framework was tested on various sizes of datasets and various features. The authors suggested Naïve Bayes is superior than other classifiers and then the size of the dataset could influence the output of the classifiers. Sayed and et al [11] discovered in their paper that the Spark ML has an attraction over Spark MLlib in the performance and accuracy of big data analytics problems. Al-Saqqa and et al [12] examined about sentiment classification of big data using Spark's MLlib. In terms of efficiency, they find that the SVM is higher than other classifiers. AL-barznji and et al [13] addressed sentiment analysis using algorithms such as Naïve Bayes and SVM to evaluate the text with Apache Spark's aids. They discovered that in all situations, the SVM is more specific. Some recent works on deep learning models using Apache Spark as follows. The authors in [16] used a deep learning approach to detect the sentiment of Arabic tweets. Their approach relies on the utilization of pre-trained word vector representations. In [14], The ensemble model that combines both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models. LSTM is a special type of RNN which can learn long-term dependencies. LSTM was developed to prevent the long-term dependency problem that exists in standard RNNs. The results show good improvements in the accuracy and f1-score measurements over some other flat approaches. Alsheikh et al. [17] proposed a deep learning model for mobile big data analytics using Apache Spark. This model suggested the deep model parallelization by slicing the MBD into several partitions in spark RDD. The results illustrated the achievement of a better implementation using deep learning models through the Spark system compared to traditional lighter models. Various methods discussed in this paper for sentiment analysis are mainly based on results regardless of time complexity. If the size of the information is small, the analysis will be finished within the time limit. But time is a significant constraint when analyzing broad corpus, so the proposed structure would decrease the time required for sentiment analysis. This approach is much simpler and fully utilizes the capabilities of Spark ML framework. It is the Spark-based large-scale approach to sentiment analysis of product review data set without the need to build a sentiment lexicon or proceed with any manual data transcription.

III. APACHE SPARK ARCHITECTURE FOR BIGDATA

Apache Spark is a platform for bigdata analytics that can offer an enhanced substitution to the Map Reduce model. In contrast to map reduce model, Spark does not push the data to a disk for every step. The data is gathered in the memory till it is fully stored. If the memory is full then the data flows over to the hard drive. Therefore, the advantage of in-memory processing applied in Spark can make its processing very fast. The architectural layout of Spark is described in Fig. 2.

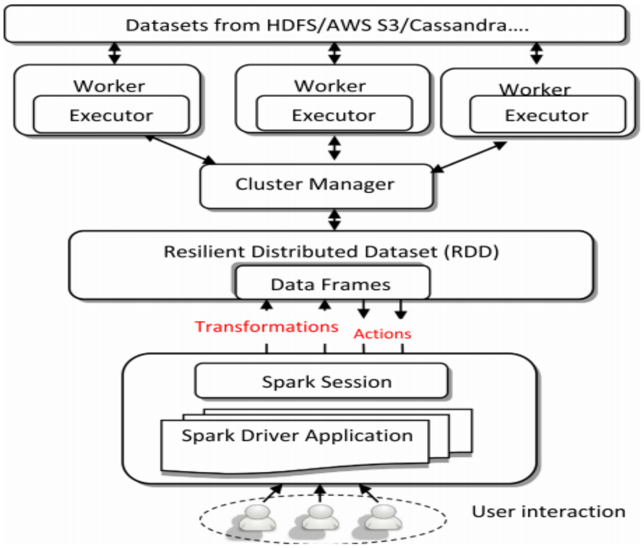


Fig. 2. Spark Architecture.

The Spark System utilizes a master/worker layout in which the master node handles the worker nodes. If the spark is mounted on the cluster, the executors mostly in worker nodes are automatically constructed and then the task is executed according to the instructions provided by the cluster manager. In the context of the Spark Session, the driver serves as the user interface from which the user transmits and receive the instructions. The Spark Session serves as the primary gateway for all communication. Spark operates at the centre on the basis of RDD idea. RDD's include applications like distributed data processing, the ability to use multiple data sources, fault tolerance and parallelism. Spark conducts two fundamental activities, i.e., transformations and actions. The transformations perform the RDD work, which transforms the input data using activities such as mapping, joining and key reduction to return the final output to RDD's and actions received the data from RDD's.

IV. SPARK MACHINE LEARNING LIBRARY

Apache Spark is a very powerful tool for analytics of big data and presents excellent performance in terms with running time. MLlib is the first library supplied with Spark for machine learning shown in Fig. 3. Unlike single node machine learning frameworks, it is much efficient and scalable. MLlib also offers distributed processing options through parallel processing as well as facilitates the use of distributed architectures for big data analytics. This criterion would reduce the processing time needed but, at same time, control the duration to evaluate analytical results. If the role of machine learning has several predictions to measure and it is highly important. It provides some of the big data tool attempts to break down each machine learning section can also be used by the distributed architecture to decrease the total running time.

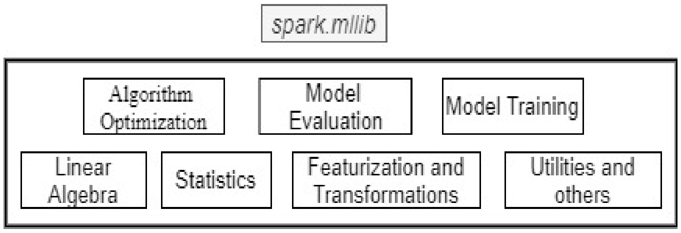


Fig. 3. Spark MLlib.

Integration is an additional benefit of Spark MLlib, this means that MLlib benefits from many software components accessible in the Spark ecosystem. It includes Spark GraphX, SQL, Spark Streaming and large spectrum of highly organized documents are freely accessible to the machine learning community. The basic machine learning utilities offered by spark MLlib are shown in Fig. 3. Spark Machine Learning Libraries provide executions of many algorithms and it is used for basic machine learning methods includes classification, regression, clustering, reduction of dimensionality, extraction and transformation of features, regular mining of patterns, and recommendations.

V. SPARK ML PIPELINES

Apache Spark ML is open-source platform for fast processing of large-scale data. Here Spark ML offers high-level APIs made on top of Data Frames for scalable data processing. In order to extract and predict the features, Spark uses the ML Pipeline to pass the data via transformers and estimator to implement the model.

Transformer: It is an algorithm to converts one Data Frame into next Data Format. Transformers are used here to convert text data with a feature vector.

Estimator: It is an algorithm to fit for the Data Frame to generate transformer. Estimators are used to train the model, which can convert input data to do the predictions.

Pipeline: The pipeline connects several transformers with estimators together to describe the ML workflow and it provides the mechanisms to build, evaluate and fine tune pipelines.

ML pipeline consisting of a set of pipeline stages to make it simpler using multiple algorithms to be merged into a single pipeline or workflow to be operated in a particular order. Spark is completely consistent with applications based on Java, as it uses Scala which runs on Java Virtual Machine. Using PySpark, it can work with RDD in Python programming language also. Further, Spark MLlib contain RDD format-based implementation of machine learning algorithms. Spark ML is based on datasets and enables us to use Spark SQL with it.

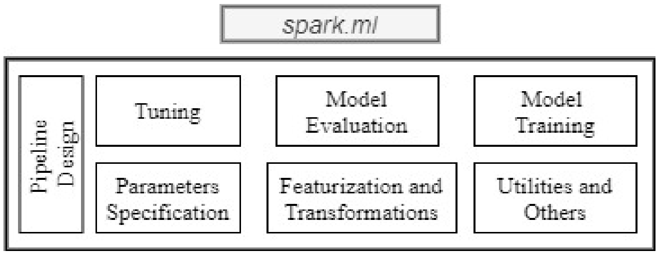


Fig. 4. Spark ML.

Feature extraction and manipulation tasks are very simple, as Spark SQL queries can now handle a lot. Typically, Dataset is in an extremely raw text, and this data usually goes through a loop or workflow where it is pre-processed, transmuted, and transformed until it is consumed for model training. The complete work flow of data transformation and its stages are fully encapsulated with the concept of ML Pipeline shown in Fig. 4. The ML libraries provide high-level API for implementation and fine-tuning of various machine learning pipeline models. It also allows us to save and load machine learning algorithms, trained models and pipelines so that it could reuse the results of previous steps to generate new results without the need to start the whole process from scratch.

VI. Apache Spark Custom DL Pipelines

The Deep Learning Pipelines is a high-level framework, which is applicable through Apache Spark MLlib. It facilitates the implementation of popular deep learning workflows. Furthermore, Databricks is an establishment started by the originators of Apache Spark provided the open-source library called DL pipelines and it support to develop deep learning models using python. It is a large-scale API integrate with the power of deep learning libraries such as TensorFlow and Keras. Both ML and DL provide a consistent collection of high-level APIs assembled on top of Data Frames to help users to build and tune functional pipelines for machine learning and deep learning models. To achieve the best outcomes in deep learning, experimenting with different training parameter values is an essential step called hyperparameter optimization. Since deep learning pipelines allow the exposure of deep learning training to Spark machine learning pipelines as a step, users can also rely on the integrated tuning work for hyperparameters.

VII. Spark ML Classifier

A. *Naive Bayes (NB)*

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, and recommendation systems. They are quick and simple to implement. This is a basic multiclass classification algorithm based on Bayes' theorem. The conditional probability from each class function is calculated first and then the theorem of Bayes is employed to predict an instance's class label. Naïve Bayes is highly suitable for large dataset. Usually, the accuracy of NBC is increased when the data size increase.

B. *Support Vector Machine (SVM)*

It is a supervised algorithm for classification. The SVM is an efficient classifier that has been successfully used in all aspects of text classification. Documents of text are denoted by a vector. The classification is achieved via defining the hyperplane that raises the margin among two categories, and the support vectors are the vectors that describe the hyperplane. Based on the mined features of the training dataset, the model seeks to find the hyperplane defined by vectors that divide the positive and negative training vectors with largest probability.

C. *Logistic Regression (LR)*

Logistic regression is a model of regression where one value out of specific number of values can be taken by the dependent variable. It determines the relationship between the instance class and the extracted input features using the logistic function, while it is commonly used for binary classification, it can also be used to solve problems with multiple classes.

D. *Random Forest (RF)*

This supervised algorithm is also known as random decision forest and is commonly used for classification, regression, and other tasks. A forest is described as a grouping of trees. It consists of a large number of separate decision trees that work together to form an ensemble. All tree in the random forest is given a class prediction, and the model prediction is made using the class with the most votes. Random decision forest has been considered a robust and reliable classifier due to the principle of bagging and bootstrapping.

VIII. Proposed CNN-SVM using Spark DL

CNN is an artificial neural network that shares their weight. This creates CNN more analogous to the Network of biological neurons and decreases the complexity of both the weight and the network model. A fully connected soft-max layer is utilized as the classification layer for sentence level sentiment classification in CNN proposed by Kim [11]. This classification layer, however, has become too simple for the task of classifying sentiments. Fortunately, the CNN pooling layer output values are considered as function vectors of the input word. They may use as the input of some other classifiers. In this article, we propose an SVM classifier based on CNN that considers CNN as an automated feature learner and SVM as the classifier of sentiment. CNN outputs, the distributed function illustrations of the input terms, are considered to be features of SVM shown in Fig. 5.

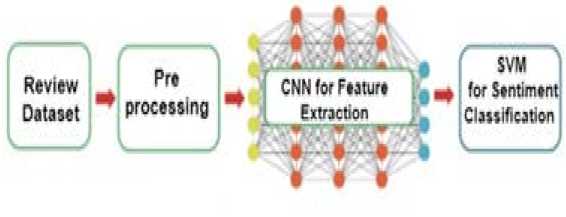


Fig. 5. Proposed Hybrid CNN_SVM Pipeline.

The CNN consists of four layers, namely the input, convolution, pooling and fully connected layer. To create feature map in the convolutional layer, all the available windows of words in the sentence are included. A max-over-time pooling operation is introduced to the function map following the convolution operation. This operation represents the pooling layer and gets a function vector of the m-dimension where the filter size is m. This operation establishes the layer of pooling and gets a vector of m-dimensional feature, where n is the total number of filters. Multiple filters of various window sizes are used in the CNN model. Then these features are moved to the last sheet, such as the fully connected layer, the output of which is the distribution of probability between labels. The pre-trained word embedding is fine-tuned throughout the training processing of the CNN by back propagation. Fine-tuning helps them to learn quite detailed word representations. If the terms may not exist in the embedding of pre-trained words, they are arbitrarily prepared. The vectors of the completely linked CNN layer are considered to be representations of the distributed sentence function, and then these representations of sentences are considered to be feature vectors of SVM classifier. The SVM classifier is trained using the feature vectors labelled with this sentiment. When this method is established for sentiment classification, the incoming sentences are converted to distributed feature representations, which are then fed into the SVM classifier for classification. It is predicted that such a pipeline model will join the benefits of CNN and SVM.

IX. Dataset

The Amazon online product review of about 100,000 reviews is used as data set for this study. This dataset contains products reviews of various domains like electronics, home appliances and books collected from amazon.com websites. To estimate the generalization error, the datasets is divided into two parts: training and testing data. After loading the datasets into the system, before applying and evaluating models the datasets are split into 80 % train dataset and 20 % test dataset randomly. In which, the training dataset is used to build a model, that is used for training the models to get predictions or recommendations. But testing dataset is independent of the training dataset, which is not used in the process of the building a model. The test dataset is to determine the efficiency of the proposed model.

X. Pre-Processing

The pre-processing of the data is the most key step. The purpose of the steps is to make data more machine-readable. Hence, uncertainty is reduced in feature extraction. In addition, to convert the streaming input to a data frame in order to run a pre-processing pipeline that includes the following steps:

Removing null reviews: This involves deleting any reviews with a null value.

Tokenization: The text is subdivided in to smaller tokens based on separator characters like white space, comma, tab, and so on in this phase.

Noise removal: This step involves removing any irrelevant information from the text that could affect the classifier's performance, including such numbers, punctuation marks, URL links, and special characters.

Stop-words removal: Non-descriptive words that can be displaced within the bag-of-words approach are known as stop words. Articles, prepositions, conjunctions, and pronouns are removed because they are not semantically necessary to characterize the viewpoint.

XI. Word Embedding

The interesting properties inside the data which you can use to make predictions are known as features. The process of converting raw data into inputs for a machine learning algorithm is known as feature engineering. For use in Spark machine learning algorithms, Features must be translated into feature vectors, which are numerical values that represent each feature's value. But for deep learning model, a neural network is a collection of neuron layers with the output of one layer being fed into the next layer. Each layer passes on the modified version of data to the next layer to promote more informative features further. Neural networks can't process direct terms; instead, they operate with word embeddings, or more precisely, feature vectors that represent certain words [15]. Neural networks can apply to any domain while learning features from the task at hand. Word2Vec is a predictive model for learning word embeddings from unstructured text that is computationally efficient. The first layer of CNN is the embedding layer converts words into real-valued feature vectors (embeddings) that take morphological, syntactical and semantic information of the words. the CNN use word embeddings feature as an input for the system. Each and every word was thus encoded as a 300-dimensional word vector that was supplied to the network. Word2vec is for word level embedding. Word level embedding is expected to obtain syntactic and semantic details, and character level embedding is projected to grab type and morphologic details. Data source on Google News (approximately words of 100 billion) is used to train the vectors in the proposed technique.

XII. Experimental Setup

Spark ML and DL have been used (open-source Bigdata tool) as development environment for performing the experiments. SparkFlow will take advantage of Spark ML's most important machine learning feature, which is the ability to combine deep learning pipelines with TensorFlow. SparkFlow allows users to train deep learning models in Spark and then link the trained model to a pipeline for smooth raw data predictions. The CNN models have been developed using TFLearn (a deep learning library) and it is pipeline with SVM using Spark ML pipeline.

Fig. 6 represents the major stages of this approach. The proposed method starts with data pre-processing and feature extraction, followed by the use of machine learning classifiers, Naïve Bayes, Support vector machine and logistic regression separately under Spark ML and proposed CNN-SVM using Spark DL environment. Finally, different metrics are used to measure the results. The PySpark library is built with the necessary Python API to run applications on top of the Spark. In order to estimate the efficiency of the proposed model, a series of tests were carried out, specifically, in terms of running time and classification results. There are five models are utilized for classification such as Naive Bayes, Logistic Regression, SVM and Random Forest and proposed hybrid CNN-SVM. Hyper parameter tuning is a technique for deciding the best parameters for achieving the highest degree of precision for the proposed model. Grid-Search with 5-fold cross validation has been applied on training data. The scalability and speed of the method is investigated in this experiment. The experiment is run four times: 1000, 2000, 4000, and 8000 reviews to see how easily Apache Spark can process data using the algorithm.

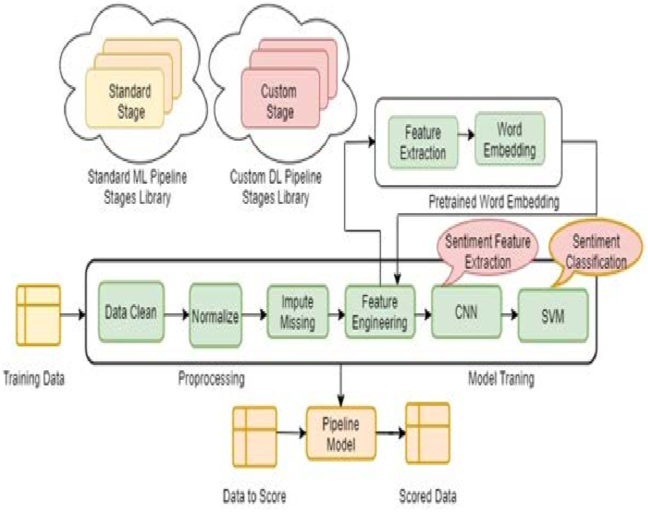


Fig. 6. Hybrid CNN_SVM using Spark DL Pipeline.

XIII. Result and Discussion

The results obtained based on processing time in Table I illustrate that in comparison to other sentiment analysis models, the proposed CNN-SVM has the fastest speed. Each algorithm's execution time is registered. There is a positive relationship between the amount of review data and processing time as it increases.

The classification accuracy of the models' performances is assessed using the random split method. Assuming that accuracy is influenced by a variety of factors, Table II displays average results of 5 algorithms based on assessment parameters such as average. For single-node systems, the values are taken into account. The model created with the CNN-SVM classifiers outshines the other classifiers. The outputs produced through this hybrid CNN_SVM Pipelined technique show higher rates of accuracy. Spark MLlib is a versatile method for analysing big data as evidenced by the findings of this research.

It presents spectacular performance in terms of running time and sentiment analysis of domain independent datasets shown in Table III. It is predictable that much greater performance is achieved in multi-node start-up configurations, as it is evaluated in ten node environments with much larger data sets. It also compared running time on growing number of nodes with varying size of data.

Table IV and Fig. 7 display the experimental outcomes. First, the computational efficiency of Spark is rising as the number of nodes in the computing cluster increases, and the subsequent experimental results indicate that the running time decreases. Secondly, the improvement of computational performance is stronger when this method is adopted to larger data. The results indicate that our proposed system performed well both in accuracy and running time.

TABLE I. Processing Time vs No of Reviews on Single Node

Algorithm	Time taken for no. of Reviews (seconds)			
	1000	2000	4000	8000
Naive Bayes	21s	29s	36s	50s
Logistic Regression	17s	18s	20s	22s
Random Forest	16s	16s	17s	19s
SVM	15s	15s	15s	16s
Proposed CNN-SVM	10s	10s	11s	11s

TABLE II. Accuracy of Various Models on Single Node

Algorithm	Accuracy for no. of Reviews				Average Accuracy
	1000	2000	4000	8000	
Naive Bayes	0.73	0.76	0.77	0.76	0.75
Logistic Regression	0.68	0.69	0.70	0.71	0.69
Random Forest	0.75	0.76	0.77	0.77	0.76
SVM	0.89	0.88	0.90	0.91	0.89
Proposed CNN-SVM	0.94	0.95	0.96	0.96	0.95

TABLE III. Evaluation Metrics for Various Domains

Domain	Performance Metrics			
	Precision	Recall	F-Score	Accuracy
Electronics	0.95	0.94	0.96	0.95
Kitchen Appliances	0.94	0.95	0.96	0.96
Books	0.94	0.93	0.94	0.94

TABLE IV. Number of Nodes vs Running Time

Nodes	Running time in seconds (50,000 Reviews)	Running time in seconds (1,00,000 Reviews)
(1)	240	461
(2)	148	266
(3)	110	180
(4)	89	154
(5)	61	102
(6)	52	84
(7)	45	72
(8)	42	63
(9)	40	55
(10)	35	50

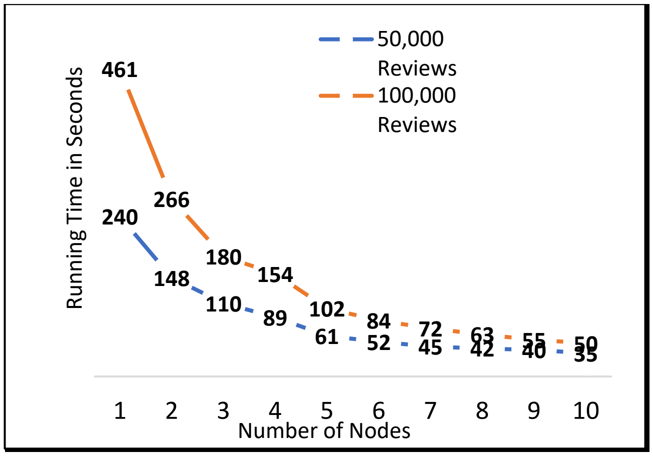


Fig. 7. Number of Nodes vs Running Time.

XIV. Conclusion

The main focus in this study was on rapidly implementing sentiment analysis on the Big Data sets. Spark MLlib has been used for handling a large volume of data as it is scalable. This paper offered new studies to classify sentiment on large amounts of data using Spark's MLlib with TensorFlow by implemented the proposed deep learning CNN-SVM model and the performance of this model is compared with different machine learning classification algorithms. Four classifiers were compared with our proposed model in terms of accuracy. The evaluation result shows that the proposed model has improved performance over the other classifiers. This work was implemented on multi node configuration with larger dataset. As part with our role in the future, we are working to perform an experimental evaluation of Spark MLlib in a number of programming languages (e.g., Python and R), and Software configurations that use a collection of large datasets with a range of data characteristics. In addition, we will develop a better deep learning model to extract optimized features in order to boost performance against other classification methods at a faster rate under large data volumes. The accuracy could be pointed for further development in future.

References

[1] S. Lenka Venkata, "*A Survey on Challenges and Advantages in Big Data*," vol. 8491, pp. 115–119, 2015

[2] Ramesh R, Divya G, Divya D, Merin K Kurian, and Vishnuprabha V, "*Big Data Sentiment Analysis using Hadoop*", IJIRST, Volume 1, Issue 11, pp. 92-98, 2015.

[3] Mohammed Guller, "*Big Data Analytics with Spark*", ISBN13 (pbk): 978-1-4842-0965-3, 2015.

[4] Nurulhuda Zainuddin, Ali Selamat," *Sentiment Analysis Using Support Vector Machine*", IEEE International Conference on Computer, Communication, and Control Technology (I4CT 2014), Kedah, Malaysia,pp.333-337, 2014.

[5] Rajat Mehta,"*Big Data Analytics with Java*", Published by Packt Publishing Ltd, ISBN 978-78728-898-0, UK, 2017.

[6] Kamal Al-Barznji, Atanas Atanassov, "*A Framework for Cloud Based Hybrid Recommender System for Big Data Mining*", a journal of "Science, Engineering & Education", Volume 2, Issue 1, UCTM, Sofia, Bulgaria, pp. 58-65, 2017.

[7] Jason Bell, "Machine Learning: Hands-On for Developers and Technical Professionals", Published by John Wiley & Sons, Inc., Indianapolis, Indiana, 2015.

[8] Baltas, A., Kanavos, A., & Tsakalidis, A. K. , " *An apache spark implementation for sentiment analysis on twitter data.*" In International Workshop of Algorithmic Aspects of Cloud Computing (pp. 15-25). Springer, Cham.

[9] Boštjan Kaluža, "*Machine Learning in Java*", first published: Published by Packt Publishing Ltd, UK, 2016.

[10] Nick Pentreath, "*Machine Learning with Spark*", Published by Packt Publishing Ltd. BIRMINGHAM – MUMBAI, 2015.

[11] Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholief, "*Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages*", A Comparative Study," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 9, pp. 674-677, Nov 2018.

[12] Samar Al-Saqqaa, b, Ghazi Al-Naymata, Arafat Awajan, "*A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Spark*," in The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, EUSPN Belgium, 2018.

[13] Kamal Al-Barznji, Atanas Atanassov, "*Big Data Sentiment Analysis Using Machine Learning Algorithms*," in Proceedings of 26th International Symposium "Control of Energy, Industrial and Ecological Systems, Bankia, Bulgaria, May 2018.

[14] S. Hochreiter and J. Schmidhuber, "*Long short-term memory*", Neural computation, Volume 9 Issue 8, November 1997, pp. 1735–1780.

[15] L. Almuqren and A. Cristea, "*Framework for sentiment analysis of Arabic text*", Proceedings of the 27th ACM Conference on Hypertext and Social Media, Halifax, Nova Scotia, Canada, July 10-13 2016, pp. 315-317.

[16] L. Al-Horaibi and M. Khan, "*Sentiment Analysis of Arabic Tweets Using Semantic Resources*", International Journal of Computing and Information Sciences, Volume 13 Issue 1, January 2017, pp. 9-18.

[17] Alsheikh, M.A., Niyato, D., Lin, S., Tan, H.-P., Han, Z., 2016. , "*Mobile Big Data Analytics Using Deep Learning and ApacheSpark*", 22–29. https://doi.org/10.1109/ MNET.2016.7474340.

Traffic Engineering in Software-defined Networks using Reinforcement Learning: A Review

Delali Kwasi Dake[1], Griffith Selorm Klogo[3]
Henry Nunoo-Mensah[4]
Department of Computer Engineering, Kwame Nkrumah
University of Science and Technology, Kumasi, Ghana

James Dzisi Gadze[2]
Department of Telecommunications Engineering
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

***Abstract*—With the exponential increase in connected devices and its accompanying complexities in network management, dynamic Traffic Engineering (TE) solutions in Software-Defined Networking (SDN) using Reinforcement Learning (RL) techniques has emerged in recent times. The SDN architecture empowers network operators to monitor network traffic with agility, flexibility, robustness and centralized control. The separation of the control and the forwarding plane in SDN has enabled the integration of RL agents in the networking architecture to enforce changes in traffic patterns during network congestions. This paper surveys major RL techniques adopted for efficient TE in SDN. We reviewed the use of RL agents in modelling TE policies for SDNs, with agents' actions on the environment guided by future rewards and a new state. We further looked at the SARL and MARL algorithms the RL agents deploy in forming policies for the environment. The paper finally looked at agents design architecture in SDN and possible research gaps.**

Keywords*—Software defined networking; reinforcement learning; machine learning; traffic engineering

I. Introduction

The emergence of fifth generation (5G) networks has propelled the growth of Internet of Things (IoT) in recent times. IoT is a rapid evolving technology that connects billions of devices to the internet [1]. With 5G, the rapid deployment of new and smart IoT applications are expected to reach 22.3 billion by 2024 and generate about 163 zettabyes (ZB) of data by 2025 [2] [3]. These new and dynamic applications are expected to benefit from the services 5G networks will provide: ultra-reliable and low latency communication (URLLC) [4], enhanced mobile broadband (eMBB) [5] and massive machine type communication, massive MIMO [6].

As shown in Fig. 1 and Fig. 2, the dynamic nature and requirements of IoT devices has necessitated a network deployment shift from the traditional networking architecture which are difficult to configure and manage to a more flexible programmable domain [7]. Software-Defined Networking (SDN) is a new networking paradigm that separates the data plane from the control plane [8] [9]. This separation makes the network more agile with centralized responsibility given to the controller [10]. The controller communicates with the application plane via the northbound APIs and the forwarding devices via the southbound APIs (OpenFlow). The automation and programmability of the SDN architecture helps to configure, secure, and optimize network resources [11] quickly whiles maintaining a good Quality of Service (QoS) [12] and Quality of Experience (QoE) [13].

Traffic Engineering (TE) in SDN involves the analysis of the networks state by the SDN controller to act on flow data through the rapid change in flow table information for forwarding devices [14]. Rerouting flows periodically to balance the loads on the network minimizes congestion and improves the overall network performance. A network experiences two kinds of traffic flows: elephant flows and mice flows [15]. The elephant flows are heavy traffic flows that requires more network resources whiles the rapid aggregation of the mice flows can equally degrade the network. These traffic flows continuously needs dynamic resource allocation for the efficient utilization of scarce network resources through TE.

With the advent of machine learning, port-based [16] and payload-based [17] flow classification techniques have become ineffective due to the dynamic port usage of IoT devices. The negative impact of packet out of order and packet loss in traditional TE techniques even worsens the case for the network operator.

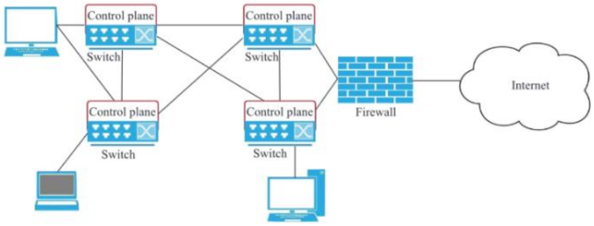


Fig. 1. Traditional Networking Architecture [7].

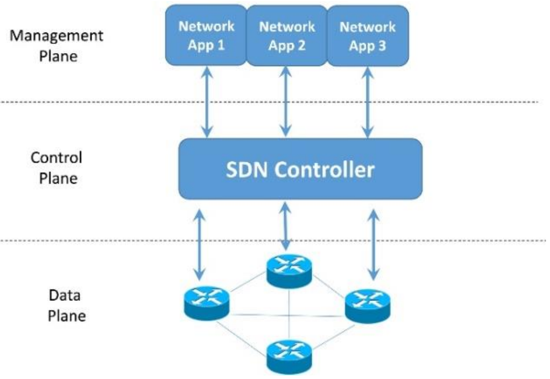


Fig. 2. Software-Defined Networking (SDN) Architecture [7].

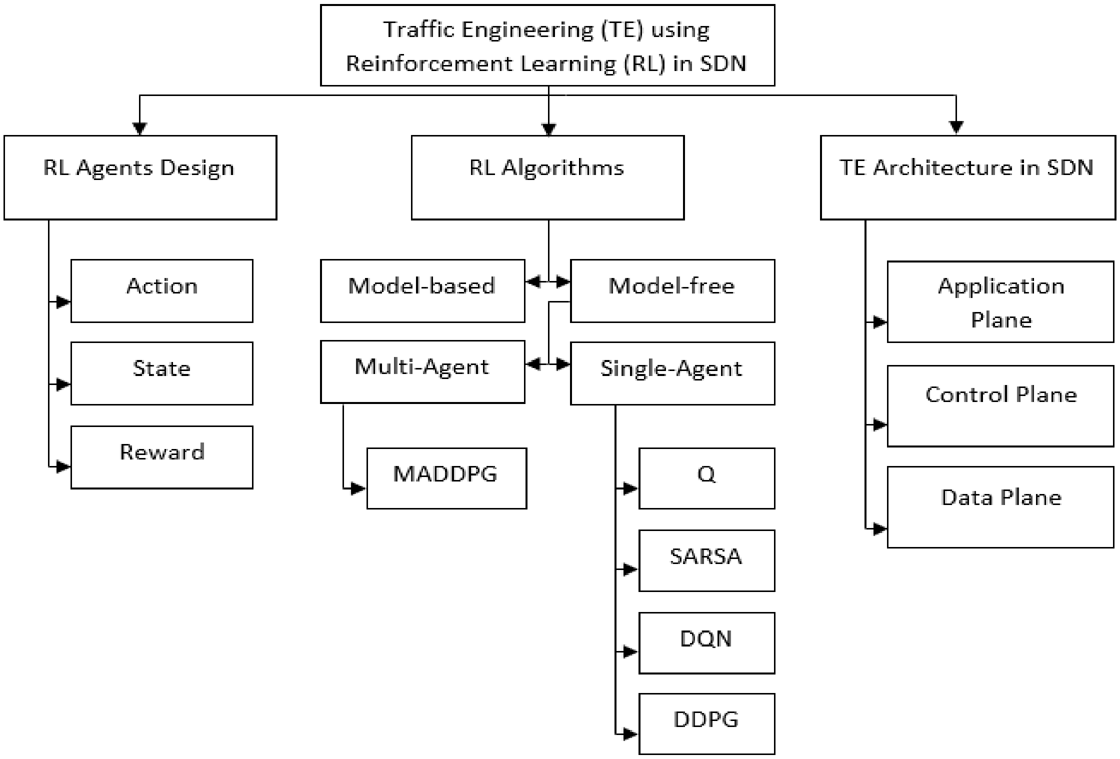


Fig. 3. TE using RL in SDN – Review Outline

Currently for TE, machine learning algorithms are adopted for intelligent flow re-routing with an efficient feature selection criterion [18] [19] in network flow analysis. Deploying these machine learning algorithms in the SDN controller will efficiently allocate network resources and formulate policies for optimal network performance with low overheads.

In this survey as outlined in Fig. 3, we reviewed popular Reinforcement Learning (RL) techniques used in SDN architecture for Traffic Engineering with limitations on parameters chosen and approaches for future research. The rest of the paper is organized as follows: Section II discusses the justification of RL for TE; Section III analysed the TE architecture integration in SDN based on policies and performance. Finally, Section IV looked at the research gaps identified from the survey.

II. Machine Learning with Reinforcement Algorithms

With the advent of Machine Learning [101] where automation modelling using data remains relevant, traditional algorithms [102][103][104] used in solving SDN-IoT related task is unfeasible. In supervised learning [103] agents are trained with a labeled dataset and later tasked to make predictions out of the learned data. Increased complexity in a dynamic environment with new IoT devices and variance in data will negatively affect the accuracy of supervised learning algorithms and predictions. Even worse is the time factor in retraining and relabeling of new data variance in an attempt to still adopt classification algorithms. With unsupervised learning [104] that uses unlabeled dataset, there is no guidance regarding the accuracy of the clustered dataset. Clustering algorithms alone is inefficient in an SDN-IoT environment that requires efficiency in diverse IoT applications. Reinforcement Learning (RL) defines the true automation of agents in an environment [20][105] with rewards as guidance on how well the agent is performing. Though complex, RL agents adapts to changing conditions in the environment by learning to solve tasks through trial-and-error approach. As the episodes progresses, agents adapt to successful actions through exploration and exploitation [106][107] on the stochastic environment. With the recent success of DeepMinds AlphaGo RL agent [108] that defeated the Go champion in 2016, the application dimensions of RL have become enormous. The only way packets can be routed intelligently in a network with varying and emerging IoT devices is to deploy RL agents to learn varying network state patterns with no exclusive data labels but with policies and actions.

III. TE using Reinforecement Learning in sdn

Reinforcement Learning (RL) is an area of machine learning where an agent is modeled to take sequence of actions informed by policies [20]. As shown in Fig. 4, the agent learns in an interactive environment and receives a reward through its actions [21, 22]. The set of actions presents a new state with corresponding reward to the agent. Unlike supervised learning [23] where a set of correct actions are provided as feedback to the agent, RL uses rewards and punishment as signals for positive and negative decisions. The goal is to use trial and error methods in getting positive rewards or build a suitable model that will maximize cumulative rewards for the RL agent.

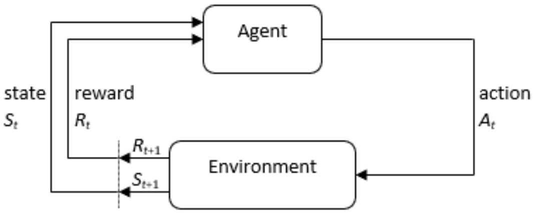


Fig. 4. Reinforcement Learning.

SDN provides centralised control with a unique advantage for intelligent TE framework implementation using RL. Network policies can easily be generated from the centralized control with corresponding TE rules to forwarding devices. With RL, the modelling of the agent's action on the environment with rewards fits into the network architecture of SDN and this expedites network control and management.

A. RL Agents Design

This section details the mathematical modelling of the state space with respect to actions and rewards. Agent design requires the environment to be monitored.

An agent based on the monitored metrics takes actions informed by policy decisions with a new state and a corresponding reward in guiding the next policy.

1) Action-State-Reward: RL agents are implemented in RL frameworks and modelled in SDN to learn critical network packet flow policies and provide routing solutions to forwarding devices. The agent takes an action on the environment and evaluates the actions based on rewards. Using its policy π , the agent performs an action *a,* which alters the environment state *s to s'*[24]. Based on the reward *r*, the agents policy is updated. In arriving at optimal policy, RL agents use Markov Decision Process (MDP) [25] to model actions on the environment with corresponding rewards. MDP is an intuitive and fundamental formalism for decision-theoretic planning (DTP) [26] and RL in stochastic domains. The MDPs have become the de facto standard formalism for learning sequential decision control problems [27].

Algorithm 1 Markov Decision Process (MDP)

```
An MDP is a 5-tuple $(S, A, P, R, \gamma)$, where;
    $S$ is a set of states
    $A$ is a set of actions
    $P\ (s, a, s')$ is the probability that action $a$ in state $s$ at time $t$
    will lead to state $s'$ at time $t + 1$
    $R\ (s, a, s')$ is the immediate reward received after a transition
    from state $s$ to $s'$, due to action $a$
        $\gamma$ is the discounted factor which is used to generate a discounted
                                reward
```

For TE in SDN, RL agents are implemented differently based on the agents policy and the metrics for measuring TE success. The actions of the agents on the environment are rated by the rewards associated with it as the episode progresses.

CFR-RL agent [28]

State Space $s_t = TM_t$ (1)

Action Space $\{0, 1, \dots, (N * (N - 1)) - 1\}$ (2)

Reward Function $r = {}^{1}/_{U}$ (3)

The CFR-RL agent resides in the controller of the SDN architecture. The RL agent uses a traffic matrix that contains the traffic demand of each flow as state. The objective is to avoid packet link congestion. As shown in equations 1 - 3, the CFR-RL agent samples K critical flows for a given stage s_t within N nodes. The CFR-RL agent then reroutes these critical flows and obtains maximum value in link utilization U as reward.

Q-DATA RL agent [29]

State Space $S_i \triangleq \{(f_i, \Delta f_i) : 0 < f_i \leq f_{capi}; -f_{capi} \leq \Delta f_i \leq f_{capi}\}$ (4)

Action Space $A_i \triangleq \{a : a \in \mathcal{F}\}$ (5)

Reward Function $\begin{cases} \frac{\sum_{x=1}^{f_i} \Theta_x}{f_i}, & 0 < f_i < f_{capi}, \\ 0, & f_i = f_{capi}, \end{cases}$ (6)

The Q-DATA agent resides in the application plane of SDN. As shown in equations 4 - 6, the Q-DATA RL agent has a defined state space where f_i is the current total number of flow entries in switch *i*; Δf_i is the number of flow entry changes between two consecutive observations and f_{capi} is the maximum number of flow entries in switch *i*. $(f_i, \Delta f_i)$ represents the state of an SDN switch *i*, as a tuple. For its action space, a represents a traffic flow matching scheme change related to a destination host and $\mathcal{F}$ denotes a list of all feasible match field combinations. With the reward function, f_i is the current total number of flow entries in the switch *i*; and Θ_x is an integer number representing the number of enabled match fields in flow entry *x*. An action has no reward if that action leads to the total number of current flow entries in the SDN switch *i* reaching the limit f_{capi}.

Mu [30]

State Space $s_i = (flow\_freq_i, flow\_recentness_i)$ (7)

Action Space $a_{flow\_freq_i}^{increase} = (flow_{freq_i}{}^{increase}, flow_{recentness_i})$ (8)

Reward Function $r_t = Compare(overhead_{current_{best}}, overhead_t)$ (9)

In [30], the RL agent resides in the controller of the SDN. As shown in equations 7-9, $flow\_freq_i$ represents the frequency of matched flows and $flow\_recentness_i$, an indication of flow duration in the memory of the switch. These are defined for the state space. The action space denotes an increase action on the flow frequency parameters. With the reward function the $overhead_{current_{best}}$ denotes the current best network control overhead obtained. A configuration with less overhead returns a positive reward, 1 to the RL agent otherwise a negative value -1 is returned. If $overhead_{current_{best}}$ and $overhead_t$ are equal, a reward value of 0 is given.

Huang [31]

State Space $s_f = (bandwidth_f, jitter_f, packet\_loss\_rate_f)$ (10)

Action Space

$$a_s = \ action_{bandwidth_f}^{routing\_path} \tag{11}$$

Reward Function

$$r_s = \{MOS_{customer}, QoE_{customer}\} \tag{12}$$

With [31], the objective of the RL agent is to maximize the cumulative QoE of customers by dynamically allocation traffic in a multimedia environment. The RL agent resides in the controller of the SDN architecture. As shown in equations 10 – 12, the state of the environment refers to the state of flows and covers the following metrics: allocated bandwidth, the delay, the jitter and the packet loss rate of flows. The action includes: the path chosen (routing path) and the bandwidth adaptation of flows. The mean opinion score (MOS) [32] used to evaluate the QoE represents the reward function. A multi-layer deep neural network (DNN) is used to map the network and application metrics to the MOS.

Choi [59]

State Space

$$s_i = (sampling\_period_i) \tag{13}$$

Action Space

$$a_i^{increase} = (\,sampling\_period_{increase}) \tag{14}$$

Reward Function

$$r_t = \begin{cases} 1, & hit\_ratio_{t-1} < hit\_ratio_t \\ 0, & hit\_ratio_{t-1} = hit\_ratio_t \\ -1, & hit\_ratio_{t-1} > hit\_ratio_t \end{cases} \tag{15}$$

In [59] RL framework is modelled to minimize the number of overflow occurrences. As shown in equations 13 - 15, the state space represents the size of the sampling period with a unit size of 500 ms. This ranges to 5,000 ms with a total of 10 states. The action space has three options: (i) increase sample period by unit size; (ii) decrease sampling period by unit size; (iii) maintain the sample period. Based on the percentage of table hits, three rewards are given. A reward of 1 is given when the measured hit rate is higher than the hit rate pre-action. If low, a reward of -1 is assigned. A reward of 0 is assigned if there is no change in the hit rate.

Fu [71]

State Space

$$State = \left\{s = \left[FT_{sw_i,t_j}, PS_{pk,sw_i,t_j}\right]\middle|\, i\ \in [1,n], j\ \in [1,m], k\ \in [1,z]\right\} \tag{16}$$

Action Space

$$Action = \{a_{p1}, a_{p2}, \dots\ a_{pk} \dots a_{pN}\} \tag{17}$$

Reward Function

$$\begin{gathered} R_{elephant} = \ \alpha * (1 - PLR) + \ \beta * TP \\ R_{mice} = \ \lambda * (1 - PLR2) + \mu * (1 - DL) \end{gathered} \tag{18}$$

In [71], flow table state and port state are responsible for collecting network statistics. The channels of the network represent the flow table utilization and its respective port rate of switches at current and previous states. For the state space modelling n, m and z respectively identifies the number of switches, moments and ports of a single switch. FT_{sw_i,t_j} represents the flow table utilization rate of switch i at the moment t_j and ranges from 0 to 1. PS_{pk,sw_i,t_j} represents the port rate of port k in switch i at the moment t_j. The action space comprises of $p_1\ to\ p_N$ which indicates all paths in the network, $a_{pk}\ \in \{0,1\}$. If $a_{pk} = 1$, the current flow is assigned to path k else $a_{pk} = 0$. For the reward function, the elephant-flows PLR, represents the average packet loss rate of elephant-flows in the network, TP is the average throughput of elephant-flows after processing. α and β are the weights of the PLR and TP respectively. With the mice-flows, $PLR2$ indicates the average packet loss rate of mice-flows and DL represents the normalized average delay. λ and μ identifies the weight of the $PLR2$ and DL , respectively.

Zhang [86]

State Space

$$s = (nc, src, dst, avail) \tag{19}$$

Action Space

$$a = (a_1,, a_i \dots, a_j, path) \tag{20}$$

Reward Function

$$r = \frac{1}{L}\sum\nolimits_{l=1}^{L}(2\frac{bl}{bw_l} - 1) - \beta\frac{2}{\pi}\arctan(\sigma) + 1 \tag{21}$$

In [86] the state comprises of four components; name of the requested content, source, destination and available link bandwidth.

With the action, a_i denotes the i^{th} destination node split ratio and relates to the content request sent to that destination node using selected transmission links. The reward is meant to improve load balance and throughput. The $\frac{1}{L}\sum_{l=1}^{L}(2\frac{bl}{bw_l} - 1)$ reveals throughput impact in relation to available normalized bandwidth. The - $(\frac{2}{\pi})\arctan(\sigma) + 1$ indicates the load balance with normalized variance of available bandwidth. A value close to 1 signals a preferred action with a reverse value close to -1, a penalty. β = 1 is a factor used to balance the throughput and the load balance.

B. RL Algorithms

In this section, we reviewed the algorithms the RL agents use to formulate policies that informs the action taken by the agent on the environment as the episode progresses. For effective TE and policy enforcement on the environment, RL agents learns to take the best actions for traffic optimization in respect to cumulative future rewards. RL algorithms are distinguished into two main classes: the model-free (direct) and model-based (indirect) methods [33, 34, 35].

1) Model-based RL methods: Model-based RL algorithms utilizes a model when the RL agent interacts with the environment. The model keeps track of transition dynamics of the network to derive optimal actions and rewards [35]. When the model is referenced, the RL agent can make predictions about the next state and reward before an action is taken. Model-based RL methods are data efficient but struggles to achieve asymptotic performance for real-world applications [36]. For model-based RL methods, the interaction between the RL agent and the environment is modeled as a discrete-time Markov Decision Process (MDP) $\mathcal{M}$ and defined by the tuple [36]:

$(S, A, p, r, \gamma, p_0, H)$. Where S is the set of states, A the action space, $p(s_{t+1}|\, s_t, a_t)$ the transition distribution, $r\ :\ S \times A\ \rightarrow \mathbb{R}$ as a reward function, $p_0\ : S \rightarrow \mathbb{R}_+$ represents the initial state distribution, γ the discount factor, and H the horizon of the process. The return function is defined as the sum of rewards $r\,(s_t, a_t)$ along a trajectory τ: = $(s_0, a_0, \dots, s_{H-1}, a_{H-1}, s_H\,)$. The goal of the reinforcement learning is to find a policy π: $S \times A\ \rightarrow \mathbb{R}^+$ that maximizes the expected return. The model-based learns the transition distribution from the observed transitions using parametric approximator $\acute{p}_{\o}(s'|s,a)$. The parameter $\o$ of the dynamic model are optimized to maximize the log-likelihood of the state transition distribution. Though model-based RL methods are data efficient, they have high computational complexity and the degree of potential error in maximizing a reward is compounded..

2) Model-free RL methods: Model-free RL algorithms do not utilize a model and thus the rewards and the optimal actions are derived through trial-and-error approach with the environment [37]. These set of algorithms operate over an unordered list of actions, with a positive or negative reward value. The RL agents that utilizes model-free algorithms increases the value associated with a positive action which helps the agent to learn from direct experience. Agents in model-free RL are represented with policy optimization and Q-learning approaches [38]. With policy optimization, the agents learns directly the policy function that maps state to action without a value function. The Q-learning approach learns the action-value function $Q(s,a)$; how good to take an action at a particular state. A scalar value is assigned over an action *a*, given the state *s* [39]. Model-free RL methods have low computational complexity but more data dependent. For TE, model-free RL methods are frequently used for RL agent sequencing and to implement policies on the environment.

3) Single Agent Reinforcement Learning (SARL): In a SARL, there is only one agent that interacts with the environment to maximize rewards. The SARL implementation is suitable for simple network management with slower convergence and learning experience. The SARL implemented algorithms are either value-based, policy-based or both [48]. As shown in Fig. 5, the SARL through the SDN controller collects information from the environment through the forwarding devices.

The agent upon receiving the state information performs a set of actions on the environment through the SDN controller. These actions are guided by policy algorithms. The episode results in a new state and rewards.

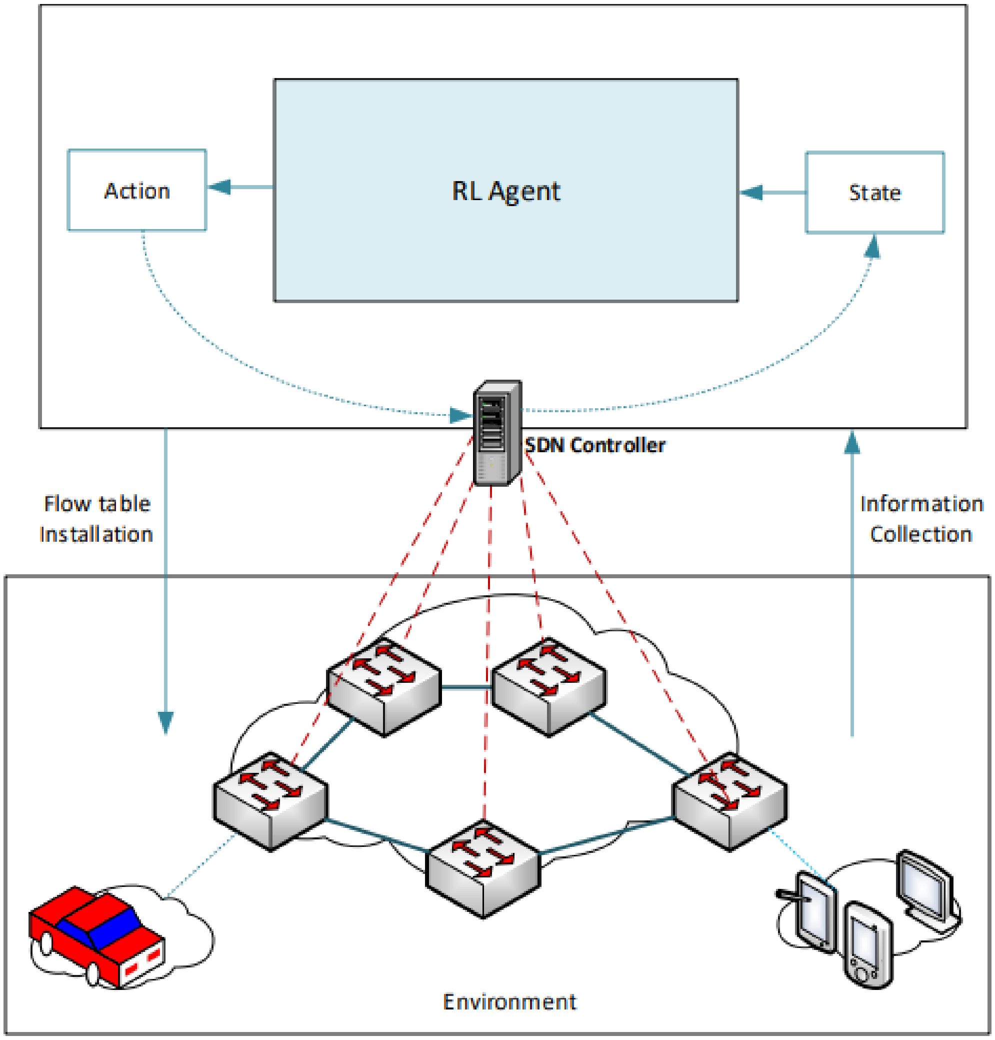


Fig. 5. SARL.

a) Q-learning Algorithm: Q-learning [40] is an off-policy, value-based algorithm that takes a random actions based on the $\epsilon - greedy$ policy, where the probability of a random decision is determined by the value of epsilon ϵ. During the learning phase, the Q-learning agent initializes the Q-table for all state-action pairs and updates it using:

$$\begin{aligned}\mathcal{Q}\_(t+1)\,(s\_t, a\_t\,) = {} & \mathcal{Q}(s\_t, a\_t\,) + \alpha[\,\mathcal{R}\_i\,(s\_t, a\_t\,) \\ & + \gamma max \mathcal{Q}\_t\,(s\_(t+1), a\,) \\ & - \mathcal{Q}\_t\,(s\_t, a\_t\,)]\end{aligned} \tag{22}$$

The Q-learning agent generates the optimal policy $\pi^*(s)$ for a state s representing an action a that needs to be taken to maximize the value of the $\mathcal{Q}_*(s,a)$ function, $\pi^*(s) = arg\,max_a\, \mathcal{Q}_*(s,a)$.

Algorithm 2 Q-learning [40]

```
Inputs: $\mathcal{F}$; for a state-action pair $(s,a)\ \forall s \in S_i, a \in \mathcal{A}_i$,
initialize a $Q$-table entry arbitrarily; initialize values of $\alpha$, $\gamma$ and $\epsilon$,
respectively.
loop
    Current state $s_t$.
    Executive action $a_t$ according to an exploratory policy ($\epsilon$).
    Obtain a new state $s_{t+1}$ and an immediate reward $\mathcal{R}_i$.
    Update the $Q$-table entry for $\mathcal{Q}(s_t, a_t)$.
    Update $s_t \longleftarrow s_{t+1}$.
end loop
Outputs $\pi^*(s) = \arg max_a\, \mathcal{Q}_*(s,a)$.
```

Phan *et al.* [29] proposed the Q-learning algorithm in maximizing traffic flow monitoring in SDN switches. It embeds a Support Vector Machine (SVM) [49] algorithm in the application plane of the SDN architecture to predict the performance degradation of the switches as the episode progresses. To reduce the long-term control plane overhead capacity limitation of Ternary Content Addressable Memory (TCAM) in OpenFlow switches, [30] proposed a Q-learning algorithm for SDN flow entry management. The framework determines the forwarding rules that remains in the flow table of the SDN switches and those processed by the controller in case of a table-miss on the switches. In [50] a Q-learning algorithm is proposed to reduce the latencies and improve the bandwidth utilization in the UbuntuNet Alliance National Research and Education Network (NRENs) SDN switches. The proposed framework adapts forwarding devices by learning from experience using multipath propagation. In dealing with bandwidth overhead caused by Dijkstra's shortest path first module [51] in an OpenDayLight (ODL) architecture meant for efficient packets delivery, [52] proposed a congestion prevention mechanism using Q-learning in SDN. With [52], the set threshold values are defined in SDN controllers to enable threshold bandwidth detections. The optimal path chosen is delivered to the OpenVSwtiches (OVS) after Q-routing by the controller during network congestion. To balance the network load in SDN, [53] proposed a Q-learning approach to reduce the number of unsatisfied users in a 5G network architecture. The researchers used a flow admission control technique with a fairness function to enhance the per-flow resource allocation in the network. In [54] a load balancing architecture is proposed for SDN networks that uses supervised Bayesian Network (BN) to solve the problem of Q value local maximum [55] in a Q-learning RL algorithm. The combination of the BN in Q-learning helps the controller select the most optimal strategy for network load balancing during congestion. For TE load balancing optimization in master controllers, [56] proposed a dynamic switch migration algorithm to slave controllers using Q-learning in SDN. The switch migration problem (SMP) is modeled and used to redefine the Q-learning parameters. The Q-learning is then used to learn the current status of SDN to select the best switches for load migration. For an efficient path selection technique in load balancing, [57] proposed a Q-learning algorithm for path selection and flow forecasting [58]. It has an integrated centre that uses Deep Neural Networks (DNNs) to process uncertain network traffic and uses Q-learning to resolve the optimal path based on the results of the DNN. The DNN path selection are obtained from the bandwidth utilization ratio, packet loss rate and transmission latency which forms the inputs to the DNN. The output which is fed into the Q-learning is derived from the corresponding link score. For timely eviction of inactive flow entries and to avoid overflows in the memory of SDN switches, [59] proposed a Q-learning User Datagram Protocol (UDP) [60] flow eviction strategy for UDP flows. The Q-learning is used to dynamically resize the sampling period as the most critical parameter in the RL architecture. This advertently maximizes the table hit rates of the UDP flows in the SDN.

b) State-Action-Reward-State-Action (SARSA): SARSA [61] is an on-policy algorithm which uses the action performed by the current policy to learn the Q-value. As shown in Eq. 23 [61] and Eq. 24 [40], the update rule for SARSA varies from that of Q-learning algorithm in the execution of actions. In SARSA, update estimates are based on the same action taken whiles in Q-learning, the update estimates are based on the number of possible actions that maximizes the post-state Q function, $\mathcal{Q}(s_{t+1}, a')$.

$$\begin{aligned}\mathcal{Q}(s_t, a_t) \longleftarrow {} & \mathcal{Q}(s_t, a_t) + \alpha[r_{t+1} + \gamma \mathcal{Q}(s_{t+1}, a_{t+1}) \\ & - \mathcal{Q}(s_t, a_t)]\end{aligned} \tag{23}$$

$$\begin{aligned}\mathcal{Q}(s_t, a_t) \longleftarrow {} & \mathcal{Q}(s_t, a_t) + \alpha[r_{t+1} + \gamma max \mathcal{Q}(s_{t+1}, a') \\ & - \mathcal{Q}(s_t, a_t)]\end{aligned} \tag{24}$$

For dynamic load balancing in multiple controllers due to switch migration conflicts, [62] proposed a SARSA-Bayesian RL algorithm for a multi-controller cluster design in SDN. With knowledge of the real-time load and controller's communication consumption, a request response model using the Bayesian [63] algorithm is combined with the SARSA RL mini-framework in a switch migration technique to the lighter controller. For a multi-layer hierarchical SDN to be effective in handling traffic, [64] proposed the SARSA algorithm for QoS provisioning. With each pre-flow, the switch contacts the SDN controller. The controller uses the SARSA algorithm to implicitly detect the QoS requirement of each flow and computes the corresponding optimum traffic path based on the needed QoS requirement. The next hop in the switch forms the basis for the next action from the source to the destination switch. To convey a massive IoT data through a limited bandwidth efficiently [65] proposed a SARSA algorithm for resource allocation through cognitive communications in the SDN-enabled environment. The SARSA agent communication is modelled with a buffer metric that manages the aggregator's output queue transmissions and reflects dynamically in the IoT data demands. This modification targeted at Publish/Subscribe (Pub/Sub) paradigms preserves the Pub/Sub bandwidth with less computational resources. In order to adapt VS-routing [67] optimization to SDN networks, [66] proposed a network hop count technique to improve the reward function of SARSA algorithm. The VS-routing introduces an $\varepsilon - Greedy$ function in the network hop count which is calculated to select the optimal route and avoid the long package queue of network links in the SDN architecture.

c) Deep Q-Network Algorithm (DQN): With the advent of Artificial Neural Networks, (ANNs) a class of RL agents that utilizes Q-learning with Deep Neural Networks (DNNs) [41] in discrete domains for TE has emerged. DQN uses feedforward neural networks with three components: (i) Neurons that are interconnected using direct links to form a network, (ii) Weights associated with each connection, (iii) Layers consisting of a number of neurons and multiple hidden layers.

Algorithm 3 Deep Q-Network Algorithm [41]

```
Pre-condition:
    Initialize experience memory $M$
    Initialize action-value pair $Q$ with random weights
    Initialize state $s_t$
    Initialize goal $\mu$
Procedure:
    improvement = 0
    repeat
        for (step = 0; step < learning_iteration; step++)
            Get action $a_t$ from $s_t$ using $\epsilon - greedy$ policy
            Get parameter $param_t$ from $s_t$ using $\epsilon - greedy$ policy
            $\epsilon = \epsilon -$ (step / learning_iteration)* $\epsilon$
            Take action $a_t$ on $param_t$ and receive reward $r$, control
            overhead $c_t$
            Observe new state $s_{t+1}$
            Store experienced memory $(s, a, r, s_{t+1})$ into $M$
            Sample $n$ random transitions $(s', a', r', s'')$ from $M$
            Update $transition\ tt \longleftarrow r' + \gamma * \max(s'' - a'')$
            Update the $param_t$ of $\theta_i$
            Train the $Q$ network using $loss = (tt - Q(s', a'))^2$
            $improvement =$ get_improvement ( $best_t, worst_t$)
        end for
    until improvement > $\mu$
```

The DQN has an experience memory for storing experienced transitions (s, a, r, s') unlike the Q-learning. The discount factor γ and the state of the Q-Network in the ith iteration, θ_i are used to update the experienced transitions with a training principle using a loss function. The $\epsilon - greedy$ policy helps select the action based on the highest Q-value associated with that action after the training. For the RL agent to choose random actions, the ϵ value is set to 1 at the start of the learning process but decreased over time in order to maintain a fixed exploration rate. The DQN keeps track of the chosen parameter corresponding to the Q-value of each action with a terminal, μ..

In [28], the DQN is used to learn a policy to select critical flows based on a given traffic matrix. The Critical Flow Re-routing-Reinforcement Learning (CFR-RL) agent then reroutes the selected flows for a balanced link utilization using Linear Programming (LP). For an efficient SDN flow entry level management with a TCAM enabled OpenFlow switches [30] proposed a DQN algorithm to obtain the flow entries and reduce the long-term control plane overhead between the SDN switch and the controller. The DQN agent automatically finds the values of decision parameters that effectively selects the candidates rule in the switch's flow table for a higher table-hit rate. For flexible network management through TE, [68] proposed a DQN based dynamic controller placement caused by flow fluctuations in SDN. The D4CPP agent in [68] integrates historical network data into the controller deployment. The real-time switch-controller mapping decisions is then triggered with inherent adaptation to the dynamic flow fluctuations in the network. For effective TE among distributed controllers in SDN, [69] proposed a DQN based switch and controller selection scheme for switch migration and switch-aware reinforcement learning-based load balancing (SAR-LB). The SAR-LB adopts the utilization ratio of diverse resource types in both controllers and switches as inputs to the neural network for a dynamic load distribution among the controllers in the network. Yao *et al.* [70] proposed a DQN-based energy-efficient routing solution for full load software-defined data centers. The optimization is for the DQN to find energy-efficient routing paths and load-balancing between controllers in reducing energy consumption in the network. The enhanced DQN-based energy-efficient routing (DQN-EER) algorithm learns directly from experience. At the same coordinated time, it selects the arriving flows and the energy-saving control path at the in-band control mode whiles detecting the energy-saving routes for the data center. Fu *et al.* [71] proposed the detection of mice and elephant flows in an SDN-enabled data center using two DQNs. The DQNs are built and trained to generate efficient routing strategies using convolutional neural networks (CNNs) [72][73] to avoid possible network congestion. For efficient latency management in SDN, [74] proposed a DQN agent that inherently predicts optimal traffic paths and future traffic demands through the SDN switches. Whiles formulating the flow rules placement policy as an Integer Linear Program (ILP), [74] used a traffic prediction module with a long short-term memory (LSTM) [75][76] neural networks algorithm. To further minimize network delay, a proposed DQN-TP (traffic prediction)-based heuristics defect-tolerant routing (DTR) [77] algorithm interacts dynamically with the DQN agent module in the controller of the SDN architecture.

d) Deep Deterministic Policy Gradient (DDPG): In combining policy gradient and Q-learning, Deep Deterministic Policy Gradient (DDPG) [42][79] is used as an off-policy, actor-critic technique consisting of two modes; actor and critic as shown in Fig. 6. The actor is the policy network and the critic, the Q-value for training the actor network.

Algorithm 4 Deep Deterministic Policy Gradient (DDPG) Algorithm [42]

```
Input: Initial policy parameters $\theta$, Q-function parameters $\phi$, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\theta_{targ} \leftarrow \theta, \phi_{targ} \leftarrow \phi$
repeat
    Observe state $s$ and select action $a = \text{clip}\ (\mu_\theta(s) + \epsilon, a_{Low}, a_{High})$, where $\epsilon \sim \mathcal{N}$
    Executive $a$ in the environment
    Observe next state $s'$, reward $r$, and done signal $d$ to indicate whether $s'$ is terminal
    Store $(s, a, r, s', d)$ in replay buffer $\mathcal{D}$
    If $s'$ is terminal, reset environment state
    if it's time to update then
        for however many updates do
            Randomly sample a batch of transitions, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
            Compute targets
                $y(r, s', d) = r + \gamma(1 - d) Q_{\phi\text{targ}}\ (s', \mu_{\theta\text{targ}}(s'))$
            Update Q-function by one step of gradient descent using
                $\nabla_\phi \frac{1}{|B|} \sum_{(s,a,r,s',d)\in B} (Q_\phi\ (s, a) - y\ (r\ s', d))^{2}$
            Update policy by one step of gradient ascent using
                $\nabla_\phi \frac{1}{|B|} \sum_{s\in B} (Q_\phi\ (s, \mu_\theta(s))$
            Update target networks with
                $\theta_{\text{targ}} \leftarrow \rho\phi_{\text{targ}} + (1 - \rho)\phi$
                $\theta_{\text{targ}} \leftarrow \rho\theta_{\text{targ}} + (1 - \rho)\theta$
        end for
    end if
until convergence
```

DDPG uses DQNs replay buffer to gather offline unrelated experiences obtained by the agents whiles performing actions on the environment. At each time step, the actor and the critic are updated by uniformly sampling a minibatch from the replay buffer. DDPG uses soft target, θ_{targ} updates rather than directly copying the weights to the target network. DDPG further utilizes batch normalization which helps normalize each dimension across the samples in a mini-batch to have unit mean and variance. DDPG algorithm is suitable for continuous action space and state representations.

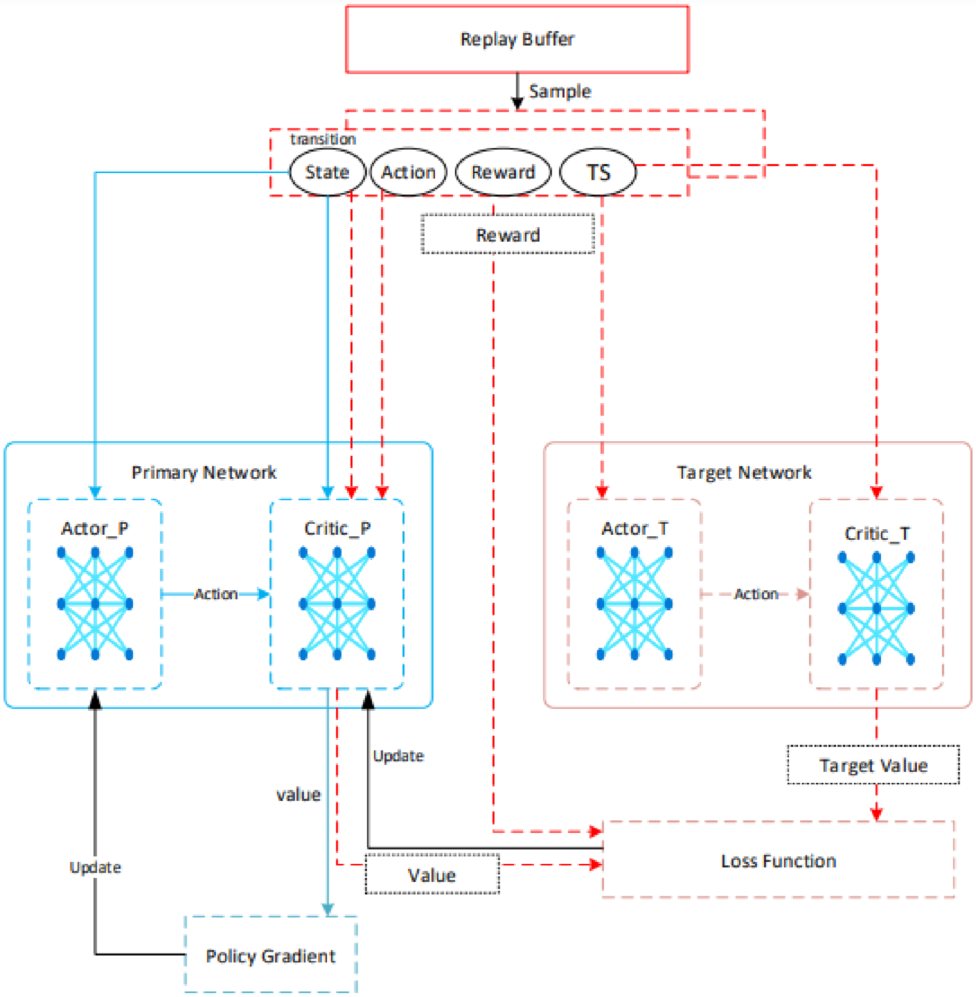


Fig. 6. Actor-Critic Model of DDPG [31].

In [31], the DDPG algorithm is used for multimedia traffic control with the objective of maximizing cumulative Quality of Experience (QoE) for network users. The DDPG agent enforces bandwidth adaptation and path chosen actions for all multimedia flows in the SDN-enabled environment. To maximize the QoE for users, a multi-layer deep neural network is used to map the network and application metrics to the mean opinion score (MOS) [78] obtained from users. Stampa *et al.* [80] proposed a DDPG agent for dynamic routing in SDN. The architecture embeds an integrated fully-connected feed-forward neural network (FFNN) [81] in the framework to re-define the feature extraction of the actor-critic network. To improve the learning rate of DDPG for effective routing optimization, [82] proposed a dynamic planning of the experience pool capacity with respect to the current iteration number. This accelerates the growth rate of the previous pool by reducing its capacity in affecting subsequent learning rates. In [83] a deep-reinforcement-learning-based quality-of-service (QoS)-aware secure routing protocol (DQSP) is proposed using DDPG algorithm. The DQSP adds an intelligent layer above the control layer which generates the routing policy and evaluates the network performance through the rewards obtained by the DDPG policy. The DQSP protocol guards against gray hole attack [84] and DDoS [85] whiles ensuring an efficient routing planning through the environment-aware module of the control layer. Zhang *et al*. [86] proposed a DDPG-based intelligent content-aware TE (iTE) which leverages on information centric networking (ICN) [87] to optimize traffic distribution in SDN. The DDPG agent together with other TE algorithms are embedded in a parallel decision-making (PDM) module in the controller. This module receives the cache information and the link bandwidth from the switches to activate and update its neural networks with a reward feedback. In [88] a DDPG-based network scheduler for deadline-specific SDN heterogenous networks is proposed. The DDPG agent receives a deadline-ware data transfers from the SDN switches and schedules the flows by initializing a pacing rate at the source of the deadline flows. The actor-critic model in the DDPG agent handles larger and a more generalized scheduling problem that maximizes and assigns the aggregated utility value to each flow if the deadline is met. For intelligent routing in software-defined data-centers (SD-DCN), [89] proposed a deep reinforcement learning based routing (DRL-R) consisting of DDPG-DQN agent to perform a reasonable routing adapted to the network state. DRL-R agent efficiently allocates cache and bandwidth in the network to improve routing performance by delay reduction. This is done through the quantification of the overall contribution score in the network and a change in the routing metric from a single link state to the resource-combined state.

4) Multi-Agent Reinforcement Learning (MARL): In MARL systems, multiple agents collectively learn and collaborate in a deterministic or a stochastic environment [90, 91, 92]. Multi-agent systems are seen in domain applications including: network resource management, computer games, distributed networking, cloud computing and intrusion detection systems. Experience sharing and faster convergence has necessitated a shift in research direction from SARL to MARL in recent times. With a coordinated policy, multi-agents learn and optimize towards an accumulated global reward [93, 94] in the network framework. As a result, the dynamics in state transitions in MARL are dependent on the joint action of all active agents as shown in Fig. 7.

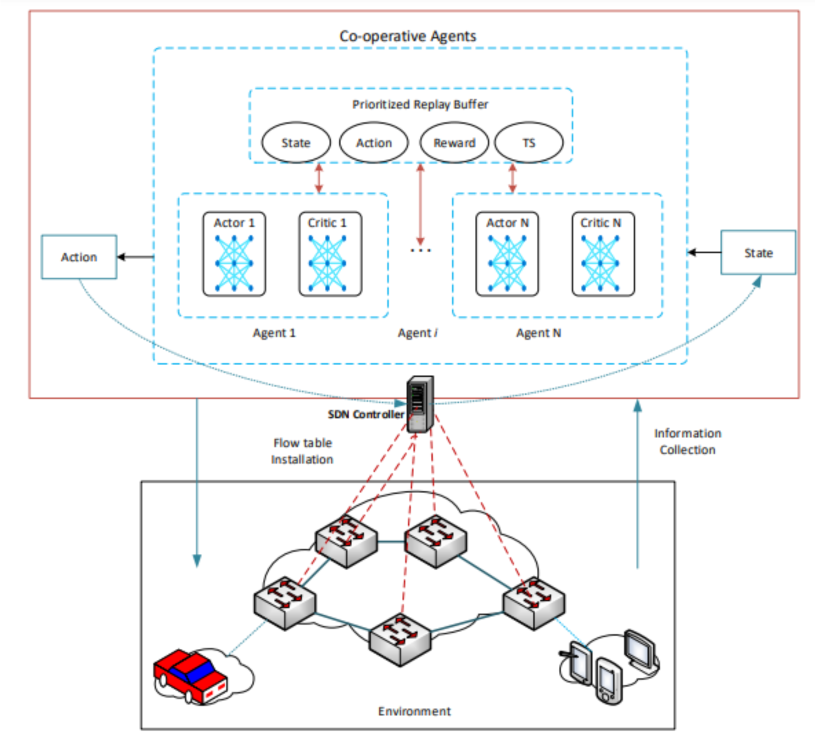


Fig. 7. MARL

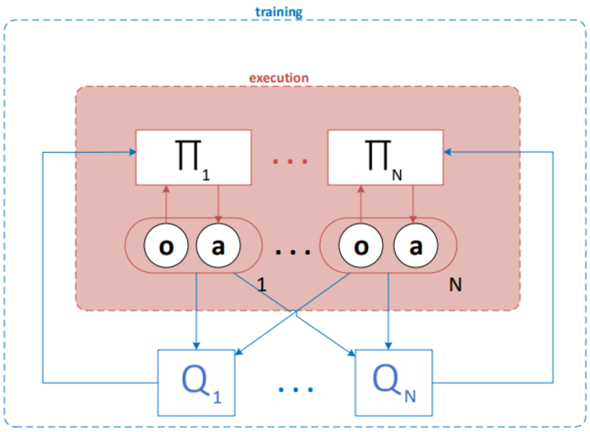


Fig. 8. MADDPG.

a) Multi-Agent Deep Deterministic Policy Gradient (MADDPG): MADDPG [95, 96, 97] is an actor-critic multi-agent extension of DDPG where the critic network is augmented with information from other agents in a decentralized execution. In MADDPG actor-critic architecture, each agent has its own actor and critic network. The critic network of each agent has full visibility of the actions and observation of other agents.

The actor network on the other hand only executes the action for its local agent given the state. In Fig. 8, the actor π_n takes an observation, o as state to give an action, a whiles the critic network, Q_n takes an observation and the action of the actor, to train the actor. The critic has dependent view from other critic networks whiles training the actor network.

Algorithm 5 Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [97]

```
for episode = 1 to $M$ do
    Initialize a random process $\mathcal{N}$ for action exploration
    Receive initial state $x$
    for $t = 1$ to max-episode-length do
        for each agent $i$, select action $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_t\ w.r.t.$ the current policy and exploration
        Executive actions $a = (a_1, \ldots., a_N$ and observe reward $r$ and new state $x'$
        Store $x, a, r, x'$ in replay buffer $\mathcal{D}$
        $x \longleftarrow x'$
        for agent $i = 1$ to $N$ do
            Sample a random minibatch of $S$ samples $x^j, a^j, r^j, x'^j)$ from $\mathcal{D}$
            Set $y^j = r_i^j + \gamma Q_i^{\mu'}\left(x'^j, a'_1, \ldots., a'_N\right)|a'_k = \mu'_k(o_k^j)$
            Update critic by minimizing the loss $\mathcal{L}(\theta_i) = \frac{1}{S}\sum_j(y^j - Q_i^{\mu}(x^j, a_1^j, \ldots., a_N^j))^2$
            Update actor using the sampled policy gradient:
                $\nabla_{\theta i} J \approx \frac{1}{S}\sum_j \nabla_{\theta i}\mu_i\left(o_i^j\right)\nabla a_i Q_i^{\mu}(x^j, a_1^j, \ldots., a_N^j)|a_i = \mu_i(o_i^s)$
        end for
        Update target network parameters for each agent $i$
            $\theta'_i \longleftarrow \tau\theta_i + (1 - \tau)\theta'_i$
    end for
end for
```

In [98], a MADDPG-based traffic control and multi-channel reassignment (TCCA-MADDPG) algorithm is proposed for the core backbone network in SDN-IoT. The TCCA-MADDPG algorithm reduces the channel interference between links by considering the policies of other neighbouring agents using a cooperative multi-agent strategy. To maximize network throughput and minimize packet loss rate and time delay, the TCCA-MADDPG uses a joint traffic control mechanism modelled with a partially observable markov decision process (POMDP) to optimize traffic performance. Yuan *et al.*, [99] proposed a dynamic controller assignment using MADDPG for effective TE in Software Defined Internet of Vehicles (SD-IoV) [100]. For controllers to make local decision in coordination with neighboring controllers, a real-time distributed cooperative assignment approach is used via the actor-critic model of the MADDPG. To get a faster MARL global convergence whiles minimizing delay, a centralized training approach using global information to attain optimal local assignment is adopted in the model development.

C. TE Architecture in SDN

In this section, we looked at the design placement of the RL agents in the SDN architecture and the communication principles adopted with the controller. The architecture of RL systems varies based on the RL agent policy algorithms, the actions selected and the environment. The agent frameworks are designed to enhance positive rewards and proactively prevent network performance degradation through forwarding devices. Different components of the RL agents design in SDN are situated in the application plane, control plane and the data plane.

1) RL agent in control plane: For easier policy formulation and faster communication between the controller and RL agent, most TE SDN designs [28][30][31][52][54][57][70][86][89] situate the RL agent in the control plane of the SDN architecture. In [28], the CFR-RL agent resides in the controller and uses a neural network trained with reinforcement algorithm [43] to map a traffic matrix to a combination of critical flows. After training, the CFR-RL applies the critical flow selection policy to each real time traffic matrix provided by the controller. The SDN controller then reroutes the selected critical flows by installing and updating flow entries of the switches whiles the remaining flows continue the normal route using Equal-Cost Multi-Path (ECMP) [44] TE technique by default. In [30], the RL agent is deployed in the controller and utilizes the flow match frequency and the flow duration to determine the flow entries that should be kept on the switch. To maximize the long term reward, the RL agent lowers the configuration overhead and the number of table-miss events. To achieve the expected reward, the RL agent splits the pool of flow entries into two parts: the local switch entries and the remote controller entries. This will reduce the control plane overhead given the Ternary Content-Addressable Memory (TCAM) [45] size of the SDN switches. With [31] the RL agent is the controller and serves as the centralized control to collect stats, make decision and take actions. The state reflects the situation in the environment and covers metrics: allocation of bandwidth, delay, jitter and the packet loss rate of flows. The action involves the path chosen and the bandwidth adaption for multimedia flows. The reward is the QoE received from the environment. To evaluate the QoE, the multi-layer deep neural network is used to map the network and application metrics to MOS [46]. [52] also proposed the controller is the RL agent and programmed with the Q-learning algorithm to detect network congestion and find optimal path to be delivered to the OpenVSwitch (OVS). In [57] the control layer has an intelligent center connected to the SDN controller. For efficient load balancing, the intelligent center uses the Q-learning algorithm to find optimal paths and returns aggregated path routing decisions to the controller. The DQN-EER architecture [70] has the RL agent programmed in the SDN controller using the DQN algorithm. The DQN is modified with deep convolutional neural networks (CNNs), empirical replay to train the agent and independent target networks to train the primary critic network. In [86] the intelligent content-aware traffic engineering (iTE) RL agent is deployed in the controller of the SDN architecture. It received cache information from the ICN-enabled switches and uses parallel execution module embedded with multiple DRL-based TE algorithms to determine the best routing paths for the flows in the network.

2) RL agent in Application Plane: For easier system failure checks in SDN, [29] [71][83] TE frameworks situate the RL agent in the application plane. The Q-DATA [29] framework architecture has a built-in forwarding application located in the control plane and a Q-DATA application residing in the SDN application plane. Initially, the built-in forwarding application module is instructed by the Q-DATA application through a REST API to apply the Full Matching Scheme (FMS) strategy at the switches. The Q-DATA application has a statistics collector module which periodically collects raw information about traffic flows at the SDN switches from the SDN controller. The statistics is then forwarded to a statistics extractor and distributor module for extraction and distribution to other modules. The SVM based performance degradation prediction module anticipates the performance degradation of the SDN switches before it occurs and provides the prediction results to the Q-learning based traffic flow matching policy creation module and the MAC matching only scheme control module. The MAC matching only scheme control module monitors and checks conditions for a traffic flow matching scheme change to FMS in the SDN switches. In [71] the AI Plane is used as the Application Plane in the SDN architecture. The RL agent is embedded in the AI Plane and uses the DQN to learn the best optimal routing paths for the mice and elephant flows by obtaining the flow type, network state information and network performance evaluation from the control plane of the SDN architecture. In [83], the DQSP architecture has an agent layer that is embedded in the application layer of the SDN architecture. The DQSP agent through the controller is aware of the underlying network environment and generates routing policies for the controller to executive. It receives the reward evaluation and adjusts policy parameters until optimal routing strategy is achieved.

TABLE I. TE IN SDN USING RL – SUMMARY OF FINDINGS

TE in SDN	Agent Algorithm	Main Contribution	MDP	Limitations	Plane
[29]	Q-learning, Support Vector Machine	The authors proposed an enhanced traffic flow monitoring in SDN using Q-learning and Support Vector Supervised Machine Learning Algorithm	Yes	The statistics tracker should have factored in control link and data link capacity utilization of the SD Networks	Application
[30]	Q-learning, Deep Q-Network (DQN)	The authors addressed the TCAM capacity issue in OpenFlow switches by determining which forwarding rules remains in the flow table and those processed by the SDN controller	Yes	The $\epsilon - greedy$ policy should have given more value for exploration to balance the dynamics of the action taken.	Controller
[50]	Q-learning	The authors improved bandwidth utilization and reduced flow latencies – NRENs case study network	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[52]	Q-learning	The authors addressed network congestion in SDN by reselecting flow paths and changing flow table using predefined threshold	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Controller

[53]	Q-learning	The authors introduced fairness function in SDN for load-balancing in peak traffic conditions	Yes	One type of user that should not be ignored is a compromised user with network intrusions.	Not stated
[54]	Q-learning, Bayesian Network	The authors used the Bayesian network to predict the degree of congestion and Q-learning for optimal action decision in SDN load-balancing framework	Yes	The rate of packet-in messages from the switches is enough parameter to predict the load congestion to the controller. Using the Bayesian Network will impede the idea of Reinforcement Learning	Controller
[56]	Q-learning	The authors proposed a dynamic switch migration algorithm with Q-learning in scaling the load on SDN controllers	Yes	No reward graph per episode to define the training and validation accuracy of the agent.	Not stated
[57]	Q-learning	The authors used an integrated DNN in Q-learning for load-balancing in SDN through flow forecasting	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Controller
[59]	Q-learning	The authors maximized table hit rates in a UDP flow entry eviction strategy in SDN by dynamically resizing sampling periods of critical parameters.	Yes	The scope of the state space definition is limited. Aside the size of the sampling period, the state of flows in the network will be an added metrics since UDP operates at the transport layer.	Not stated
[62]	SARSA, Bayesian Network	The authors proposed a switch migration prediction method based on Bayesian network and used with SARSA algorithm for overload-lighter load controller migration.	Yes	Comparing the modified SARSA algorithm to Q-learning in the research will have given a more comparative insight into the results of the research.	Not stated
[64]	SARSA	The authors proposed a QoS-aware adaptive routing scheme using SARSA to provide fast convergence in QoS provisioning in SDN	Yes	The reward function of the MDP is not well defined. Secondary, comparing the results with other known algorithms will have given more credence to the α, γ values	Not stated
[65]	SARSA	The authors proposed a resource allocation technique in massive IoT through cognitive communication in SDN-enabled environment	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[66]	SARSA	The authors proposed a network hop count technique in SDN to improve VS-routing through $\varepsilon - Greedy$ function	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[68]	DQN	The authors proposed a flexible network management through dynamic controller placement technique in SDN	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[69]	DQN	The authors used a DQN based switch and controller selection scheme for switch migration in distributed SDN controllers	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[70]	DQN	The authors used DQN to find energy-efficient routing paths and load-balancing between SDN controllers	Yes	Though energy-saving and load balancing are metrices defined in this research, the extent of a controller's ability to balance the load can be added to the reward functionality.	Controller
[71]	DQN	The authors used two DQN agents to detect mice and elephant flows in an SDN-enabled data center	Yes	A comparative analysis using packet-in and packet-out messages in defining the state-action-reward pair will have added higher scope to the research.	Application
[74]	DQN	The authors proposed a DQN agent that predicts optimal traffic paths and future traffic demands using LSTM neural networks.	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[31]	DDPG	The authors proposed an SDN architecture to maximize QoE using DDPG agent to enforce bandwidth adaption and path chosen for all multimedia flows	Yes	There is little mathematical modelling of the DDPG algorithm used in this research. The pseudocode is not stated mathematically for this research. The parameters for simulation set up was not well defined in this	Controller

				research	
[80]	DDPG	The authors adopted the DDPG agent for dynamic routing in feature extraction with FFNN in the actor-critic network of the agent.	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[82]	DDPG	The authors proposed a DDPG-EREP algorithm with dynamic planning of the experience pool capacity using the current iteration number of the sampling size	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Not stated
[83]	DDPG	The authors proposed a DQSP using DDPG algorithm with added intelligent layer above the control layer for routing policy optimization in SDN	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Application
[86]	DDPG	The authors proposed an iTE which leverages on ICN to optimize traffic distribution in SDN through the PDM module in the controller	Yes	The action space definition should have included the flow path selection procedure aside the split ratio for the i-th destination node.	Controller
[88]	DDPG	The authors used a DDPG agent to receive a deadline-aware data transfers from SDN switches and schedules subsequent flows by initiating a pacing rate at the source of the flows	Yes	This research can be extended to multi-path routing using AOMDV protocol	No stated
[89]	DQN, DDPG	The authors proposed a DRL-R based on DDPG-DQN agent to allocate cache and bandwidth in the SDN to improve routing performance	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	Controller
[98]	MADDPG	The authors proposed TCCA-MADDPG algorithm to reduce the channel interference between links by considering the policies of neighbouring agents using multi-agent strategy	Yes	The TCCA-MADDPG should have been compared with DDPG and not DQN since both TCCA-MADDPG and DDPG work in continuous environment.	Not stated
[99]	MADDPG	The authors proposed a MADDPG for effective traffic load engineering in SDN-IoV using a real-time distributed cooperative assignment approach via the Actor-Critic network	No	Since MDP was not used to mathematically define the network parameters, the measuring metrics for success is not well defined.	No stated

IV. Open Research Issues

In this section, we looked at the research gaps identified after the review. From the review summary shown in Table I, it is conclusive that, SDN-based TE solutions using RL agents has the potential to eliminate completely network degradation and provide a network recommender system for end users. From this review, some future research issues exist.

A. RL Agent Implementation

From the review RL agents are designed and situated at the control or application plane of the SDN architecture. For a more efficient and pro-active TE solutions, new SDN design architectures can situate the RL agent as mini-embedded applications adapted to dedicated forwarding devices with oversight from the SDN controller. With performance comparison based on end-to-end delay and response time [47], data plane based RL agents will enable a faster network congestion detection and prevention since the agents are closer to the forwarding devices.

B. RL Agent Algorithm

For TE, most RL agents use model-free based algorithms for policy enforcement and rewards. Though model-based algorithms have high computational complexities, a hybrid architecture that enables the RL agent to select either algorithm based on reward has a research value. Using trial-or-error and referencing a model will give more intelligence to the RL agent. The agent will have the capacity to decide the algorithm to activate based on network complexity and the priority of applications.

C. Multi-Agent Reinforcement Learning

For faster convergence and collaborative learning, MARL solutions in TE though complex is the future in solving network related routing and load-balancing in SDN architecture. The advent of connected devices will only increase with time. MARL agents from review have limited research [98][99] TE solutions in SDN. MARL when proposed efficiently can segment the network into smaller units with multi-agent capabilities.

V. Conclusion

Software-Defined Networks (SDN) has emerged to give more control in network management by separating the control layer from the forwarding devices. This separation has given a centralized programmable supervisory role to the controller and a flexible management of network flows in forwarding devices. In regulating the behaviour of data transmitted over the network, we discussed the relevance of Reinforcement Learning in SDN for Traffic Engineering. This paper explained major reviews using RL techniques in network traffic management and the action of agents on the environment for rewards and new states. The review further detailed the mathematical modelling of agents and environment using the Markov Decision Process (MDP). We illustrated with diagrams SARL and MARL agents and detailed their importance in regards to TE.

With Reinforcement Learning, agents are modelled in a controlled loop to take sequence of actions on the environment to receive future rewards and a new state. The agent must exploit and explore the stochastic environment through determined actions that will lead to a faster convergence. From the review, the paper offers future research options for optimal Traffic Engineering solutions in SDN.

References

[1] Zanella, Andrea, et al. "Internet of things for smart cities." *IEEE Internet of Things journal* 1.1 (2014): 22-32

[2] Salem, Mohammed A., et al. "M2M in 5G Communication Networks: Characteristics, Applications, Taxonomy, Technologies, and Future Challenges." *Fundamental and Supportive Technologies for 5G Mobile Networks*. IGI Global, 2020. 309-321.

[3] Mattisson, Sven. "Overview of 5G requirements and future wireless networks." *ESSCIRC 2017-43rd IEEE European Solid State Circuits Conference*. IEEE, 2017.

[4] Ji, Hyoungju, et al. "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects." *IEEE Wireless Communications* 25.3 (2018): 124-130.

[5] Busari, Sherif Adeshina, et al. "5G millimeter-wave mobile broadband: Performance and challenges." *IEEE Communications Magazine* 56.6 (2018): 137-143.

[6] Jungnickel, Volker, et al. "The role of small cells, coordinated multipoint, and massive MIMO in 5G." *IEEE communications magazine* 52.5 (2014): 44-51.

[7] Alencar, Felipe, et al. "How Software Aging affects SDN: A view on the controllers." *2014 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2014.

[8] Kim, Hyojoon, and Nick Feamster. "Improving network management with software defined networking." *IEEE Communications Magazine* 51.2 (2013): 114-119.

[9] Yeganeh, Soheil Hassas, Amin Tootoonchian, and Yashar Ganjali. "On scalability of software-defined networking." *IEEE Communications Magazine* 51.2 (2013): 136-141.

[10] Kim, Hyojoon, and Nick Feamster. "Improving network management with software defined networking." *IEEE Communications Magazine* 51.2 (2013): 114-119

[11] Pruss, R. M., Mcdowall, J. E., Medved, J., & Abrahams, L. (2015). *U.S. Patent No. 9,047,143*. Washington, DC: U.S. Patent and Trademark Office.

[12] Karakus, Murat, and Arjan Durresi. "Quality of service (QoS) in software defined networking (SDN): A survey." *Journal of Network and Computer Applications* 80 (2017): 200-218.

[13] Kassler, Andreas, et al. "Towards QoE-driven multimedia service negotiation and path optimization with software defined networking." *SoftCOM 2012, 20th International Conference on Software, Telecommunications and Computer Networks*. IEEE, 2012.

[14] Mahboob, Tahira, Young Rok Jung, and Min Young Chung. "Optimized Routing in Software Defined Networks–A Reinforcement Learning Approach." *International Conference on Ubiquitous Information Management and Communication*. Springer, Cham, 2019.

[15] Perera, Menuka, Kandaraj Piamrat, and Salima Hamma. "Network Traffic Classification using Machine Learning for Software Defined Networks." *Journées non thématiques GDR-RSD 2020*. 2020.

[16] Bernaille, Laurent, et al. "Traffic classification on the fly." *ACM SIGCOMM Computer Communication Review* 36.2 (2006): 23-26.

[17] Finsterbusch, Michael, et al. "A survey of payload-based traffic classification approaches." *IEEE Communications Surveys & Tutorials* 16.2 (2013): 1135-1156.

[18] Khondoker, Rahamatullah, et al. "Feature-based comparison and selection of Software Defined Networking (SDN) controllers." *2014 world congress on computer applications and information systems (WCCAIS)*. IEEE, 2014.

[19] Dey, Samrat Kumar, and Md Mahbubur Rahman. "Flow based anomaly detection in software defined networking: A deep learning approach with feature selection method." *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*. IEEE, 2018.

[20] Sutton, Richard S., and Andrew G. Barto. "Introduction to reinforcement learning. Vol. 135." *MIT press Cambridge* 5 (1998): 21-22.

[21] Szepesvári, Csaba. "Algorithms for reinforcement learning." *Synthesis lectures on artificial intelligence and machine learning* 4.1 (2010): 1-103.

[22] Stampa, G., Arias, M., Sánchez-Charles, D., Muntés-Mulero, V., & Cabellos, A. (2017). A deep-reinforcement learning approach for software-defined networking routing optimization. *arXiv preprint arXiv:1709.07080*.

[23] Møller, Martin Fodslette. "A scaled conjugate gradient algorithm for fast supervised learning." *Neural networks* 6.4 (1993): 525-533.

[24] Torrey, Lisa, and Matthew Taylor. "Teaching on a budget: Agents advising agents in reinforcement learning." *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 2013.

[25] Puterman, Martin L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[26] Boutilier, Craig, Thomas Dean, and Steve Hanks. "Decision-theoretic planning: Structural assumptions and computational leverage." *Journal of Artificial Intelligence Research* 11 (1999): 1-94.

[27] Van Otterlo, Martijn, and Marco Wiering. "Reinforcement learning and markov decision processes." *Reinforcement learning*. Springer, Berlin, Heidelberg, 2012. 3-42.

[28] Zhang, Junjie, et al. "CFR-RL: Traffic engineering with reinforcement learning in SDN." *IEEE Journal on Selected Areas in Communications* 38.10 (2020): 2249-2259.

[29] Phan, Trung V., et al. "Q-DATA: Enhanced Traffic Flow Monitoring in Software-Defined Networks applying Q-learning." *2019 15th International Conference on Network and Service Management (CNSM)*. IEEE, 2019.

[30] Mu, Ting-Yu, et al. "SDN flow entry management using reinforcement learning." *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 13.2 (2018): 1-23.

[31] Huang, Xiaohong, et al. "Deep reinforcement learning for multimedia traffic control in software defined networking." *IEEE Network* 32.6 (2018): 35-41.

[32] Khan, Asiya, Lingfen Sun, and Emmanuel Ifeachor. "QoE prediction model and its application in video quality adaptation over UMTS networks." *IEEE Transactions on Multimedia* 14.2 (2011): 431-442.

[33] Zhang, Hongming, and Tianyang Yu. "Taxonomy of Reinforcement Learning Algorithms." *Deep Reinforcement Learning*. Springer, Singapore, 2020. 125-133.

[34] Buşoniu, Lucian, et al. "Approximate reinforcement learning: An overview." *2011 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*. IEEE, 2011.

[35] Polydoros, Athanasios S., and Lazaros Nalpantidis. "Survey of model-based reinforcement learning: Applications on robotics." *Journal of Intelligent & Robotic Systems* 86.2 (2017): 153-173.

[36] Clavera, Ignasi, et al. "Model-based reinforcement learning via meta-policy optimization." *Conference on Robot Learning*. PMLR, 2018.

[37] Degris, Thomas, Patrick M. Pilarski, and Richard S. Sutton. "Model-free reinforcement learning with continuous action in practice." *2012 American Control Conference (ACC)*. IEEE, 2012.

[38] Song, Zhao, and Wen Sun. "Efficient model-free reinforcement learning in metric spaces." *arXiv preprint arXiv:1905.00475* (2019).

[39] Akrour, Riad, et al. "Model-free trajectory optimization for reinforcement learning." *International Conference on Machine Learning*. PMLR, 2016.

[40] Sutton, Richard S., and Andrew G. Barto. "Introduction to reinforcement learning. Vol. 135." *MIT press Cambridge* 5 (1998): 21-22.

[41] Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013).

[42] Kumar, Arun, Navneet Paul, and S. N. Omkar. "Bipedal walking robot using deep deterministic policy gradient." *arXiv preprint arXiv:1807.05924* (2018).

[43] Williams, Ronald J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine learning* 8.3-4 (1992): 229-256.

[44] Chiesa, Marco, Guy Kindler, and Michael Schapira. "Traffic engineering with equal-cost-multipath: An algorithmic perspective." *IEEE/ACM Transactions on Networking* 25.2 (2016): 779-792.

[45] Salisbury, B. "TCAMs and OpenFlow-what every SDN practitioner must know." *See http://tinyurl. com/kjy99uw* (2012).

[46] Khan, Asiya, Lingfen Sun, and Emmanuel Ifeachor. "QoE prediction model and its application in video quality adaptation over UMTS networks." *IEEE Transactions on Multimedia* 14.2 (2011): 431-442.

[47] Chin, Tommy, Mohamed Rahouti, and Kaiqi Xiong. "Applying software-defined networking to minimize the end-to-end delay of network services." *ACM SIGAPP Applied Computing Review* 18.1 (2018): 30-40.

[48] Nachum, Ofir, et al. "Bridging the gap between value and policy based reinforcement learning." *arXiv preprint arXiv:1702.08892* (2017).

[49] Suthaharan, Shan. "Machine learning models and algorithms for big data classification." *Integr. Ser. Inf. Syst* 36 (2016): 1-12.

[50] Chavula, Josiah, Melissa Densmore, and Hussein Suleman. "Using SDN and reinforcement learning for traffic engineering in UbuntuNet Alliance." *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2016.

[51] Barbehenn, Michael. "A note on the complexity of Dijkstra's algorithm for graphs with weighted vertices." *IEEE transactions on computers* 47.2 (1998): 263.

[52] Kim, Seonhyeok, et al. "Congestion prevention mechanism based on Q-leaning for efficient routing in SDN." *2016 International Conference on Information Networking (ICOIN)*. IEEE, 2016.

[53] Tennakoon, Deepal, Suneth Karunarathna, and Brian Udugama. "Q-learning approach for load-balancing in software defined networks." *2018 Moratuwa engineering research conference (MERCon)*. IEEE, 2018.

[54] LIANG, Siyuan, et al. "Load Balancing Algorithm of Controller Based on SDN Architecture Under Machine Learning." *Journal of Systems Science and Information* 8.6 (2020): 578-588.

[55] Wang, Ke, Wai-Choong Wong, and Teck Yoong Chai. "A MANET routing protocol using Q-learning method integrated with Bayesian network." *2012 IEEE International Conference on Communication Systems (ICCS)*. IEEE, 2012.

[56] Min, Zhu, Qu Hua, and Zhao Jihong. "Dynamic switch migration algorithm with Q-learning towards scalable SDN control plane." *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2017.

[57] Yu, Chen, et al. "Intelligent optimizing scheme for load balancing in software defined networks." *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE, 2017.

[58] Smith, Brian L., and Michael J. Demetsky. "Traffic flow forecasting: comparison of modeling approaches." *Journal of transportation engineering* 123.4 (1997): 261-266.

[59] Choi, Hanhimnara, et al. "UDP Flow Entry Eviction Strategy Using Q-Learning in Software Defined Networking." *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 2020.

[60] Nadeau, Thomas D., and Ken Gray. *SDN: Software Defined Networks: an authoritative review of network programmability technologies*. " O'Reilly Media, Inc.", 2013.

[61] Hausknecht, Matthew, and Peter Stone. "Deep reinforcement learning in parameterized action space." *arXiv preprint arXiv:1511.04143* (2015).

[62] Yang, Shike, Haobin Shi, and Hengsheng Zhang. "Dynamic Load Balancing of Multiple Controller based on Intelligent Collaboration in SDN." *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE, 2020.

[63] Li, Zhihua, et al. "Bayesian network-based virtual machines consolidation method." *Future Generation Computer Systems* 69 (2017): 75-87.

[64] Lin, Shih-Chun, et al. "QoS-aware adaptive routing in multi-layer hierarchical software defined networks: A reinforcement learning approach." *2016 IEEE International Conference on Services Computing (SCC)*. IEEE, 2016.

[65] Arruda, Carlos E., et al. "Enhanced Pub/Sub Communications for Massive IoT Traffic with SARSA Reinforcement Learning." *arXiv preprint arXiv:2101.00687* (2021).

[66] Yuan, Zhengwu, et al. "Research on Routing Optimization of SDN Network Using Reinforcement Learning Method." *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*. IEEE, 2019.

[67] Saraph, Girish P., and Pushpraj Singh. "Traffic engineering using new VS routing scheme." *2004 IEEE International Conference on Communications (IEEE Cat. No. 04CH37577)*. Vol. 2. IEEE, 2004.

[68] Wu, Yiwen, et al. "Deep Reinforcement Learning for Controller Placement in Software Defined Network." *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020.

[69] Yeo, Sangho, et al. "Achieving Balanced Load Distribution with Reinforcement Learning-Based Switch Migration in Distributed SDN Controllers." *Electronics* 10.2 (2021): 162.

[70] Yao, Zan, Ying Wang, and Xuesong Qiu. "DQN-based energy-efficient routing algorithm in software-defined data centers." *International Journal of Distributed Sensor Networks* 16.6 (2020): 1550147720935775.

[71] Fu, Qiongxiao, et al. "Deep Q-learning for routing schemes in SDN-based data center networks." *IEEE Access* 8 (2020): 103491-103499.

[72] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." *2017 International Conference on Engineering and Technology (ICET)*. Ieee, 2017.

[73] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." *arXiv preprint arXiv:1511.08458* (2015).

[74] Bouzidi, El Hocine, Abdelkader Outtagarts, and Rami Langar. "Deep reinforcement learning application for network latency management in software defined networks." *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019.

[75] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." *Thirteenth annual conference of the international speech communication association*. 2012.

[76] Lipton, Zachary C., et al. "Learning to diagnose with LSTM recurrent neural networks." *arXiv preprint arXiv:1511.03677* (2015).

[77] Pitaksanonkul, Anucha, et al. "DTR: A defect-tolerant routing algorithm." *Proceedings of the 26th ACM/IEEE Design Automation Conference*. 1989.

[78] Streijl, Robert C., Stefan Winkler, and David S. Hands. "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives." *Multimedia Systems* 22.2 (2016): 213-227.

[79] Silver, David, et al. "Deterministic policy gradient algorithms." *International conference on machine learning*. PMLR, 2014.

[80] Stampa, Giorgio, et al. "A deep-reinforcement learning approach for software-defined networking routing optimization." *arXiv preprint arXiv:1709.07080* (2017).

[81] Schmidt, Wouter F., Martin A. Kraaijveld, and Robert PW Duin. "Feed forward neural networks with random weights." *International Conference on Pattern Recognition*. IEEE COMPUTER SOCIETY PRESS, 1992.

[82] Lu, Xiaoye, et al. "SDN routing optimization based on improved Reinforcement learning." *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*. 2020.

[83] Guo, Xuancheng, et al. "Deep-reinforcement-learning-based QoS-aware secure routing for SDN-IoT." *IEEE Internet of Things Journal* 7.7 (2019): 6242-6251.

[84] Dhawan, Mohan, et al. "SPHINX: detecting security attacks in software-defined networks." *Ndss*. Vol. 15. 2015.

[85] Ashraf, Javed, and Seemab Latif. "Handling intrusion and DDoS attacks in Software Defined Networks using machine learning techniques." *2014 National Software Engineering Conference*. IEEE, 2014.

[86] Zhang, Qingyi, et al. "Intelligent Content-Aware Traffic Engineering for SDN: An AI-Driven Approach." *IEEE Network* 34.3 (2020): 186-193.

[87] Dannewitz, Christian, et al. "Network of information (netinf)–an information-centric networking architecture." *Computer Communications* 36.7 (2013): 721-735.

[88] Ghosal, Gaurav R., et al. "A Deep Deterministic Policy Gradient Based Network Scheduler For Deadline-Driven Data Transfers." *2020 IFIP Networking Conference (Networking)*. IEEE, 2020.

[89] Xu, Chunlei, Weijin Zhuang, and Hong Zhang. "A Deep-reinforcement Learning Approach for SDN Routing Optimization." *Proceedings of the 4th International Conference on Computer Science and Application Engineering*. 2020.

[90] Buşoniu, Lucian, Robert Babuška, and Bart De Schutter. "Multi-agent reinforcement learning: An overview." *Innovations in multi-agent systems and applications-1* (2010): 183-221.

[91] Christianos, Filippos, et al. "Scaling Multi-Agent Reinforcement Learning with Selective Parameter Sharing." *arXiv preprint arXiv:2102.07475* (2021).

[92] Omidshafiei, Shayegan, et al. "Learning to teach in cooperative multiagent reinforcement learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

[93] Busoniu, Lucian, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2 (2008): 156-172.

[94] Gupta, Jayesh K., Maxim Egorov, and Mykel Kochenderfer. "Cooperative multi-agent control using deep reinforcement learning." *International Conference on Autonomous Agents and Multiagent Systems*. Springer, Cham, 2017.

[95] Egorov, Maxim. "Multi-agent deep reinforcement learning." *CS231n: convolutional neural networks for visual recognition* (2016): 1-8.

[96] Chu, Tianshu, et al. "Multi-agent deep reinforcement learning for large-scale traffic signal control." *IEEE Transactions on Intelligent Transportation Systems* 21.3 (2019): 1086-1095.

[97] Lowe, Ryan, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." *arXiv preprint arXiv:1706.02275* (2017)

[98] Wu, Tong, et al. "Joint Traffic Control and Multi-Channel Reassignment for Core Backbone Network in SDN-IoT: A Multi-Agent Deep Reinforcement Learning Approach." *IEEE Transactions on Network Science and Engineering* (2020).

[99] Yuan, Tingting, et al. "Dynamic Controller Assignment in Software Defined Internet of Vehicles through Multi-Agent Deep Reinforcement Learning." *IEEE Transactions on Network and Service Management* (2020).

[100] Jiacheng, Chen, et al. "Software defined Internet of vehicles: Architecture, challenges and solutions." (2016): 14-26.

[101] Dey, Ayon. "Machine learning algorithms: a review." *International Journal of Computer Science and Information Technologies* 7.3 (2016): 1174-1179.

[102] Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervised learning." *Machine Learning* 109.2 (2020): 373-440.

[103] Sen, Pratap Chandra, Mahimarnab Hajra, and Mitadru Ghosh. "Supervised classification algorithms in machine learning: A survey and review." *Emerging technology in modelling and graphics*. Springer, Singapore, 2020. 99-111.

[104] Khanum, Memoona, et al. "A survey on unsupervised machine learning algorithms for automation, classification and maintenance." *International Journal of Computer Applications* 119.13 (2015).

[105] Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." *International Conference on Machine Learning*. PMLR, 2017.

[106] Wang, Haoran, Thaleia Zariphopoulou, and Xun Yu Zhou. "Exploration versus exploitation in reinforcement learning: a stochastic control approach." *Available at SSRN 3316387* (2019).

[107] Colas, Cédric, Olivier Sigaud, and Pierre-Yves Oudeyer. "Gep-pg: Decoupling exploration and exploitation in deep reinforcement learning algorithms." *International Conference on Machine Learning*. PMLR, 2018.

[108] Holcomb, Sean D., et al. "Overview on deepmind and its alphago zero ai." *Proceedings of the 2018 international conference on big data and education*. 2018.

How Enterprise must be Prepared to be “AI First”?

A pragmatic approach for AI adoption

Mustapha Lahlali[1], Naoual Berbiche[2], Jamila El Alami[3]
Laboratory of Systems Analysis, Information Processing and Industrial Management
Higher School of Technology 11000- Salé, Mohammed V University, Rabat, Morocco

***Abstract*—Among disruptive technologies, Artificial Intelligence (AI), Robotic Process Automation (RPA) and Machine Learning (ML) play a very important role in Businesses Transformation and continues to show great promise for creating new sources of wealth and new business models. The reality of AI in the company is not reduced to a simple process optimization. In fact, AI introduces new organizational schemes, new ways of working, new optimization niches, new services, other ways of thinking about interactions with customers and therefore a new way of doing business. It thus reshuffles competitive data and imagine innovative processes to create new business models, offering new opportunities not only for IT solution providers but also for innovators, investors and business owners. Even if the contribution of Artificial Intelligence is not to be proved, many companies face difficulties in adopting this technology, mainly due to the lack of a pragmatic approach highlighting the roles and responsibilities of the various stakeholders, especially IT professionals and business owners and the key steps to follow to make this experience a real success. This research aims to answer fundamental questions, in particular: What will bring the implementation of this technology to the business of the company? How to prepare for this adoption? and if the decision to go is confirmed, what kind of adoption approach should companies follow? and finally how can Enterprises monitor this shift to the Intelligent edge.**

***Keywords*—Artificial intelligence; machine learning; RPA; business transformation; AI adoption**

I. Introduction

Artificial Intelligence was certainly born several decades ago, but its rhythm of adoption has clearly accelerated in recent years. Indeed, this technology is no longer reserved only for the scientific world and its rich, varied and innovative applications are a delight for many companies.

This acceleration was primarily driven by the explosion in the amount of data and the digital transformation, and also by the company's enthusiasm to adopt disruptive technologies in their growth strategy. Indeed, many companies, in order to remain competitive, have been forced to rethink the way they operate by integrating technological innovation into the daily process as an enabler for creating value.

Artificial Intelligence fully expresses its potential through various functionalities: it optimizes existing processes, automates routine tasks, allows assisted monitoring, detects abnormal phenomena, predicts future actions, and interacts more and more "naturally" with humans, thanks to recent developments in natural language processing.

This article aims to illustrate how AI can bring a tangible value to the enterprise by answering pragmatic questions. It shows, in a pragmatic approach, how companies can meet the challenges and takes benefits of adopting this technology by making the most of its contributions while minimizing the associated risks.

II. Artificial Intelligence in Business: For What Purposes?

The development of Artificial Intelligence (AI) technology widens the limits of business practice for companies, thus promoting the transformation of information technology to optimize decision-making and operations. Recent research findings [1][2] presents several main developmental trends of the technology and the resulting challenges and show how dynamic AI capabilities can improve operational efficiency and business performance.

AI is definitively not a technological subject. It’s, at the first level, a business matter aiming to improve the business performances and capabilities. Fig. 1 shows the main expectations of business managers towards the adoption of AI.

Artificial Intelligence can support three important major business needs: processes automation, cognitive signals, and cognitive engagement [4].

A. Process Automation

Robotic Process Automation paradigms (RPA) bring the opportunity to integrate robots inside the enterprise chain of value. RPA is not aiming to replace human, it’s mainly designed to increase human productivity by automating repetitively and costly tasks [5]. Basically, RPA robots correspond to software programs acting as a digital worker and ensuring many kinds of tasks including:

- Automating customer relationship management and ensuring clear and comprehensive customer data collection and structuring.
- Ensuring that multiple data are correctly updating different systems, for example, blocking immediately customer identity in case of ID lost reclamation by customers.
- Verifying the integrity of data manipulated in different systems to ensure the authenticity of information managed in different business lines, for example, verifying automatically inside the billing system, the amount of a settlement check provided inside the supplier bills.

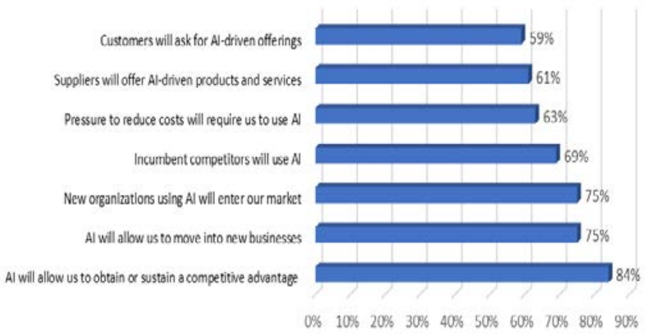


Fig. 1. Reasons for Adopting AI [3].

RPA is usually considered as the least expensive and easiest to implement module of the cognitive technologies and typically brings a quick and high return on investment.

Some enterprises launched RPA pilots in accounts payable, automatic settlement check controlling and validation, automation of IT daily tasks by avoiding manual operations and minimizing human intervention.

In all cases, RPA is not a Human predator [6]. It's a new way to let Human making valuable tasks and being supported by robots to do well and correctly daily tasks. The cost killing objective is usually forecasted by enterprises beside the use or not of the RPA. For this reason, the outsourcing/offshoring sector will be challenged by this paradigm and maybe the job losses impact can be felt more in those areas [7].

B. *Cognitive Signals*

Cognitive signals can detect patterns in vast volumes of data and interpret their meaning (e.g., predicting purchasing habits, identifying credit card frauds, determining trends, qualifying data quality,..).

Cognitive signals provided by Machine Learning are stronger and more sophisticated than those provided by analytics for those three principal reasons:

- The huge amount of data.
- The ability for Machine Learning-based systems to learn from the data collected.
- The prediction capabilities to anticipate features.

It should also be noted that Machine Learning bring new capabilities and features like text and voice recognition. In the past, a lot of efforts had to be made to collect and to manipulate structured data. Today, Machine learning brings a significant contribution to do these tasks more intelligently for both structured and unstructured data thanks to the important developments in computing power.

Cognitive signal applications are typically used to improve performance on jobs and tasks only machines can do, so we can therefore imagine a fruitful collaboration between humans and machines.

C. *Cognitive Engagement*

Smart assistants are revolutionizing the daily lives of individuals and businesses. With their ability to interpret and understand human language through Machine Learning capabilities, they enable the following opportunities:

- Intelligent agents [8] ensuring customer assistance every day by supporting simple to complex queries via quick and concise responses.
- The product offering is enhanced by learning from the various customer questions addressed to chatbots, allowing the company to anticipate customer demands and to differentiate itself from the competition.
- Internally, the distance between the company and its employees is reduced by setting up intelligent channels of exchange and support for daily queries (e.g., Administrative HR assistance, IT help desk,…).

Fig. 2 shows the different types of AI adoption within Companies. It gives a clear visibility on the current use and forecasts of the future uses of the intelligent technologies on which the AI is based to transform the business processes.

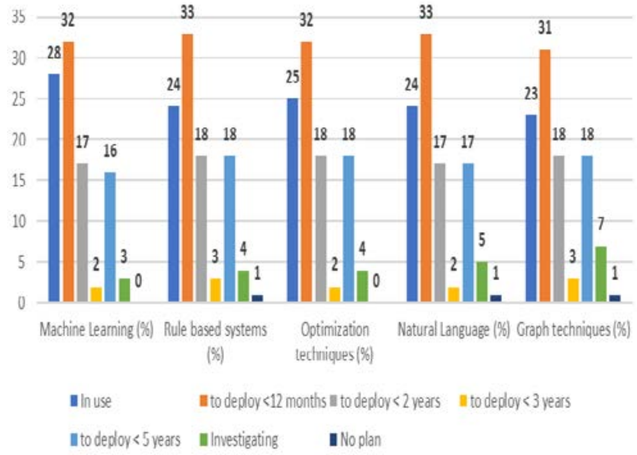


Fig. 2. AI Adoption by Type [9].

III. Challenges Facing AI Integration and Adoption

Companies must put in place the necessary prerequisites for the success of its AI integration projects, which are related, in particular, to technological, financial, organizational and human aspects.

It should be noted that once the technology adoption process is successful, the impact on the business can be very important. For example, in the retail industry, a lot of champions made a significative growth by adopting IA like Amazon and Walmart [10].

A. *Understanding the Technologies*

Before going in depth and embarking on AI adventures, companies must ensure that they have a thorough understanding of the technology and be able to make a distinction between different paradigms and their potential for value creation. For example, RPA offers no ability intelligence or learning, it automates the cumbersome and repetitive tasks, unlike the deep learning where there is a strong element of intelligence. In all cases, AI-driven companies will have the potential to compete by looking beyond the initial product or service to the full potential of the technology available [11].

From talent management perspective, companies must prepare task forces which are necessary to manage these AI projects. From Data Scientist to AI Architect, each role has a specific mission to leverage one or multiple sides of the technology in order to succeed in AI projects. Unfortunately, this is not enough. The construction of an ecosystem made up of experts and industrialized product offerings is considered as a fundamental pillar in the mastery of technology and a guarantee to ensure its viability.

B. Financial Effort

Experimentation of the Machine Learning and Data Intelligence subjects require a significant financial effort which must be justified from the point of view of viability. In fact, to make this successful, companies are called upon to mobilize a lot of expertise which requires large expenses. The whole challenge lies in the adequacy between the expenditure and the return on investment which is generally difficult to ensure given the high level of uncertainty that characterizes a large number of AI projects.

Also, thanks to the learning techniques offered by AI (e.g. supervised, unsupervised, by reinforcement) [12] [13], many opportunities are now available to companies to create new business models and offer a new form of value. In this context, the predictive aspect remains an undoubtedly promising axis for designing intelligent systems capable of reasoning, analyzing and perhaps making decisions.

Some questions will arise in terms of risk management accompanying this revolution, but from a business point of view, several possibilities arise for switching approaches from classic IT solutions towards intelligent and augmented solutions.

Faced with increased competition, IT Departments will find themselves at the center of all interests and will transform from a support unit to a profit unit because in the era of intelligence, there mission will not be to support the trades but to transform the whole operational model of the company by making intelligence an axis of wealth. In this perspective, IT teams will find themselves in a new management style of collaborative work and agile implementation with a record speed of change and transformation and also with a drastic impact on the product development methodology and the implementation of intelligent journeys.

It is almost certain that the “Augmented CIO (Chief Information Officer)” is a CIO who will use “intelligence” as a workhorse to enhance its contribution to the company's strategy.

C. Organization Challenges: Internal and External Dependencies

Faced with the diversity of profiles and the complexity of AI subjects, many organizational challenges will be the object of challenge. Some companies have adopted the Chief AI Officer (CAIO) function as an unifying role to create synergy between business, IT and statisticians and make teamwork more fluid and more constructive.

Most AI topics are expert and advanced computing topics which require the collaboration of several profiles from different structures, whether internal to the company or located in its ecosystem (e.g. service providers, partners, IT experts). In this situation, managing relationships between different teams is almost considered as a daily challenge. For example, in terms of big player dependencies, many suppliers are positioning themselves in this segment without necessarily having a solid foundation with commercial promises that are not necessarily tenable. For this reason, many companies consider that Open source-based solutions are an issue that will not depend on the big players, but at the same time it requires the loyalty of internal experts which is not at all an easy task. In this perspective, the real challenge will consist in finding a balance between business ambition, technological capacity and a better organization to co-construct intelligent value.

D. Thinking Platforms

AI is drastically transforming the technological DNA of the company. In fact, the rate of dependence on experts and IT giants remains too high. In this sense, the company must adopt a new approach. The first one is based on internal Labs to develop and retain its own internal experts by associating them with the company's businesses. The second point concerns the implementation approach and methodology. In fact, supplying the process of creating AI products is totally different from traditional IT products, hence from a technical standpoint, the company must adopt a reasoning based on platforms and not on isolated solutions. These platforms will aim to industrialize the entire chain from data collection to the operationalization of machine learning models with the possibility of mixing between the private cloud and the public cloud to optimize production capacities and have a very significative return on investment.

E. Skills and Task Forces

Companies that want to win the race for a successful AI adoption need to be truly convinced that this victory comes first and foremost through the provision of skills, competencies and talents. Furthermore, developing and maintaining smart talents is a very difficult HR subject because it’s about innovation in HR management practices to create suitable policies to smart task forces highly targeted by competitors. Enterprise should protect their internal skills by supporting them and rewarding their efforts.

The availability of qualified resources is a precondition for a successful leap towards the era of intelligence. Different modes of organization are possible such as internal competence centers and Tech accelerators. The intangible capital of the company represented in its human wealth is definitely a sure catalyst for success.

The sheer diversity and complexity of AI projects combined with a requirement for rapid time to production create the need to find key AI roles to achieve successful AI projects. Fig. 3 shows main AI specialist’s roles and responsibilities [14].

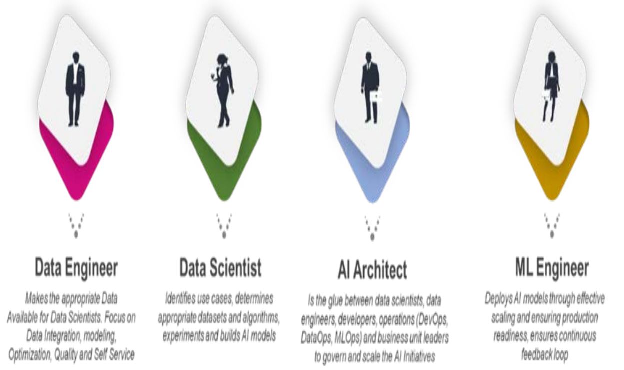

Data Engineer
Makes the appropriate Data Available for Data Scientists. Focus on Data Integration, modeling, Optimization, Quality and Self Service
Data Scientist
Identifies use cases, determines appropriate datasets and algorithms, experiments and builds AI models
AI Architect
Is the glue between data scientists, data engineers, developers, operations (DevOps, DataOps, MLOps) and business unit leaders to govern and scale the AI Initiatives
ML Engineer
Deploys AI models through effective scaling and ensuring production readiness, ensures continuous feedback loop

Fig. 3. AI's Core Roles and Responsibilities.

F. *Ethics*

A question that comes up quite often in AI implementation projects concerns the limitation of ethical responsibilities for decisions that might be made by an algorithm based on machine intelligence [15].

For example, administrative, social or financial decisions can be subject to a complaint. Indeed, the decisions taken by human can consider emotional et behaviors factors, even subjective in some cases, that a machine cannot necessarily make. All the effort therefore remains to clearly define the tolerance thresholds and to delimit responsibilities between human and machines.

IV. AI Initiatives and Projects: The Way of Success

The success of an initiative based on AI begins with the identification of opportunities, the determination of the use cases to better achieve these opportunities and finally the choice of the right technology and solutions to meet the expected results.

A. *Creating a Portfolio of Projects and Identifying Opportunities*

AI programs [16] are becoming more and more present in companies' IT portfolios. These programs consist of a several use cases and projects generated, most of the time, after an experimentation phase. This structuring approach helps the company to choose and prioritize AI initiatives and projects according to business needs, market opportunities and needs, team capacity and also taking in account prioritization of intelligence technologies according to the context and providers maturity. The management of these programs allows top management to continuously monitor the evolution of the company's intelligence quotient.

B. *Targeting the use Cases*

To approach AI projects, companies must start with an exploration phase which consists in identifying concrete use cases where intelligence is essential and also which constitute a real axis of improvement and wealth creation.

These use cases will be essentially inspired by the strategy of the company, its operational objectives, the market opportunities and, above all, the trajectories identified to ensure growth.

Basically, the use cases must be proposed and especially well defined in feasibility term correlated to the expected value. The result of this work will offer the company a measured trajectory to appropriate the technology and move slowly but surely towards the intelligent enterprise.

C. *Selecting Tools and the Technology*

One of the major challenges to be faced is to properly identify the right technology through which intelligence will be showcased.

For example, chatbots or digital assistants can be used in certain contexts where the company must be very close to its customers and respond in record time to their requests. As for facial recognition, which is very data-intensive, it is of interest to use in contexts of identification or detection of abnormal behavior. In all cases, AI architects must use the right channel to better achieve the desired impact.

V. Three Steps to Integrate AI in Enterprise Business Applications

Once AI-based initiatives are successful, companies can move up to integrate these use cases into their process. To achieve this objective, companies must first launch pilot projects to clearly identify all aspects related to the implementation of the technology (technical prerequisites, financial impacts, risks, organizational aspects, Human Resources, impact on the Ecosystem,..). The second step relates to the optimization, review and redesign of the processes. The third and last phase obviously relates to the full-scale generalization.

A. *Setting up Pilot Projets*

The creative appetite for intelligent products and services is always faced with the contradiction between promises and real creative capabilities. For this reason, companies from the Test & learn perspective, must test as part of Proof of Concepts (PoCs) the Business impact on a reduced scope before any full-scale realization.

The SENSE-THINK-ACT [17] model can be a suitable tool to challenge AI PoCs, by preparing Data, designing products and creating AI powered solutions.

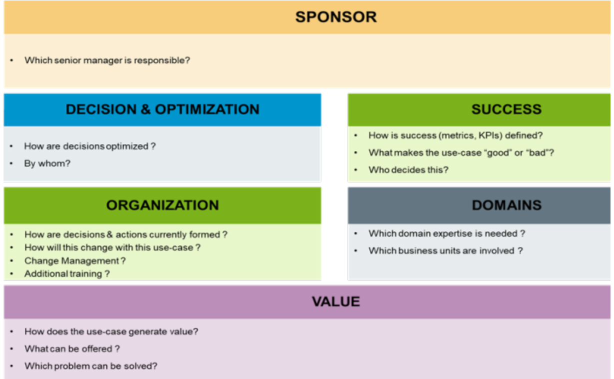
The phase of testing and validating the added value of the technology on the business processes remains a key stage in AI adoption. Companies must in no case derail the rule under the pretext of submitting to external pressure or simply following the fashion effect.

B. *Redesigning Processes*

The implementation of any projects based on Artificial Intelligence does not mean necessarily the fact that the company should automates its processes in an iso functional way. To get the most out of the technology, companies must think about reshaping their way of doing things (thinking out of the box), so that machines and Humans can perform in perfect harmony, because, at the end, each actor should compensate the weaknesses of the other.

Hironori Takeuchia and Shuichiro Yamamoto [18] worked on a Generic business–AI alignment model giving the capacity to integrate AI within the Enterprise architecture and giving a strong and didactic way to facilitate AI integration with business processes.

Today, every business realizes the drastic impact of enterprise architecture on the leap to intelligent enterprise. This integration is illustrated in Fig. 4 according to the Enterprise AI Canvas [19].



SPONSOR

- Which senior manager is responsible?

DECISION & OPTIMIZATION

- How are decisions optimized ?
- By whom?

SUCCESS

- How is success (metrics, KPIs) defined?
- What makes the use-case "good" or "bad"?
- Who decides this?

ORGANIZATION

- How are decisions & actions currently formed ?
- How will this change with this use-case ?
- Change Management ?
- Additional training ?

DOMAINS

- Which domain expertise is needed ?
- Which business units are involved ?

VALUE

- How does the use-case generate value?
- What can be offered ?
- Which problem can be solved?

Fig. 4. Enterprise AI Canvas.

Working in a common framework under solid governance is far from being a choice. Indeed, the orchestration of flows and the orchestration of exchanges is necessary to successfully integrate AI into the company and ensure that all professionals and AI engineers work in perfect synergy.

C. Scaling Up

Several experiences have shown that many companies having passed "successfully" proof of concepts stage, have unfortunately failed to make it a real use case, because the mastery of technology and IT talents cannot be considered as the only factors guaranteeing success [20].

Indeed, before any generalization of an experimentation, companies must put in place the necessary prerequisites for its success, in particular: a perfect synergy and convergence of point of view and vision between the business owners and the IT professionals, an adequate governance framework and a solid change management to allow the appropriate use of the future system.

The Scaling up approach is a mix between the AI adoption degree and the knowledge of the technology as shown in Fig. 5:

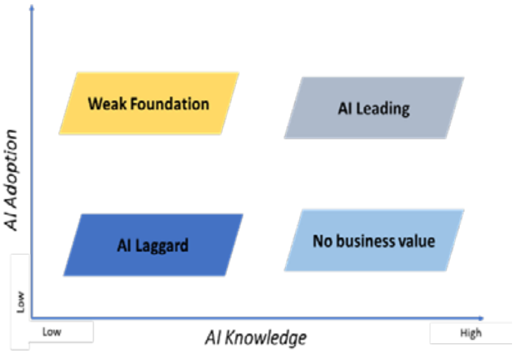


Fig. 5. Blueprint for AI Scaling up.

VI. Conclusion

Today, no one can deny the importance and the contribution of Artificial Intelligence in the transformation of our way of acting, driving our processes and especially of creating value. To take maximum benefits from this technology, companies must prepare the prerequisites for a successful adoption, in terms of understanding the technology, coordinating initiatives, making data available and implementing innovative use cases.

In that way, to make the implementation of AI projects a happy story, company faces many challenges and it is brought to change management issues due to the massive adoption of disruptive technologies, the use of structured and unstructured data and the transformation in the ways of doing business.

Our work aims to implement a Model of AI adoption inside the Enterprise. We call it 2PS Model (Pilot, Project and Scale) giving the opportunity to leverage AI adoption by implementing holistic approach. Each step must be well prepared and well executed to ensure success.

Within this study, we tried to answer a couple of questions regarding enterprise readiness to Adopt AI. At this stage, aspect concerning Explainable AI and Trust AI are not discussed in this paper. In future publication we will challenge those aspects as perspective and evolution of our thinking about this problematic.

Our approach can be compared with some frameworks like Google AI readiness framework which deal with AI adoption from Learn, Lead & Access perspective by challenging tactical, strategic and transformation aspects. We deal with the AI adoption from strategic point of view to give the core capacities to enterprise to deal with AI disruption. The Google framework supposes that enterprise is smart enough to go further in its AI journey. In that way, our work aims to prepare the enterprise to be able to leverage frameworks (like google or other) to create maximum value for user experience and a better contribution for enterprise growth.

Finally, AI platforms are basically based on Data, Machine Learning and Models Operations. So, as another perspective of this research, we aim to analyze how to leverage those capabilities to give a strong enabler for AI adoption through the implementation of a case study to illustrate the disruptive impact of the AI on business process. We aim within the future work demonstrate the power of this technology, particularly these aspects related to Data Science and Machine Learning, on improving both the internal performance of the company and its interactions with its eco-system.

References

[1] Y. Chen, z. Lin: "Business intelligence capabilities and firm performance: a study in china", International journal of information management, August 2020, p. 4.

[2] S. m. Correia loureiro, j. Guerreiro, l. Tussyadiah: "Artificial intelligence in business: state of the art and future research agenda", Journal of business research, November 2020.

[3] S. Ransbotham, d. Kiron, p. Gerbert, m. Reeves: "Reshaping business with artificial intelligence closing the gap between ambition and action", MIT sloan management review, October 2019, p. 5.

[4] T. H. Davenport, R. Ronanki : "On AI, Analytics, and the New Machine Age", Harvard business review, 2018, p 11.

[5] J. G. Enríquez, a. Jiménez-ramírez, f. J. Domínguez-mayo, j. A. García-garcía: “Robotic process automation: a scientific and industrial systematic mapping study”, Ieee access, March 2020, pp 1-4.

[6] Wil m. p. Van der Aalst, m. Bichler, a. Heinzl: “Robotic Process Automation”, Springer, May 2018, p 2.

[7] J. Surowiecki: “The great tech panic: Robots won’t take all our jobs”, www.wired.com, 2018.

[8] T. Boobier: “AI and the future of banking”, Wiley finance series, May 2020, pp 261-265.

[9] S. Sicular: “The hype cycle for artificial intelligence 2020 reflects the state of AI in the enterprise”, Gartner, October 2020, p. 1.

[10] F. D. Weber, r. Schütte: “State-of-the-art and adoption of artificial intelligence in retailing”, Digital Policy, Regulation and Governance, 2019, p. 9.

[11] K. Lakhani, m. Lansiti: “Competing in the age of AI: strategy and leadership when algorithms and networks run the world”, Harvard business review, January 2020, pp 8-10.

[12] Cigref: “Artificial intelligence in companies: strategies, governance and data intelligence challenges”, October 2018, p. 4.

[13] R. e. Neapolitan, x. Jiang: “Artificial Intelligence, With an Introduction to Machine Learning, SECOND EDITION”, CRC Press, 2018, PP 89 – 330.

[14] L. Goasduff: “How to staff your AI team”, Gartner, November 2020, p.2.

[15] B. W. Wirtz, J. C. Weyerer, C. Geyer: “Artificial Intelligence and the Public Sector—Applications and Challenges”, International Journal of Public Administration, 2018, p. 11.

[16] E. Brynjolfsson, a. Mcafee: “The business of artificial intelligence”, Harvard business review, July 2017.

[17] N. Soni, e. K. Sharma, n. Singh, a. Kapoor: “Impact of artificial intelligence on businesses: from research, innovation, market deployment to future shifts in business models”, International conference on computational intelligence and data science (iccids), April 2019, p. 8.

[18] H. Takeuchi, s. Yamamoto: “Business analysis method for constructing business–AI alignment model”, 24th International conference on knowledge-based and intelligent information & engineering systems, 2020, pp 4-9.

[19] U. Kerzel: “Enterprise AI canvas integrating artificial intelligence into business”, Arxiv, September 2020, pp. 10-11.

[20] B. Mccarthy, s. Tamim: “Building AI powered organization”, harvard business review, August 2019, pp 9-13.

The Effect of Augmented Reality in Improving Visual Thinking in Mathematics of 10th-Grade Students in Jordan

Dr. Fadi Abdul Raheem Odeh Bani Ahmad
Assistant Professor of Education Technology
Middle East University, Amman
Jordan

***Abstract*—Augmented reality is one of the key issues in the area of improving visual thinking in science courses such as Mathematics. Augmented reality also offers a significant and effective role in the educational process. The current study aimed to investigate the effect of augmented reality in improving visual thinking of 10th-grade students in mathematics in Jordan. To achieve the objectives of the study, the methodology used includes the application of the semi-experimental approach and augmented reality technology. The methodology used also includes preparing a test to measure visual thinking comprising (20) multiple-choice items used as a pre-and post-test, and its validity and reliability are verified. The study sample consists of (57) female students purposefully selected from the 10th-grade students at the Jerash Model Schools for the first semester of 2020/2021. The study sample is divided into two groups as follows: one is an experimental group consisting of (28) female students taught by the augmented reality technology, and the second is a control group consisting of (29) female students taught in the traditional method. The results of the study show that there are statistically significant differences at the level of (α = 0.05) in the development of visual thinking in favor of the experimental group students taught by the augmented reality technology. The study also shows that there are differences in the performance of the experimental group students in each skill of visual thinking.**

Keywords—Augmented reality technology; visual thinking development; 10th grade; mathematics

I. Introduction

Augmented reality offers a significant and effective role in the educational process because it is one of the up-to-the-minute and developed technological innovations that depend on smartphones. Augmented reality enables the student to deal with an imaginary or near-reality environment based on simulation between the student and a third-dimensional electronic environment because education is one of the areas strongly affected by technology and digitization.

At present, school and university students no longer want to learn by reading books and copying texts, but rather wish to exploit the advantages of technology for use in the classroom. Therefore, augmented reality has emerged as a development of virtual reality that requires special tools and professionalism for third-dimensional design programs. The augmented reality can be attained by less professional programs, or by using online libraries that contain a lot of previously designed third shapes. Moreover, augmented reality in the classroom contributes to making students busier with the curriculum, learning more deeply, and interacting and cooperating better. It is indicated that augmented reality furnishes the learners with multiple options in representing information dynamically and quickly, helping to increase learners' motivation, and developing their academic achievement [1]. Augmented reality is a form of technology that enhances the real world by merging the textbook with technology, especially a mobile phone. From the viewpoint of emphasizing the educational theories in the domain of educational technology on the significance of using visual thinking and its various tools such as images, shapes, graphs, and diagrams in the education process, this study investigates the possibility of the effect of augmented reality technology in the development of visual thinking among students during the educational learning process.

Augmented reality has always provided a significant and effective role in the educational process because it is one of the contemporary and developed technological innovations that depend on smartphones. With that, an investigation is conducted by the researchers to explore how to address the poor achievement of 10th-grade students in mathematics by interviewing several teachers. The results of the investigation show that the reason for the poor achievement of 10th-grade students is the stagnation that characterizes mathematics and is reflected in the difficulty faced by students in understanding and recognizing the content, and this is evidenced by their low achievement and motivation towards learning in general and learning mathematics in particular. Accordingly, it is necessary to move towards learning that simplifies the presentation of scientific content and lessons in a manner that meets their desires and takes into account their differences, so augmented reality is the choice from the researchers' point of view because it does not require effort, time, and cost to enhance, enrich, and support the environment of the educational learning process. In light of the recommendations of the study, it is noted that augmented reality motivates students to conduct interactive experiments and develops their visual thinking [2]. Also, it is maintained that augmented reality is used in Japan and Austria in the museums of mathematics and science to develop remembering and prediction among learners [3]. It is also recommended that augmented reality contributes to achieving happiness among students in the primary stage, which increases their motivation and ability to learn and read [4]. Based on the investigative process carried out by the researchers, the study problem of the study lies in investigating the effect of augmented reality in improving the visual thinking of 10th-grade students in mathematics in Jordan.

Due to the problem of the study, the following questions are articulated: What is the effect of augmented reality on developing visual thinking in mathematics among 10th-grade students in the Jerash governorate? The second question: What is the effect of augmented reality on the development of each of the visual thinking skills (visual reading skill, visual discrimination skill, scientific deduction skill, the skill of analyzing and interpreting the visual shape) in mathematics among 10th-grade students in Jerash governorate?

With that, the following objectives formatted to answer the questions of the study: Explore the effect of augmented reality on developing visual thinking in mathematics among 10th-grade students in Jordan and examine the effect of augmented reality on the development of each of the visual thinking skills (visual reading skill, visual discrimination skill, scientific deduction skill, the skill of analyzing and interpreting the visual shape) in mathematics among 10th-grade students in Jordan.

Given the significance of augmented reality technology in developing visual thinking among students who have learned to use it, the significance of this study lies in the scarcity of Arab and foreign studies that have investigated the effectiveness of augmented reality technology in visual thinking. Importantly, this study is the first at the level of Jordan and the third at the regional level investigating the increasing knowledge in the area of augmented reality and visual thinking and its effect on students. Accordingly, this paper presents an analysis of the effect of augmented reality in improving visual thinking in mathematics of 10th-grade students in Jordan.

II. Literature Review

In light of the nature of the study, it is necessary to demarcate the literature review into four parts as follows: use of augmented reality technology, visual thinking, visual thinking skills, and visual thinking tools.

A. *Use of Augmented Reality Arachnology*

Augmented reality technology is considered the product of the development of virtual reality, which is the integration of real reality with augmented information, whether it is still images, videos, or texts taking into account that real reality (RR) cannot be ignored to enhance the perception to develop visual thinking of the learner in mathematics. Augmented reality technology is distinguished from virtual reality in that augmented reality is the cutting-edge, easiest, and latest technology used in the educational learning process. Among its key characteristics are providing clear, flawless, and accurate information, the possibility of entering information easily and effectively, the ability to interact between the teacher and the learner in a positive, flexible, and less expensive manner. As put by [5], augmented reality provides an interactive environment that the real world does not offer.

Augmented reality also employs multimedia as digital content augmented with 2D and 3D animations or videos. Moreover, it is indicated that augmented reality takes into account individual differences among learners and is a safe way for children while the impact of learning with the learner's sense of fun [6]. This is confirmed by [4] that the animation used in augmented reality contributes to achieving educational objectives and furnished children with fun and happiness and that the appearance of 3-dimensional models while learning increases their motivation towards learning.

Augmented reality is characterized by the complementarity between the media based on this type of education, which aims to enhance and enrich the real world. The media-based augmented reality such as icons, printed images, and interactive videos contributes to raising the efficacy of academic achievement, allows students to control the presentation, and repeats the scenes that embody the information, ensuring the realization of the principle of visual analysis by controlling the speed of presentation [7]. The augmented reality contributes to achieving the highest degree of interaction between the scientific content and the learner, which is difficult to achieve in traditional learning, especially in learning abstract scientific concepts in mathematics, which are difficult for the learner to understand and perceive, as augmented reality contributes to the survival of the learning effect and the achievement of self-learning [8].

The concept of augmented reality is one of the up-to-date concepts and indicates that augmented reality has expanded in real reality by adding layers of computer-generated information to the real environment, and this information may be texts, videos, or static and animated graphics [9]. Likewise, it is asserted that augmented reality combines the virtual environment and the real environment in a 3-dimensional image for the user to see as a real-world [10]. Moreover, it is believed that augmented reality is a system that allows the integration between the real and the virtual world in the same place, where they interact together at the same time, as students use computers to attain rich environments and meaningful multi-media content related to the context of the educational content [11]. As specified by [12], the fundamental point is the difference between virtual augmented reality and augmented reality is immersion, as virtual augmented reality uses masks and special glasses that immerse the learners in learning and try to convince them that they are inside a real reality, so it is a reasonably convincing illusion, while augmented reality does not ignore the real world but uses a computer to improve it. Augmented reality technology is distinguished from virtual reality in that augmented reality is the cutting-edge, easiest, and latest technology used in the educational learning process [13]. Among its key characteristics are providing clear, flawless, and accurate information, the possibility of entering information easily and effectively, the ability to interact between the teacher and the learner in a positive, flexible, and less expensive manner. As put by [5], augmented reality provides an interactive environment that the real world has not provided.

Importantly, augmented reality provides an opportunity for students to interact with the educational content in general and in teaching and learning mathematics in particular because it includes abstract and complex scientific concepts, where augmented reality can simulate these concepts and embody the information and this is reflected in the development of thinking skills, especially visual thinking, which is one of the most complex types of thinking. This useful and practical aspect of augmented reality has been confirmed by previous studies, including [14] and [15]. Therefore, the development of thinking is one of the most important points that educational institutions focus on to achieve it among students by harnessing all their capabilities in developing thinking of students, which is reflected in students in the conscious and appropriate interaction with life and the circumstances surrounding it. Visual thinking is one of the types of thinking that is no less important than other types of thinking because it depends on the five senses that connect the person with the world around him and we cannot express it in writing [16]. As defined by [17], it is the merging of the senses with thinking to help in explaining a situation, and it is the greatest method of perception and has a direct impact on developing skills and acquiring competencies in teaching and learning. However, it is shown the effectiveness of augmented reality based on mobile learning in developing academic achievement and motivation in science subjects for fourth-grade students is on the rise [6]. As for the study [18], it evaluates student’s achievement and motivation during a high school augmented reality math activity that focuses on dimensional analysis. The use of augmented reality in mathematics and a combination of print and augmented reality also known as interactive printing are demonstrated. Participants in the quasi-experimental study are 61 students and the study instruments are pre and post-achievement tests. The results of the study have supported claims that the use of technology in a mathematics lesson and augmented reality increase student’s achievement and enhance the student’s motivation to learn mathematics and that the impact of technologies on the conceptual use of mathematics demonstrates the need for continued exploration to determine the impact of technologies not only on overall mathematical achievement but also on the specific type of conceptual mathematical activity.

The study aimed to benefit from the ability of virtual reality and augmented reality to visualize 3-dimensional shapes in mathematics where the study has investigated the possibility of using virtual reality and augmented reality techniques to teach the lesson of engineering solids to primary school children [19]. A 30-student sample randomly selected from fourth, fifth, and sixth grades is divided into three groups comprising a control group and two experimental groups. The results show that applying new technologies in teaching virtual and augmented reality improves interaction and students' interest in teaching mathematics, which contributes to increasing learning efficacy and understanding mathematical concepts when compared to traditional teaching methods.

This study identified the effect of augmented reality technology on the achievement of high school students in a unit in the science subject and their attitudes towards teaching the science subject [20]. To achieve the objectives of the study, the quasi-experimental approach is used, and the study is applied to a sample divided into two groups; an experimental group taught using augmented reality technology and a control group taught using the traditional method. The results show that students of the experimental group are more positive and happy about learning and want to continue to use augmented reality applications in the future. The results also conclude that the students show no signs of anxiety when using augmented reality applications. The trends of secondary school learners in Turkey towards augmented reality applications have been explored [21]. The study has also identified the relationship between the trend towards augmented reality and achievement in the solar system unit using augmented reality technology for 4 weeks. The results show that learners have positive trends towards augmented reality technology. They also show that there is a statistically significant relationship between the trend towards augmented reality and achievement. An analytical study of the results of (50) research papers in three databases, namely: Web of Science, Link Scopus, and Springer during 2008-2018 [22]. Factors such as the advantages, uses, challenges, and scope of augmented reality in the educational domain, and the positive or negative effects of its use in learning are analyzed to include students at various educational levels. After analyzing the results, it is concluded that the use of augmented reality for comprehensive education in mathematics is the area in which most studies have been conducted. The study recommends the necessity of conducting pieces of research in other areas of the importance of augmented reality in education.

B. Visual Thinking

Visual thinking is defined as one of the types of thinking that results from what we see around us, whether intentional or unintentional and it is the one that depends on the sense of sight because it is the means of communication that captures images and sends them to the mind and interacts with them, either by distinguishing, analyzing, interpreting or deducting. Visual thinking is constructed on experimental research into the way of thinking of learners by focusing on developing their abilities in changing and translating the visual language carried by the visual shape into verbal, written, or spoken language. This means that the relationship between the cognitive structure and the cognitive representation is a reciprocal relationship based on the influence and effect from the inside through which knowledge is strongly represented because visual representation supports visual thinking in terms of the ability to see the internal relationships of the presented shape and reveal the relative relationships in the dimensions of the shape and the development of inference skills [23].

In the same vein, the impact of augmented reality technology on teaching concepts related to the relationship between the earth and the sun to university students-majoring in geography has been identified [24]. The researchers have used the experimental approach on a 30-student sample from the University of Washington, where models designed with augmented reality technology are presented for the concepts under study. They have also employed worksheets as study instruments to evaluate the development of concepts among students applied pre and post the same group. The study shows that the students’ understanding of realistic concepts related to the relationship between the earth and the sun greatly improves due to the effectiveness of models designed with augmented reality technology. Also, the study shows that the use of this technology contributes to replacing false alternative perceptions of concepts with correct concepts.

The effect of augmented reality and its ability to facilitate the learning of chemistry for students has been revealed so that they can understand abstract concepts [25]. The study has employed the semi-experimental approach on a 96-student sample majoring in organic chemistry at the University of Washington, where they are divided into three groups: (26) students taught through books only, (26) students taught through augmented reality only, and (22) students taught through augmented reality in cooperative pairs. The study instruments are represented in the questionnaire, a chemistry self-efficacy measure, and an achievement test. The study concludes that the performance of a group of students taught using augmented reality only does much better than students taught without using augmented reality, and other students taught using augmented reality in cooperative pairs. The study recommends the need to support education with augmented reality technology in various educational stages other than the university stage.

C. Visual Thinking Skills

There is a multiplicity of visual thinking skills and differs from one study to another according to the nature of the educational situation, as the skills of visual thinking are demonstrated in the following: the skill of visual reading, the skill of visual discrimination, the skill of perceiving spatial relations, and the skill of visual closure [26]. It is asserted that the skills of visual thinking can be summarized as follows: the skill of describing and recognizing the shape, the skill of analyzing and interpreting the shape, the skill of correlating relationships in the shape, the skill of perceiving and clarifying ambiguity, and the skill of inferring meanings [27]. Abd Al-Reda and Fadel (2019) also maintain that the skills of visual thinking are demarcated into the skill of visual reading which is the ability to determine the dimensions and nature of the shape or image presented [28]. It is also the lowest skill of visual thinking skills, the skill of visual discrimination which is the ability to recognize the shapes or images and distinguish them from other shapes or images, the skill of scientific deduction which is the ability to attain new meanings and arrive at scientific concepts and principles through the presented shape, image, or map, and the skill of analyzing and interpreting the visual shape which is the ability of the individual to focus on minute details, pay attention to macro and microdata, and clarify the meanings of shapes, words, symbols, signs and figures, and approximate relations among them.

D. Visual Thinking Tools

At a time when the visual thinking skills are numerous, their tools are also numerous, as the modern digital tools that supported it help and work to capture ideas and organize information, as elucidated by the study of [29]. Visual symbols are likened to three tools: pictures which are one of the most accurate tools in communication, and they are a visual symbolic component of ideas and reality, symbols which are the most widely used and widespread tools in communication, and they are all indications of something and its proxy, and diagrams of shapes which illustrate an idea, are expressed in simple shapes or expressed in lines and include drawings related to pictures, concepts, and comics. Visual thinking is constructed on experimental research in the way of thinking of learners by focusing on developing their capabilities in changing and translating the visual language carried by the visual shape into a verbal language, written or spoken to develop communication skills and creative and logical thinking skills that achieve the learner's confidence in dealing with complex, ambiguous, and diverse opinions. Likewise, these skills enhance perception through discussions that take place across its processes to develop aesthetic practice [30].

Additionally, the ability to think visually overlaps with the skill of critical thinking that helps in solving problems and comprehend concepts, and at long last, visual thinking is seen as a more complex mental activity than the rest of the levels of thinking, as it depends on the representation of the displayed figure with symbols, diagrams, and pictures. Thus, visual thinking is one of the most important matters in the educational learning process as specified by [31] indicating the importance of teaching visual thinking skills to the learner because it works to raise the level of the learner's ability to communicate with others and increase mental ability. Visual thinking skills are regarded as a key to different types of thinking, including innovative thinking and critical thinking. The role of visual thinking skills in raising the level of motivation among the learners is unforgettable, and this is what encourages learners to teach and learn, works to develop science processes such as observation, interpretation, and analysis, helps them to self-learning that takes into account the individual differences between them and also assists to achieve objectives of science such as interpretation and prediction.

Both [28] indicate that there is great importance for visual thinking and its effect on the educational process, as it has replaced a lot of verbal information in pictures and visual shapes, the most important of which is the learner’s ability to develop visual language skills and comprehend all visual messages surrounding individuals inside the classroom, which indicates the scientific and technological development. Also, it can be emphasized that visual thinking is of great importance for people in general and for learners in particular in expanding their perceptions of understanding and imagination, and this helps them speed up the process of understanding and validating the information and facilitating its retrieval [32].

The educational theories in the area of educational technology have emphasized the importance of using visual thinking and its various tools such as images, shapes, diagrams, and graphs in the education process as the most important theory that focuses on the importance of using graphics in the education process is the theory of information processing. As put by [33] based on the learning styles used by the learner, thinking is classified into (Visual-Aural-Kinesthetic Model) which is auditory thinking, emotional thinking, and visual thinking. Visual thinking in the previous classification mentioned by [33], explained in detail by [34], and graphically represented by [35] helps in transferring the information to the human brain and processing inside it, and then expressing it in various methods. It turns out that the sense of sight controls the largest amount of information transferred to the brain, which indicates that thinking that depends on the sense of sight and is called visual thinking is the most used and the most important among the thinking styles in this classification. As defined by [36], visual thinking is an individual's skill to visualize and present an idea or information using pictures and drawings instead of the much stuffing and redundant wordiness used in communication.

Visual thinking skills are considered three main skills, namely vision, visualization, and drawing, where sub-skills branch out of these three main skills, and this is what [37] emphasized, as the three main skills are the origin of all skills and sub-skills change according to the types of sciences that belong to them. The visual sub-skills of computer science, for example, are different from the visual sub-skills of mathematics. Among the studies examining augmented reality and its effect on visual thinking is [38], which has investigated the effectiveness of a program based on augmented reality technology in developing visual thinking skills in science subjects for 9th-grade students in Gaza. Due to the nature of the study, the experimental method has been used. The study instrument is a one-group design with a pre-post measurement, where the study population consists of all the 6894 9th-grade students in schools of the Ministry of Education in the Gaza Strip. The Yarmouk Basic School (A) for boys in Gaza Governorate is randomly selected, and one class is also randomly selected, as the study sample consists of (43) students from the 9th-grade school. The study shows that there are statistically significant differences at the level of ($\alpha = 0.01$) between the average scores of the students in the pre and post applications on the visual thinking test. The study also shows that the employment of the program based on augmented reality technology achieves high effectiveness (Black's Modified Gain Ratio = 1.2) in developing visual thinking skills.

Against this, based on the importance of the primary stage among the stages of the educational process, the use of augmented reality technology has an effective effect on expanding the students’ perceptions of visualization and visual thinking. Therefore, the basic 10th-grade students are selected to teach them mathematics because this subject is the closest to reality, needs sensory experiences, and focuses on developing students’ thinking skills in general and visual thinking in particular. Intending to investigate the effect of augmented reality technology on visual thinking, augmented reality hopefully has great importance and effectiveness in the development of the educational learning process, as it is expected to provide an interactive learning environment filled with vitality, positivity, and excitement.

III. Research Methodology

Due to the nature of the study, the quasi-experimental approach is used for its suitability for the study. The independent variable (augmented reality technology) is subjected to test and its effect on the dependent variable (visual thinking) is measured for the students of the experimental group.

A. *Study Sample*

The study sample is purposefully selected from Jerash Model Schools of the Education Directorate of Jerash Governorate. The study sample consists of (57) male and female students randomly distributed into two sections: one of them is experimental taught using augmented reality technology and consists of (28) male and female students, and the second is the control group taught by the traditional method and consists of (29) male and female students.

B. *Study Instrument, Validity and Reliability*

The study instrument comprises a visual thinking test consisting of (20) items of a multiple choice of geometry unit in mathematics for the 10th grade of the academic year (2020/2021). This test is prepared and developed by researchers of the paper, where four skills are identified: visual reading skill, visual discrimination skill, scientific deduction skill, the skill of analyzing and interpreting the visual shape. A specification table is prepared and illustrated in Table I.

TABLE I. Specifications for the Visual Thinking Test

Skill	Number of Questions	Percentage (%)
Reading skill	5	25%
Visual discrimination skill	6	30%
Scientific deduction skill	5	25%
Skill of analyzing and interpreting the visual shape	4	20%
Total	20	100%

To verify the validity of the test (apparent and content), it is presented in its initial form to a group of 10 experienced and specialized validators, and their observations are strongly considered. Also, the reliability of the test is verified by applying the test to the reliability sample outside the study sample, consisting of (31) male and female students through the calculation of the Kuder-Richardson 20 (KR-20) reliability coefficient, and the Test-Retest Reliability coefficient, and the results of the analysis are shown in the table.

Table II shows that all the values of reliability coefficients are high, enhancing the accuracy of the instrument and its suitability for application to achieve the purposes of the study.

TABLE II. Results of Reliability Coefficients Analysis

Study Instrument	Calculated Reliability Coefficient	
	Kuder-Richardson 20 (KR-20)	Test-Retest
Achievement Test	0.923	0.931
Visual Thinking Test	0.961	0.934

IV. Results and Discussion

Results and discussion of the first question of the study “What is the effect of augmented reality on developing visual thinking in mathematics among 10th-grade students in Jerash governorate?”

To answer the first study question, the values of the arithmetic means and standard deviations of the performance of the two groups (experimental and control) in the pre and post-performance are calculated. Table III shows the results of the analysis.

Table III show that there is a noticeable convergence between the two groups in the pre-performance of visual thinking skills, while there are apparent differences between the two groups in the post-performance of visual thinking skills, as it is noticed that the value of the post arithmetic means of the performance of the experimental group students is (18.50) which is the highest compared to the post arithmetic mean of the performance of the control group students which is (11.50).

This result is due to the effect of augmented reality technology in developing visual thinking in the teaching of the geometry unit for several reasons, including the technology's ability to display the model and check it from all directions and its movement, which has helped students attain the approximate sense of realism. This technology also assists to attract students' attention, as it is accompanied by sounds, videos, colors, and 3-dimensional images enhanced with information and concepts related to pictures, which helps in developing visual thinking skills, allowing interaction with the application individually or collectively and giving them sufficient opportunity for visual thinking, and positively increasing their interaction. Importantly, the application helps in providing images of geometric shapes that are unavailable in the 10th-grade textbook, which broadens the perceptions of thinking in general among the students. The positive result of the current study in bringing about diversity and a paradigm shift in favor of the use of augmented reality technology in visual thinking is consistent with the results of many studies, including the studies of [24], [25], and [39] that have investigated the effectiveness of a program based on augmented reality technology in developing the students' achievement.

Results and discussion of the second question of the study "What is the effect of augmented reality on the development of each of the visual thinking skills (visual reading skill, visual discrimination skill, scientific deduction skill, the skill of analyzing and interpreting the visual shape) in mathematics among 10th-grade students in Jerash governorate?"

Table IV shows the values of the arithmetic means and standard deviations of the performance of the members of the experimental group in the pre-and post-test according to each skill.

It is evident from the results of the analysis of Table IV that the values of the arithmetic means of the performance of the members of the experimental group in the post-test are high (4.00 -5.20) in all skills compared to their performance in the pretest. The visual discrimination skill is ranked first with an arithmetic mean of (5.20) and a standard deviation of (0.76), while the visual reading skill is ranked second with an arithmetic mean of (4.73) and a standard deviation of (0.52). Also, the scientific deduction skill is ranked third with an arithmetic mean of (4.57) and a standard deviation of (0.50). As for the last rank, it is the skill of analyzing and interpreting the visual shape with an arithmetic mean of (4.00) and a standard deviation of (0.00).

To find out the indications of the differences between the arithmetic means of skills, the Analysis of Covariance (ANCOVA) is used as illustrated in Table V.

The results mentioned in Table V show that there is a difference in the post-performance between the two groups, where all the values of (F) are statistically significant at the level of ($\alpha = 0.05$). In detail, the difference is in favor of the experimental group, as the arithmetic means of their performance in the visual thinking skills test are higher compared to the arithmetic mean of the performance of the control group. The value of the practical significance of the combined visual thinking skills is (0.674). This indicates that (67.4%) of the variance between the two groups is due to the effect of the experiment. This result is to the fact that visual thinking is a mental capacity associated with the visual perceptual aspects, where it occurs when there is mutual harmony between what the learner sees in terms of shapes, drawings, and relationships, and what happens in terms of linking and mental output based on seeing the presented drawing and this is what the augmented reality technology focuses on, and this is what is confirmed by [23] and [26].

TABLE III. VALUES OF THE ARITHMETIC MEANS AND STANDARD DEVIATIONS OF THE PERFORMANCE OF THE TWO GROUPS IN THE PRE- AND POST-PERFORMANCE

Visual Thinking Skills	Group	Sample Size	Number of Items	Pre-Performance		Post-Performance	
				AM	SD	AM	SD
Total Performance	Experimental	30	20	11.43	7.30	18.50	1.43
	Control	30		11.07	7.09	11.50	3.25

TABLE IV. VALUES OF THE ARITHMETIC MEANS AND STANDARD DEVIATIONS ARRANGED IN DESCENDING ORDER DUE TO THE ARITHMETIC MEANS

	Pre-Performance		Post-Performance		Evaluation
Visual thinking skills	AM	SD	AM	SD	
Visual discrimination skill	1.43	0.86	5.20	0.76	High
Visual reading skill	1.70	1.49	4.73	0.52	High
Scientific deduction skill	1.50	0.78	4.57	0.50	High
The skill of analyzing and interpreting the visual shape	1.83	1.05	4.00	0.00	High
Total Performance	11.07	7.09	18.50	1.43	High

TABLE V. RESULTS OF THE ASSOCIATED ANALYSIS OF COVARIANCE (ANCOVA)

Source of Variance	Sum of Squares	Degree of Freedom	Average of Squares	Value of F	Level of Sig.	Eta=Squared
Visual Reading Skill	0.067	1	0.067	0.074		0.695
	117.544	1	117.544	129.679	0.000*	
	51.666	57	0.906			
	169.333	59				
Visual Discrimination Skill	1.328	1	1.328	1.203		0.624
	104.495	1	104.495	94.636	0.000*	
	62.939	57	1.104			
	170.933	59				
Scientific Deduction Skill	0.699	1	0.699	1.052		0.649
	69.952	1	69.952	105.294	0.000*	
	37.868	57	0.664			
	108.983	59				
The skill of analyzing and interpreting the visual shape	1.569	1	1.569	3.019		0.564
	38.4	1	38.4	73.869	0.000*	
	29.631	57	0.52			
	69.6	59				
Total Performance	11.957	1	11.957	1.931		0.674
	729.659	1	729.659	117.806	*0.000	
	353.043	57	6.194			
	1100.000	59				

* means a statistically significant function at the level of statistical significance (α = 0.05).

The way of thinking of students who have learned with augmented reality focuses on developing their abilities in translating the visual language embedded in the visual shapes presented in the lessons into verbal written or spoken language, developing communication skills that achieve the learner's confidence in dealing with complexity, ambiguity, and diversity of opinions, as well as increasing and enhancing perception through discussions that take place across their operations to develop aesthetic practice, and this will help in developing the visual thinking skills, and this is confirmed by [38] and [39]. The information offered in the study unit (geometry) in the textbook of the 10th grade has helped to develop the visual reading skill among the students of the experimental group represented by the ability to determine the dimensions and nature of this shape on the one hand. On the other hand, it has helped students to perceive the shape and distinguish it from other shapes and this represents the development of the visual discrimination skill, and this is confirmed by [17]. The development of the scientific deduction skill is represented by the ability of students to attain new meanings and scientific concepts and principles through the shape. Students learn, through this technology, to focus on the minute details, pay attention to geometric shapes and the subject of the study unit, and get geometric properties through their shapes. The augmented reality technology has its effect on the visual representation of ideas from shapes and drawings as an incentive for students to discover the meaning of the presented contents in the front of them, and this leads to better thinking, a progressive development towards creativity, and an indication of the conceptual developmental structure. Reading the visual shape aims to understand the meaning and includes understanding in reading the shape, linking, symbol, meaning, and organizing the read ideas. Reading the visual shape is a set of activities that allow the analysis of the information given in the form of conceptual connections in the presented shape, that is, a set of activities linking new information with previously acquired data stored in memory as these understanding models are closely related to the representation of the presented shape and or drawing.

V. CONCLUSION

The current study investigates the effect of augmented reality in improving visual thinking of 10th-grade students in mathematics in Jordan. In light of the results of the study, the objectives of the study, i.e. exploring the effect of augmented reality on developing visual thinking in mathematics among 10th-grade students in Jordan and examining the effect of augmented reality on the development of each of the visual thinking skills (visual reading skill, visual discrimination skill, scientific deduction skill, the skill of analyzing and interpreting the visual shape) in mathematics among 10th-grade students in Jordan have been achieved. Thus, the current study recommends using augmented reality technology in education in general, and in mathematics education in particular, organizing the content of mathematics curricula in line with augmented reality technology, which helps in developing visual thinking skills, paying attention to designing electronic lessons in mathematics and producing them using augmented reality applications, and holding training courses for mathematics teachers and encouraging them to learn how to integrate technology with education and develop visual thinking.

VI. Future Works

The area of augmented reality needs to be extended in other current works, with new approaches in the future work. This domain can be applied in other scientific areas such as physics or other detailed issues in mathematics.

Acknowledgments

The author is grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

References

[1] Catenazz, N. & Sommaruga, L. (2013). Social media: challenges and opportunities for education in modern society, mobile learning and augmented reality: new learning opportunities, International Interdisciplinary Scientific Conference, Vol. 1 No1.

[2] Christie, Rodgers. (2014). Augmented reality books and the reading motivation of fourth-grade students, Union University School of Education, UMI. ProQuest LLC.

[3] Klopfer, E. (2008). Augmented learning: Research and design of mobile educational games. Cambridge, MA: MIT Press.

[4] Rabia M. Yilmaz and SevdaKucuk, YukselGoktas. (2015). Are augmented reality picture books magic or real for preschool children aged five to six? British Educational Research Association. Journal of Educational Technology and Society, 2(7). 22-36.

[5] S. Cuendet, Q. Bonnard, S. Do-Lenh, and P. Dillenbourg. (2013). Designing augmented reality for the classroom. Computers and Education, 68(2), 557–569.

[6] Chiang, T., Yang, S. & Hwang, G. (2014). An augmented reality-based mobile learning system to improve student's learning achievements and motivations in natural science inquiry activities. Educational Technology & Society, 17 (4), 352-365.

[7] Wang, J. & Hartley, K. (2003). Video technology as a support for teacher education reform. Journal of Technology and Teacher Education, 11(1), 105-138.

[8] Derry, S.J. (2007). Guidelines for Video Research in Education: Recommendations from an Expert Panel. Chicago: Data Research and Development Center.

[9] Glockner, H., Jannek, K., Mahn, J., and Theis, B. (2014). Augmented reality in logistics: Changing the way we see logistics: A DHL perspective.

[10] Bower, M., Howe, C. (2018). Augmented reality in education. Educational Media International, 51(1), 15-1.

[11] M. AKC, Ayır and G. Akc, Ayır. (2017). Advantages and challenges associated with AR for education: A systematic review of the literature. Educational Research Review, 20(3).

[12] Kulkarni, S. & Takawale, N. (2016). Comparative Study of Augmented Reality and Virtual Reality. International Journal of Innovative Research in Computer and Communication Engineering, 4 (1): 1-6.

[13] Liarokapis, F., Anderson, E. (2010). Using augmented reality as a medium to assist teaching in higher education. In Proceedings of the 31st Annual Conference of the European Association for Computer Graphics (Euro graphics 2010), Education Program (pp. 9-16). Norrköping, Sweden: Euro graphics Association.

[14] Ivanova, M, & Ivanov, G. (2011). Enhancement of learning and teaching in computer graphics through marker augmented reality technology. International Journal on New Computer Architectures and Their Applications, (IJNCAA), 1(1), 176-184.

[15] Núñez, M., Quirós, R., Núñez, I., Carda, J. B., Camphor, E., Mauri, J. L. (2008). Collaborative augmented reality for inorganic chemistry education. WSEAS International Conference. Proceedings-Mathematics and Computers in Science and Engineering. Hera lion, Greece: ACM Digital Library.

[16] Dilek, G. (2010). Visual thinking in teaching history: Reading the visual thinking skills of 12-year-old pupils in Istanbul. Education, 38(3), 257-274. ERIC Document Reproduction Service No. EJ898020.

[17] Carrascals. S. (2019). Acquisition of competencies for sustainable development through visual thinking. A study in rural schools. Sustainability, 11(8), 23-37 Https://doi.org/10.3390/su11082317.

[18] Estapa, A.T. and Nadolny, L.N. (2015). The effect of augmented reality enhanced mathematics lessons on student achievement and motivation. Journal of STEM Education: Innovations and Research, 16(3), 40-48. ERIC Document Reproduction Service No. EJ1078415.

[19] Demitriadou, E., Stavroulia, K.E. and Lanitis, A. (2019). Comparative evaluation of virtual and augmented reality for teaching mathematics in primary education. Education and Information Technologies, 25, (2), 381-401. Https://doi.org/10.1007/s10639-019-09973-5.

[20] Sahin, D. and Yilmaz, R.M. (2020). the effect of Augmented Reality Technology on middle school students' achievements and attitudes towards science education. Computer Education, 144(2), 10-22. Https://doi.org/10.1016/j.compedu.2019.103710.

[21] Sirakaya, M., & Cakrnak, E. K. (2018). Investigating student's attitudes toward Augmented Reality. MOJET: Malaysian Online Journal of Education Technology 1(6).

[22] Quintero J, Molera C, Juamperez J, Redecillas S, Meavilla S, Nuñez R, García-Volpe C, Del Toro M, Garcia-Cazorla Á, Ortega J, Segarra Ó, de Carpi JM, Bilbao I, Charco R. (2019). Augmented reality in educational inclusion. A systematic review on the last decade; 10: 1835. Doi: 10.3389/fpsyg.2019.01835.

[23] Saqr, Nasih, Hussein Salem. (2018). the effectiveness of using visual thinking networks in the development of visual thinking skills and academic achievement in mathematics for children with learning difficulties at the primary stage. Journal of Educational Sciences, 26(1), 210-247.

[24] Shelton, B., & Hedley, N. (2010). Using augmented reality for teaching earth-sun relationships to undergraduate geography students. In The - 476 - First IEEE International Augmented Reality Toolkit Workshop, Darmstadt, Germany, IEEE Catalog Number: 02EX632 ISBN: 0-7803-7680-3. DOI: 10.1109/ART.2002.1106948.

[25] Chen, Y. (2013). Learning protein structure with peers in an ar enhanced learning environment. [Ph.D. Thesis]. The University of Washington. The USA.

[26] Suleiman, Mohamed El-Sayed. (2018). the effectiveness of a multimedia program based on the systemic approach according to the Davis model in developing visual thinking skills and cognitive achievement among hearing-impaired students. International Educational Specialized Journal, 7(4), 1-21.

[27] Jabr, Yahya Saeed. (2010). the effect of employing the strategy of the metacognitive learning course on the development of concepts and visual thinking skills in the sciences for the tenth primary students. [Unpublished Master Thesis]. College of Education, Islamic University, Gaza, Palestine.

[28] Abd Al-Reda, Athraa Abd al-Rahim, and Fadel, Sri Muayad. (2019). Visual thinking among students at College of Education of Waist University. The 11th International Scientific Conference, April/2019 at Waist University, Iraq.

[29] Abdulaziz, Safwat Hassan. (2018). The effect of using info graphics in teaching science on achievement and development of visual thinking skills and attitude towards them among elementary school students in the State of Kuwait. Concept Journal of In-depth Psycho-Philosophical and Humanistic Studies, 2(2), 42-63.

[30] Haciomeroglu, E. & Chicken, E. (2012). Visual Thinking and Gender Differences in High School Calculus. International Journal of Mathematical Education in Science and Technology, 43(3), 303-313. ERIC Document Reproduction Service No. EJ992909.

[31] Razooqi, Raad Mahdi and Abdul Karim, Soha Ibrahim. (2015). Thinking and its Patterns, 1st Edition Amman: Dar Al-Masirah.

[32] Genovesi, J. S. (2011). An exploratory study of a new educational method using live animals and visual thinking strategies for natural science teaching in museums. Ph.D. Dissertation]. Drexel University, ProQuest.

[33] Sword, L. (2005). The Power of Visual Thinking. Gifted and Creative Services Australia, Retrieved Dec 26, 2020 from http://www.starjump.com.au/.

[34] Thomas F. Hawk & Amit J. Shah. (2007). Using Learning Style Instruments to Enhance Student Learning, Decision Sciences Journal of Innovative Education, Vol 5 NO. 1. DOI: 10.1111/J.1540-4609.2007.00125.X.

[35] Eicher, J.; Johns, J.; & Bearley, W. (2009). Neuron-Linguistic Communication Profile Online. HRDQ Assessment Center. Retrieved Oct 25, 2020, from http://www. hrdqstore.com/assets/images/products/NCP/NeurolinguisticCommunication-Profile-Online-Assessment-Sample-Report.pdf.

[36] Wileman, R. E. (2000). Visual Communicating, Englewood Cliffs, N-Educational Technology Publication, Biscoe Electronic.

[37] Grandin, T. (2006). Thinking in Pictures. Retrieved from http://www.grandin.com/inc/visual.thinking.html on Oct 25, 2020.

[38] Ahmad, Awadallah. (2016). the effectiveness of a program based on augmented reality technology in developing visual thinking skills in science subjects among ninth-grade students in Gaza. [Unpublished Master Thesis]. Al-Azhar University, Gaza, Palestine.

[39] Issa, Sami Abdul Hamid Muhammad. (2018). The use of mobile augmented reality technology with various support patterns (fixed/flexible) in developing some visual thinking skills among middle school students. Education Technology: Studies and Research, 37, (2), 151-193.

Speeding up an Adaptive Filter based ECG Signal Pre-processing on Embedded Architectures

Safa Mejhoudi[1], Rachid Latif[2]
Amine Saddik[3], Wissam Jenkal[4]
Laboratory of Systems Engineering and Information Technology, ENSA, Ibn Zohr University, Agadir, Morocco

Abdelhafid El Ouardi[5]
SATIE, Digiteo Labs
Paris-Saclay University
Orsay, France

***Abstract*—Medical applications increasingly require complex calculations with constraints of accelerated processing time. These applications are therefore oriented towards the integration of high-performance embedded architectures. In this context, the detection of cardiac abnormalities is a task that remains a high priority in emergency medicine. ECG analysis is a complex task that requires significant computing time since a large amount of information must be analyzed in parallel with high frequencies. Real-time processing is the biggest challenge for researchers, when talking about applications that require time constraints like that of cardiac activity monitoring. This work evaluates the Adaptive Dual Threshold Filter (ADTF) algorithm dedicated to ECG signal filtering using various embedded architectures: A Raspberry 3B+ and Odroid XU4. The implementation has been based on C/C++ and OpenMP to exploit the parallelism in the used architectures. The evaluation was validated using several ECG signals proposed in MIT-BIH Arrhythmia database with a sampling frequency of 360 Hz. Based on an algorithmic complexity study and a parallelization of the functional blocks which present significant workloads, the evaluation results show a mean execution time of 7.5 ms on the Raspberry 3B+ and 0.34 ms on the Odroid XU4. With an efficient parallelization on the Odroid XU4 architecture, real-time performance can be achieved.**

Keywords—ECG signal denoising; ADTF algorithm; OpenMP programming; embedded architectures

I. Introduction

ECG is an essential element in the diagnosis and the detection of cardiovascular disease or also in the monitoring of patients [1]. However, it is often correlated with different types of noise, which generates a distortion of the signal and a loss of valuable information. To simplify the interpretation task, several processing and filtering algorithms are proposed in the literature [2-6]. Digital Filters (FIR and IIR), Empirical Mode Decomposition (EMD), Wavelet Transform denoising techniques as Discrete Wavelet Transform (DWT) and the Adaptive Dual Threshold Filter (ADTF) [7-8].

Using digital filters, some useful information in the signal can be affected, particularly the R wave [3,9]. The EMD and the DWT give satisfying results, but they are characterized by their algorithmic complexity that requires more hardware resources and important computation time [10-11]. The ADTF proposed by W. Jenkal et al. in [7] has a great denoising capacity, especially when the signal is mixed with the high-frequency noises. The advantage of this technique is the very low complexity. Hardware implementation of this algorithm on FPGAs is presented in [8] using Hardware Description Language (VHDL). The author divided the algorithm into 3 main blocks: The first consists of Real-time data loading module with an acquisition frequency of 360 Hz as he uses signals from MIT-BIH database. The second is used for ADTF features calculation. The third is for Test and Assignment. To respect real-time processing, a frequency of 3.6 kHz is used in the second and third blocks, which is ten times greater than the acquisition frequency. This implementation indeed respects real-time, but this is related to the processing frequency which is fixed by the author not given by the architecture.

Abbreviations

ADTF: Adaptive Dual Threshold Filter
DWT: Discrete Wavelet Transform
DT-WT: Dual-Tree Wavelet Transform
ECG : Electrocardiogram
EMD : Empirical Mode Decomposition
EEMD : Ensemble Empirical Mode Decomposition
EEMD-GA: EEMD and Genetic Algorithm
FIR: Finite Impulse Response
FPGA: Field Programmable Gate Array
GPU: Graphics Processing Unit
Ht: Higher Threshold
IIR: Infinite Impulse Response
IMF : Intrinsic Mode Functions
LA : Left Atria
Lt: Lower Threshold
LV: Left Ventricle
MSE: Mean Square Errors
PRD: Percentage Root-mean-square Difference parameter
RA: Right Atria
RV: Right ventricle
SNR: Signal to Noise Ratio
VHDL: VHSIC Hardware Description Language
WGN: White Gaussian Noise

This work deals particularly with the real-time ECG denoising. Real-time systems differ from others by considering time constraints and compliance, which is as important as the result's accuracy. In other words, the system should not just deliver exact results; it should also provide them within a set timeframe. This notion takes on its importance when it comes to human health. So, the design of processing algorithms that meet these constraints is therefore essential. In this context, the processing system is considered a real-time system. It must make the necessary correction for sample *n* before the arrival of sample *(n + 1)*. This constraint implies rigorous requirements in terms of performance and speed. In this work, different signals from the Physionet MIT-BIH arrhythmia database are used, with 360 Hz of sampling frequency to test the reliability of the algorithm. This means that the time constraint is of the order of 2.77 ms. The Matlab simulation gives a processing time average of 150ms for one sample, which is too far from being in real-time.

In order to ensure real-time processing, OpenMP parallel programming is used, which gives excellent results compared to C/C++ naive implementation and Matlab implementation. So, for Desktop with OpenMP programming, just 0.34 ms is needed to process one sample. But desktop is not always a good solution in biomedical monitoring because of its size, weight, and power consumption, especially when the objective is the home monitoring to facilitate access to care for the elderly or in regions lacking medical personnel, prevent hospitalizations and improve patient control and quality of life.

We opted to use Low-cost embedded architectures such as XU4 and Raspberry. The given results show that the optimal choice is the XU4 board with an average processing time of 2.34ms instead of 7.5 ms in the case of Raspberry 3B, which greatly satisfies the time and energy consumption constraints.

The paper is formulated as follow:

1) The first section exposes an overview of ECG signal denoising techniques and related work.

2) The second section depicts a detailed evaluation of the ADTF algorithm and describes its implementation.

3) The third section puts on view the results and discussion of the processing performance of the embedded implementations on different embedded architectures.

4) Lastly, conclusion and future work are the objects of the fourth section.

II. ECG Signal Processing: An Overview

A. *ECG Signal Processing*

The ECG signal or electrocardiogram is a widely used exam in cardiology field. It describes the electrical activity of the human heart and has a high clinical value for diagnosing cardiac arrhythmias. From the ECG signal processing, several parameters can be extracted. As a rule, different waves' durations and shapes can indicate some cardiac abnormalities [3]. Fig. 1 represents the formation of the ECG signal, which reflects the different deflections and contractions of the heart muscles, making it possible to diagnose a patient's cardiac state. As shown in the Fig. 1, RA, LA, RV, and LV represent respectively the right and left Atria and ventricle.

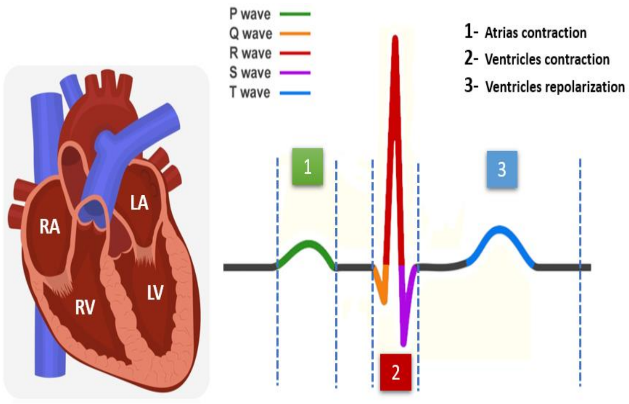


Fig. 1. ECG Signal Formation.

To make the best use of ECG data in large quantities, intelligent diagnostic systems have appeared. These systems can improve the quality of the signal; extract useful information, and offer a diagnosis that can help doctors make the right decisions.

The biomedical engineering revolution forces researchers to enhance the automatic diagnosis by optimization of ECG processing algorithms in order to ensure real-time monitoring of cardiac data [12-16]; and their implementation on embedded systems as recent technological resources [8,17-19].

The ECG signal is characterized by low frequency and a small amplitude. So, it is often affected by various kinds of noise as interference due to electrical appliances, high-frequency noises produced by muscles activity, and low-frequency noises of body movements in relation with respiration, which distorts its morphology, resulting a wrong diagnostic of the heart state of the patients [5,20]. To overcome this problem, the ECG signal must first go through a precise and effective preprocessing step.

The preprocessing step aims to remove or reduce the different noises. In this context, many methods are used such as Digital Filters (FIR/IIR) [21-22], Empirical Mode Decomposition (EMD) and Ensemble EMD (EEMD) based techniques [10, 23-25], Discrete Wavelet Transform (DWT) [26-27, 3], Dual-Tree Wavelet Transform (DT-WT) [28] and Adaptive Filtering [29,8].

B. *Related Works based ECG Denoising*

Digital filters are represented by Finite Impulse Response (FIR) and Infinite Impulse Response (IIR). They are used to denoise ECG signals. Their names are originally linked to the mathematical definition, and their expressions are given respectively by (1) and (2):

$$Y(n) = \sum_{k=0}^{N-1} \mathrm{b(q)X(n-q)} \qquad (1)$$

$$y(n) = \sum_{k=1}^{M} a_k\, y(n-k) + \sum_{k=0}^{N} b_k\, x(n-k) \qquad (2)$$

The implementation of FIR filters can be done without feedback as it is shown in Fig. 2 where X(n) presents the input signal, Y(n) is the filtered signal, Z^{-1} operator is a delay in the Z transformation, N is the filter order, and b_q are the coefficients of the filter transfer function. Various windowing techniques are used, as example: Rectangular window, Kaiser window, Hamming window, Hanning window and Blackman window. IIR filters are designed using filters as Chebyshev filter, Butterworth filter, Inverse Chebyshev filter [30]. The major difference between them is that FIR filters are stable for any input signal. However, IIR filters can alter to unstable due to the feedback as shown in Fig. 3. Where a_n and b_m are the filter transfer function coefficients.

Most of the works used FIR or IIR filters by selecting a bandwidth related to the utile data from the ECG signal. From different papers, it can be deduced that that FIR filter with Kaiser Window eliminates noises from ECG signal with less alteration in the waveform [4,20]. FIR filters' problem is the high computational due to the number of coefficients needed to achieve excellent denoising result and a group of delay in response, which is the main challenge in real-time systems [4].

Wavelet methods have proven to be more common and effective than FIR/IIR filters [31]. Wavelet methods simultaneously characterize time and frequency information. They decompose the signal to different resolutions using low pass filters (H[n]) to get the approximations (A) and high pass filters (g[n]) to get the details (D) as it's explained in Fig. 4. The ECG signals are denoised using thresholding techniques. But they have some limitations since they reduce the signal amplitude, which can affect the R waves.

To overcome this limitation, methods based on EMD are used, the signal is disintegrated into a sequence of intrinsic mode functions (IMFs), and noisy IMFs are removed, but this technique can remove some useful information when eliminating the noisy IMFs. EEMD is used to overcome this problem by removing the mode-mixing [32].

To deal with the problems of complexity in ECG denoising, W. Jenkal et al. developed a new approach inspired from image denoising [33]. This approach is an adaptive dual threshold filter which is dedicated to removing high-frequency noises [34-35]. This method aims to compute three elements for a selecting window (the average of the window, the higher threshold, and the lower threshold) and then the correction of the window's median value using the thresholding. The process is explained in the following block diagram shown in Fig. 5.

The performance evaluation of this method is made in [8] based on the SNRimp result comparison between the ADTF and techniques based on EEMD. This evaluation shows that the ADTF gives very good results compared to the EEMD denoising algorithm presented in [36] and a competitive SNRimp results to the enhanced EEMD method (EEMD-GA) published in [37].

The main characteristic of the ADTF algorithm is its low complexity compared to the cited methods. The ADTF presents a linear complexity C(n) depending just on the signal size n. A comparative study of the complexity between the EMD, the EEMD, and the ADTF methods is presented in [8]. The conclusion of this study shows that ADTF presents a low complexity, unlike EMD and EEMD, which is related to various parameters, namely, the length of the signal, the number of the noisy signals, the number of IMFs, as well as the number of sifting processes. Comparing with the DWT, it also has linear complexity, in the manner of EMD/EEMD, it's related to other parameters not only the size of the signal, we are talking about the wavelet mother's coefficients, the number of decomposition's level and also the thresholding technique [38].

In the next section, a detailed study of the ADTF is presented as well as simulation results.

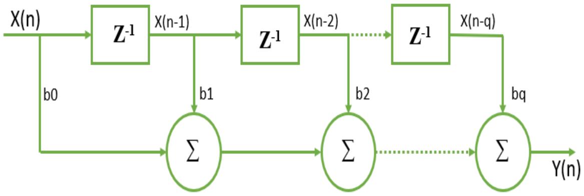


Fig. 2. FIR Filter Conception.

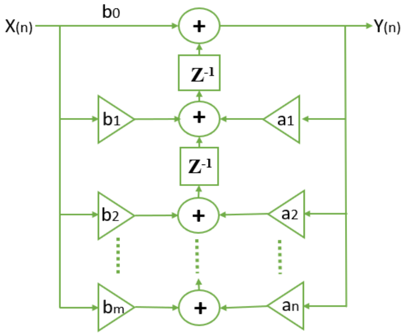


Fig. 3. IIR Filter Conception.

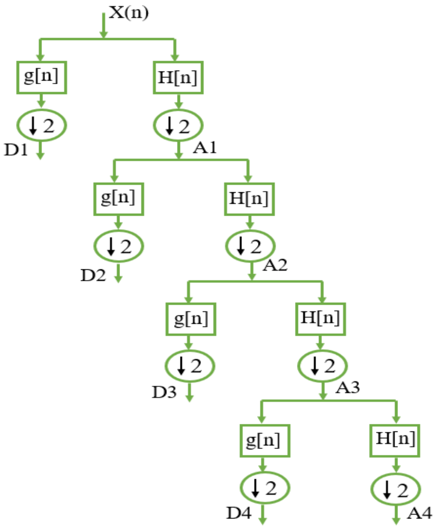


Fig. 4. DWT Decomposition.

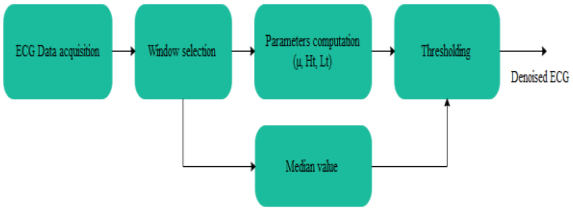


Fig. 5. ADTF Algorithm Overview.

C. The ADTF Technique

As described above, the ADTF algorithm aims to compute the average of the selected window (μ), the lower threshold (Lt), and the higher threshold (Ht). Hereinafter, they are presented respectively by (3), (4), and (5).

$$\mu = \frac{1}{W}\sum_{i=n}^{n+W} \mathrm{Input(i)} \quad (3)$$

$$\mathrm{Lt} = \mu - [(\mu - \mathrm{Min}) * \alpha] \quad (4)$$

$$\mathrm{Ht} = \mu + [(\mathrm{Max} - \mu) * \alpha] \quad (5)$$

Where the window size is W, Input (i) is the noisy signal, and n presents the signal length. Min is the minimum value of the window, Max is the maximum value of the window, and α is the thresholding coefficient with:

$$0 < \alpha < 1$$

α is essentially used to adapt the thresholding. According to the filtering process, α varies between 0% and 100%, lower values are recommended for a high concentration of noise, and higher values are tolerated in the opposite case [7].

The window selection is not arbitrary; it's used to compare the median sample to his left and right regions, as stated in [7], a window of five samples gives the best results in terms of MSE and SNRinput.

Fig. 6 shows the result of the ADTF filtering applied on the signal n° 234 from the MIT-BIH Arrhythmia database corelated with white Gaussian noises of 5 dB, using a window of 5 samples, an α coefficient of 5%. The compilation is done using Matlab R2019a.

The algorithm validation is done using Matlab coding [7], but this remains a functional validation, while in biomedical engineering, Real-time processing is required in most cases, especially ECG signal processing. The time constraint is related to the sampling frequency of the used signals. In the MIT-BIH Arrhythmia database, the signals are sampled 360 Hz which leads to an interval of 2.7 ms between the samples. Using Matlab coding, 150 ms is needed to process and correct each sample. Thusly, with Matlab implementation, the system is too far from being in real-time.

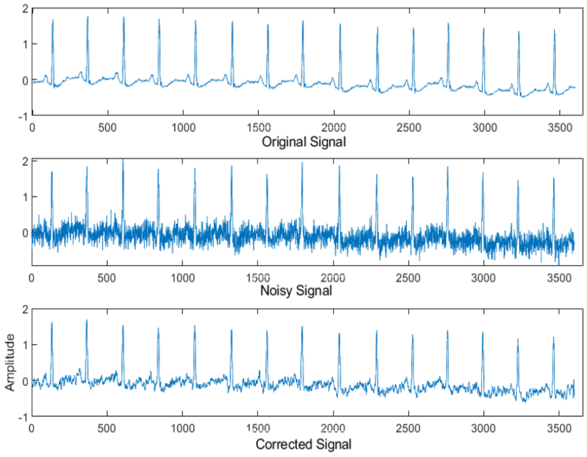


Fig. 6. The Denoising Results of the Signal n°234 of the MIT-BIH Arrhythmia Database Correlated with 5 dB of the WGN.

We have opted for C/C++ optimization using OpenMP parallelization in order to optimize the code. As follows a comparative study of the obtained results using Matlab, C/C++ non-optimized algorithm and OpenMP optimized algorithm on different architectures.

III. Results and Discussion

A. System Specification

In this paper, ECG signals from MIT-BIH Arrhythmia, the international database Physionet, which incorporates 48 half-hour records are used. These signals are converted to numerical values at 360 Hz with a resolution of 11-bits. Different signals with additive White Gaussian Noise (WGN) at SNR levels of 10dB and 20dB are used. Table I presents the used signals for both the Matlab and the C/C++ validation.

The validation of the algorithm was performed using Matlab. In order to ensure a reliable and accurate real-time processing, OpenMP programming is used in the implementation to ensure parallel programming on a shared memory multiprocessor system. The parallelism is achieved by creating a set of threads; these treads execute independently and simultaneously the appointed tasks. Using OpenMP, the program could be optimized to evaluate processing times using three different architectures.

The work is based on Raspberry 3B and XU4 architectures; the choice of these architectures was based on the low energy consumption and also a low weight for embedded application. Despite the desktop giving excellent processing time results, but it's not a good solution for biomedical monitoring because of its size and portability in the real case. Good results can also be given using TX1, TX2, and AGX Xavier boards; notably, with GPU part, we can decrease the processing time [39]. But the major problems are their cost and power consumption. Raspberry and XU4 present a low-cost embedded system with low power computation [40]. The rest of this study will allow to make the right decision by comparing the found processing times for the two architectures. Table II presents the used architectures specifications.

TABLE I. ECG Signal Records

Signal number	**The corresponding Signal from MIT-BIH**
1	Record n° 100
2	Record n° 100 + 10 dB of WGN
3	Record n° 100 + 20 dB of WGN
4	Record n° 101
5	Record n° 101 + 10 dB of WGN
6	Record n° 101 + 20 dB of WGN
7	Record n° 103
8	Record n° 103 + 10 dB of WGN
9	Record n° 103 + 20 dB of WGN
10	Record n° 113
11	Record n° 113 + 10 dB of WGN
12	Record n° 113 + 20 dB of WGN

TABLE II. DESKTOP, RASPBERRY AND XU4 SPECIFICATIONS

Type	Desktop	Raspberry	XU4
Processor	Intel® Core™ i5-4200M	Broadcom BCM2837B0	Exynos 5422 big. LITTLE
Cores	Quad	Quad	Octa
CPU	I5 4200M	ARM Cortex-A53 (ARMv8)	ARM Cortex A15/A7
GPU	HD Graphics 4600 /AMD Radeon R5 M230	Broadcom Videocore-IV	Advanced Mali
Support Language	C/C++/OpenCL/ OpenGL	C/C++	C/C++/OpenCL/OpenGL
Frequency	2.50 GHz	1.4GHz	2GHz/1.4GHz
Weight	2.6 kg	50 g	60 g
Energy	90 w	15.5 w	5W
Dimensions	377 x 250 x 34 mm	85 x 56 x 17 mm	82 × 58 × 22 mm

B. Experimental Results

To evaluate the denoising performance of the C/C++ code of the ADTF algorithm, Mean Square Errors (MSE), Percentage Root-mean-square Difference parameter (PRD), and Signal to Noise Ratio (SNR) are computed for 12 records of ECG of 10s from the MIT-BIH arrythmia database. Their expressions are presented respectively by (6), (7), and (8).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\big(Input(i) - Output(i)\big)^2 \qquad (6)$$

$$SNRout = 10 \times log10\left(\frac{\sum_{i=1}^{N}(Input(i))^2}{\sum_{i=1}^{N}(Output(i)-Input(i))^2}\right) \qquad (7)$$

$$PRD = \sqrt{\frac{\sum_{i=1}^{N}(Input(i)-Output(i))^2}{\sum_{i=1}^{N}(Input(i))^2}} \times 100 \qquad (8)$$

Where input(i) represents each input sample and output (i) the filtered sample, N is the size of signal.

The experimental results in Fig. 7, 8, and 9 show that the C/C++ program provides concrete denoising than the Matlab code in terms of good SNR with less MSE and PRD.

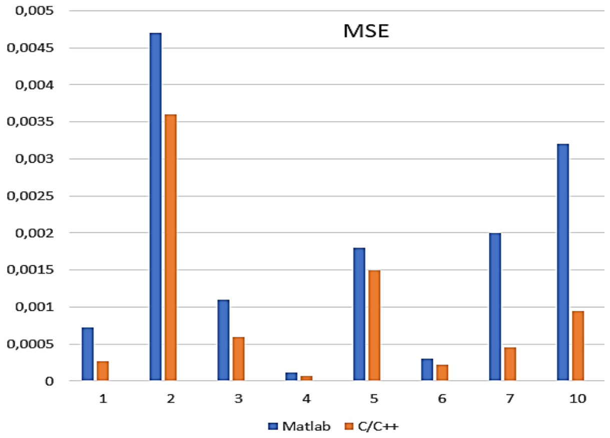


Fig. 7. MSE Comparison of Denoised Signals using Matlab and C/C++.

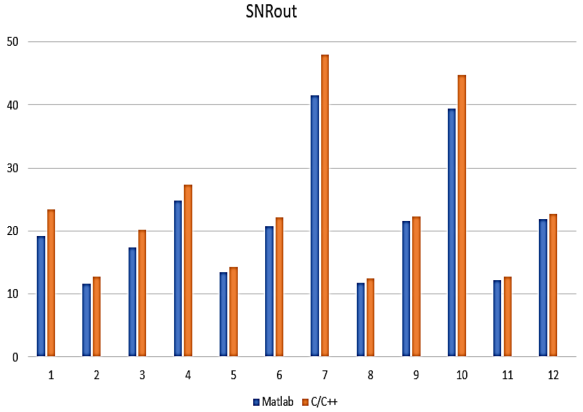


Fig. 8. SNRout Results Comparison of Denoised Signals using Matlab and C/C++.

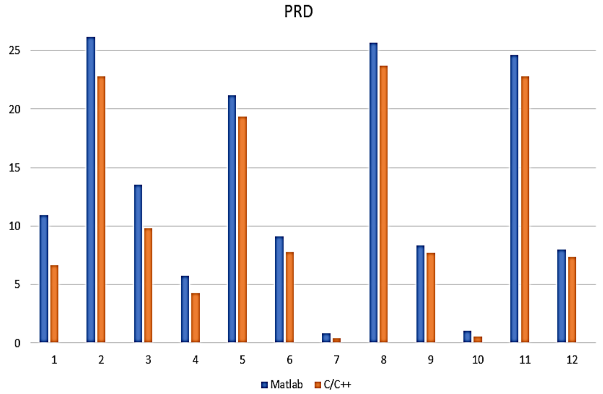


Fig. 9. PRD Comparison of Denoised Signals using Matlab and C/C++.

As shown, the C/C++ gives better performance in noise reduction. As follows, a comparison study of the execution time on Desktop is presented in Table III. It presents the needed time to process each sample of the signal based on nine iterations of time average calculation. The obtained results show that using OpenMP on the desktop, the results showed a ×4 speed-up compared to the time obtained with the naive C/C++ implementation. In Fig. 10, the execution time of Matlab is added to visualize the interest of parallel programming in this case.

TABLE III. EXECUTION TIME IN DESKTOP

C++		Open MP	
Iteration	*Time (ms)*	*Iteration*	*Time (ms)*
1	1.21	1	0.25
2	2.01	2	0.4
3	0.9	3	0.133
4	1.47	4	0.36
5	1.85	5	0.41
6	1.81	6	0.408
7	2.04	7	0.405
8	1.81	8	0.365
9	0.94	9	0.405
Average	**1.56**	**Average**	**0.348**

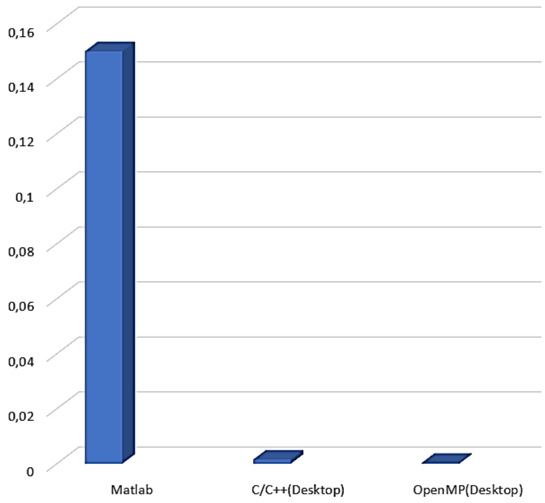


Fig. 10. Processing Times on Desktop using different Tools.

C. Performance Evaluation using Embedded Architectures

Performance evaluation implementing embedded architectures is the main of this work. After proving the algorithm performance in noise reduction, a reliable real-time implementation is necessary.

The Matlab implementation was used for the algorithm validation and results interpretation, but the very high execution times lead us to exclude any software optimization in order to speed up processing in this implementation. Then we opted for C/C++ programming. The implementation is done, as cited above using a Desktop, a Raspberry 3B and XU4 boards. The C/C++ implementations results are too much better than those given by Matlab implementation, but they stay a little far from the real-time as it appears in Fig. 11, with 1.56 ms for Desktop, 9.2ms for Raspberry and 6.8 ms for XU4.

The block diagram of the parallelized algorithm is presented in Fig. 12. The first block is the ECG signal acquisition which is not the subject of this paper. The second aims to divide to the input signal by the thread number (A). Here a test is done; if the number of signal samples is not divisible on A the system searches for the optimal signal size by adding a few samples at the end of the signal. The added samples can be calculated using the values of the last window of the signal, an average of the window can be calculated to replace missing samples. The third block consists of memory allocation and parameters initialization as the α coefficient and the window size.

Block 4 is the core of this work; it aims to execute the denoising procedure. This is where the parallelism is applied, ECG signal is divided by the number of threads, each thread runs the denoising program on a portion of the signal instead of the whole signal. The last blocks present the denoised signal's exploitation step either for additional processing, storage, display, or transmission.

To optimize the given execution times, OpenMP is used. Fig. 13 depicts the pseudo code of the OpenMP-based parallel computation algorithm.

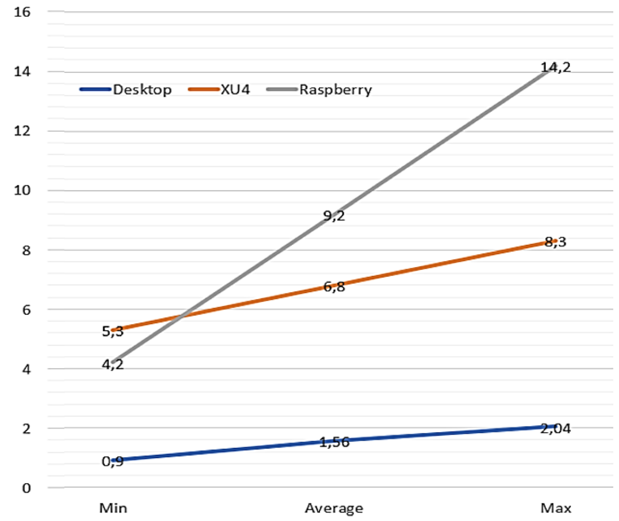


Fig. 11. Min and Max Processing Times using different Architectures.

The first step is to determine the optimal size that can give a divisible value over the number of threads. Thereafter, we set the input and output signals that will be written in the output file. This output file will be plotted subsequently using Matlab in order to display the errors and compute the different evaluation metrics.

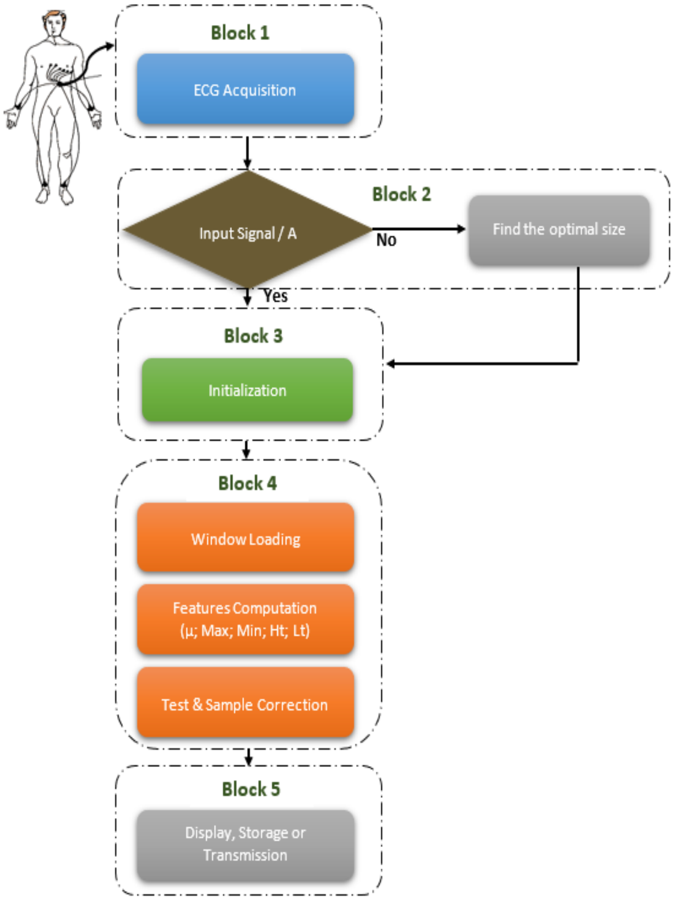


Fig. 12. The Algorithm Block Diagram.

```
ALGORITHM
#define NUMTHREADS
#define NAME_FILE "INPU_ECG"
#define NOM_SORTIE1 "OUTPUT_Signal"
#define Alpha
Input:  ECG Signal.txt
        Find the optimal signal size
        X(i)  <=  Input(i)
        i=1;
        α, W Initialisation
        Threads = NUMTHREADS
        #pragma omp parallel for shared (Min, Max,
         µ, Ht, Lt,) num_threads (threads)
          for i=1 to N-W
              Compute Min, Max, µ, Ht and Lt
              If X (i+W/2) > Ht Output (i+W/2) =Ht;
              If X (i+W/2) < Lt Output (i+W/2) =Lt;
              Else
              Output(i+W/2) = Input(i+W/2)
              i= i+1;
          End for
        Output(N-1) = X(N-1)
        Output(N) = X(N)
End.
```

Fig. 13. Pseudo Program of the Algorithm using OpenMP.

Fig. 14 shows processing time using OpenMP implementation. A time of 7.5 ms is achieved for one sample processing using Raspberry architecture, 2.34 ms using XU4 architecture, and 0.34 ms using the desktop. The time constraint posed by the acquisition system forced us to process each sample with a delay less than 1/360Hz, which implies trying to process each sample within 2.77ms. The results allowed to eliminate the choice of raspberry due to the processing times, which exceed 2.77ms. Despite their low energy consumption and weight, the time evaluation has shown that this architecture cannot process the algorithm in real-time. The desktop gave a very low processing time, 0.34ms, which shows the desktop's high performance, but the drawback here is the high-power consumption, which makes this type of system does not meet the reliability requirement. On the other hand, the XU4 architecture met the time constraint, making it the best choice for this application. In addition, its low power consumption and low weight confirm the choice.

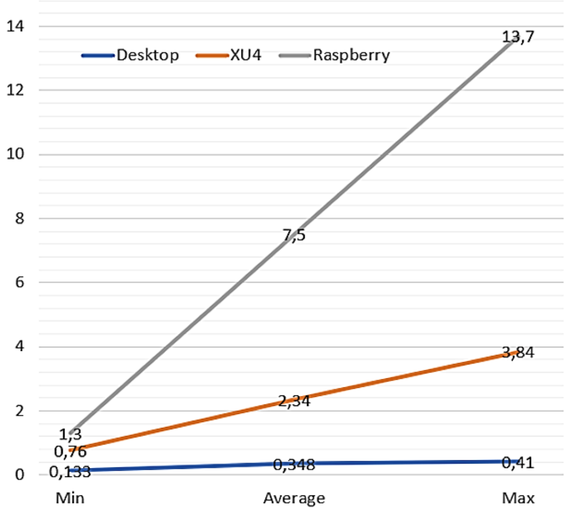


Fig. 14. OpenMP Executing Time.

TABLE IV. DIFFERENT EXECUTING TIMES

Executing time (ms)	Desktop	XU4	Raspberry
C/C++	1.56	6.8	9.2
C/C++ - OpenMP	0.34	2.34	7.5

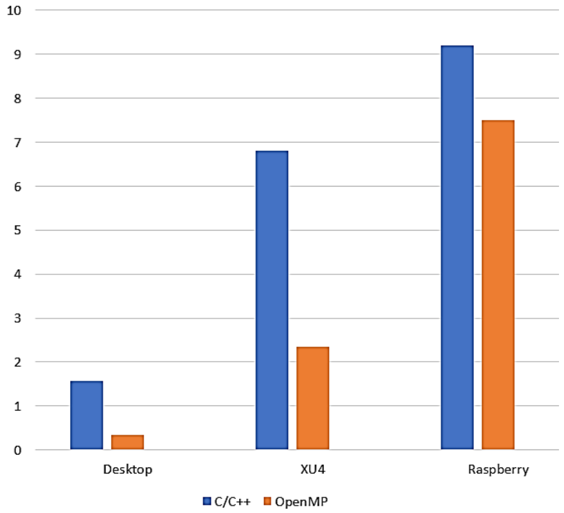


Fig. 15. Mean Processing Times (MS) based on different Architectures.

Table IV and Fig. 15 shows a comparison of all processing times using the different architectures and both C/C++ and OpenMP parallel implementation.

The optimization of the algorithm on the XU4 architecture proves to be very efficient and makes it possible to speed up processing and achieve real-time processing times in addition to its power consumption advantage.

IV. Conclusion

In this paper, a complex algorithm-based ECG signals processing is studied in order to meet the requirements of monitoring applications in terms of real time and portability on a low power architecture.

The evaluation of the algorithm using Matlab allowed validation of the algorithm and the evaluation of the different metrics (MSE, PRD, and SNR errors).

The approach followed by the algorithm parallelization on an adequate architecture is effective to process signals at 2.34ms/samples using a 360 Hz frequency acquisition.

This study opens up research perspectives to design a system integrating sensors and a SoC whose architecture is similar to that of the XU4 and which integrates an FPGA in order to carry out on-the-fly signal processing without data storage.

Acknowledgment

We would like to thank the National Centre for Scientific and Technical Research (CNRST) of Morocco for the financial support (scholarship number: 588UIZ2017).

References

[1] W. Jenkal, R. Latif, A. Toumanari, A. Dliou, O. El, and F. Mrabih, “QRS Detection Based on an Advanced Multilevel Algorithm,”

International Journal of Advanced Computer Science and Applications (IJACSA)., vol. 7, no. 1, 2016, doi: 10.14569/IJACSA.2016.070135.

[2] S. Elouaham, A. Dliou, R. Latif, and M. Laaboubi, "Filtering of Biomedical signals by using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," *Int. J. Comput. Appl.*, vol. 149, no. 7, pp. 39–43, 2016, doi: 10.5120/ijca2016911515.

[3] W. Jenkal, R. Latif, A. Toumanari, A. Dliou, O. El B'Charri, and F. M. R. Maoulainine, "An efficient algorithm of ECG signal denoising using the adaptive dual threshold filter and the discrete wavelet transform," *Biocybern. Biomed. Eng.*, vol. 36, no. 3, pp. 499–508, 2016, doi: 10.1016/j.bbe.2016.04.001.

[4] P. C. Bhaskar and M. D. Uplane, "High Frequency Electromyogram Noise Removal from Electrocardiogram Using FIR Low Pass Filter Based on FPGA," *Procedia Technol.*, 2016, doi: 10.1016/j.protcy.2016.08.137.

[5] S. Mejhoudi, R. Latif, A. Elouardi, and W. Jenkal, "Advanced Methods and Implementation Tools for Cardiac Signal Analysis," *Adv. Sci. Technol. Innov.*, pp. 95–103, 2019, doi: 10.1007/978-3-030-05276-8_11.

[6] Z. Wang, F. Wan, C. M. Wong, and L. Zhang, "Adaptive Fourier decomposition based ECG denoising," *Comput. Biol. Med.*, 2016, doi: 10.1016/j.compbiomed.2016.08.013.

[7] W. Jenkal, R. Latif, A. Toumanari, A. Dliou, and O. El B'charri, "An efficient method of ecg signals denoising based on an adaptive algorithm using mean filter and an adaptive dual threshold filter," *Int. Rev. Comput. Softw.*, vol. 10, no. 11, pp. 1089–1095, 2015, doi: 10.15866/irecos.v10i11.7821.

[8] W. Jenkal, R. Latif, A. Toumanari, A. Elouardi, A. Hatim, and O. El'bcharri, "Real-time hardware architecture of the adaptive dual threshold filter based ECG signal denoising," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 14, pp. 4649–4659, 2018.

[9] G. Tang and A. Qin, "ECG de-noising based on empirical mode decomposition," *Proc. 9th Int. Conf. Young Comput. Sci. ICYCS 2008*, pp. 903–906, 2008, doi: 10.1109/ICYCS.2008.178.

[10] M. A. Kabir and C. Shahnaz, "Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains," *Biomed. Signal Process. Control*, vol. 7, no. 5, pp. 481–489, 2012, doi: 10.1016/j.bspc.2011.11.003.

[11] T. Wang, M. Zhang, Q. Yu, and H. Zhang, "Comparing the applications of EMD and EEMD on time-frequency analysis of seismic signal," *J. Appl. Geophys.*, vol. 83, pp. 29–34, 2012, doi: 10.1016/j.jappgeo.2012.05.002.

[12] O. El B'charri, R. Latif, W. Jenkal, and A. Abenaou, "The ECG Signal Compression Using an Efficient Algorithm Based on the DWT," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 3, pp. 181–187, 2016.

[13] A. Giorgio, "A New FPGA-based Medical Device for the Real Time Prevention of the Risk of Arrythmias," *Int. J. Appl. Eng. Res.*, vol. 11, no. 8, pp. 6013–6017, 2016.

[14] Z. Zhang, Z. Li, and Z. Li, "An Improved Real-Time R-Wave Detection Efficient Algorithm in Exercise ECG Signal Analysis," *J. Healthc. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/8868685.

[15] S. Sahoo, P. Biswal, T. Das, and S. Sabut, "De-noising of ECG Signal and QRS Detection Using Hilbert Transform and Adaptive Thresholding," *Procedia Technol.*, vol. 25, no. Raerest, pp. 68–75, 2016, doi: 10.1016/j.protcy.2016.08.082.

[16] T. A. Rashid, C. Chakraborty, and K. Fraser, "Advances in Telemedicine for Health Monitoring: Technologies, Design and Applications," *Adv. Telemed. Heal. Monit. Technol. Des. Appl.*, no. June, 2020, doi: 10.1049/pbhe023e.

[17] H. W. Lim, M. Syafiq, M. Sani, A. Hashim, and Y. W. Hau, "Throb : System-on-Chip based Arrhythmia Screener with Self Interpretation," pp. 30–36, 2015.

[18] W. Shen, D. Wei, W. Xu, X. Zhu, and S. Yuan, "Parallelized computation for computer simulation of electrocardiograms using personal computers with multi-core CPU and general-purpose GPU," *Comput. Methods Programs Biomed.*, vol. 100, no. 1, pp. 87–96, Oct. 2010, doi: 10.1016/J.CMPB.2010.06.015.

[19] L. V. R. Kumari, Y. P. Sai, N. Balaji, and K. Viswada, "FPGA Based Arrhythmia Detection," *Procedia Comput. Sci.*, vol. 57, pp. 970–979, 2015, doi: 10.1016/j.procs.2015.07.495.

[20] B. Chandrakar, O. P. Yadav, and V. K. Chandra, "a Survey of Noise Removal Techniques for Ecg Signals," *Ijarcc*, vol. 2, no. 3, pp. 1354–1357, 2013, [Online]. Available: www.ijarcce.com.

[21] J. M. Łęski and N. Henzel, "ECG baseline wander and powerline interference reduction using nonlinear filter bank," *Signal Processing*, vol. 85, no. 4, pp. 781–793, 2005, doi: 10.1016/j.sigpro.2004.12.001.

[22] Z. Haque, R. Qureshi, M. Nawaz, F. Khuhawar, N. Tunio, M. Uzair, "Analysis of ECG Signal Processing and Filtering Algorithms," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, Issue 3, pp. 545-550, 2019, doi: 10.14569/IJACSA.2019.0100370.

[23] P. Nguyen and J. M. Kim, "Adaptive ECG denoising using genetic algorithm-based thresholding and ensemble empirical mode decomposition," *Inf. Sci. (Ny).*, vol. 373, pp. 499–511, 2016, doi: 10.1016/j.ins.2016.09.033.

[24] M. Rakshit and S. Das, "An efficient ECG denoising methodology using empirical mode decomposition and adaptive switching mean filter," *Biomed. Signal Process. Control*, vol. 40, pp. 140–148, 2018, doi: 10.1016/j.bspc.2017.09.020.

[25] G. Han, B. Lin, and Z. Xu, "Electrocardiogram signal denoising based on empirical mode decomposition technique: An overview," *J. Instrum.*, vol. 12, no. 3, 2017, doi: 10.1088/1748-0221/12/03/P03010.

[26] K. J. Bijwe1, S. S. Vasekar, "FPGA Implementation of DWT for ECG Signal Pre-Processing", *International Journal of Engineering Science and Computing* vol. 6, no. 8, pp. 2450–2452, 2016.

[27] E. M. El Hassan and M. Karim, "An FPGA-based implementation of a pre-processing stage for ECG signal analysis using DWT," 2014 *2nd World Conf. Complex Syst. WCCS 2014*, pp. 649–654, 2015, doi: 10.1109/ICoCS.2014.7060929.

[28] O. El B'charri, R. Latif, K. Elmansouri, A. Abenaou, and W. Jenkal, "ECG signal performance de-noising assessment based on threshold tuning of dual-tree wavelet transform," *Biomed. Eng. Online*, vol. 16, no. 1, pp. 1–18, 2017, doi: 10.1186/s12938-017-0315-1.

[29] S. Poungponsri and X. H. Yu, "An adaptive filtering approach for electrocardiogram (ECG) signal noise reduction using neural networks," *Neurocomputing,* vol. 117, pp. 206–213, 2013, doi: 10.1016/j.neucom.2013.02.010.

[30] P. Podder, M. Mehedi Hasan, M. Rafiqul Islam, and M. Sayeed, "Design and implementation of butterworth, chebyshev-i and elliptic filter for speech signal analysis," *arXiv*, 2020, doi: 10.5120/17195-7390.

[31] P. N. Malleswari, C. Hima Bindu, and K. Satya Prasad, "An investigation on the performance analysis of ECG signal denoising using digital filters and wavelet family," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 166–171, 2019.

[32] S. Thakran, "A hybrid GPFA-EEMD_Fuzzy threshold method for ECG signal de-noising," *J. Intell. Fuzzy Syst.*, vol. 39, no. 5, pp. 6773–6782, 2020, doi: 10.3233/JIFS-191518.

[33] V. Gupta, V. Chaurasia, and M. Shandilya, "Random-valued impulse noise removal using adaptive dual threshold median filter," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 296–304, 2015, doi: 10.1016/j.jvcir.2014.10.004.

[34] W. Jenkal, R. Latif, A. Elouardi, and S. Mejhoudi, "FPGA Implementation of the Real-Time ADTF process using the Intel-Altera DE1 Board for ECG signal Denoising," *Proc. 2019 IEEE World Conf. Complex Syst. WCCS 2019*, 2019, doi: 10.1109/ICoCS.2019.8930780.

[35] S. Mejhoudi, R. Latif, W. Jenkal, and A. Elouardi, "Real-Time ecg signal denoising using the adtf algorithm for embedded implementation on fpgas," *Proc. 2019 IEEE World Conf. Complex Syst. WCCS 2019*, 2019, doi: 10.1109/ICoCS.2019.8930771.

[36] K. M. Chang, "Arrhythmia ECG noise reduction by ensemble empirical mode decomposition," *Sensors,* Vol. 10, Issue. 6, 2010, pp.6063-80.

[37] P. Nguyen, J. M. Kim, "Adaptive ECG denoising using genetic algorithm-based thresholding and ensemble empirical mode decomposition," *Information Sciences*, Vol. 373, 2016, pp. 499-511.Y.

[38] L. Wu, D. Agrawal, and A. El Abbadi, “A comparison of DFT and DWT based similarity search in time-series databases,” pp. 488–495, 2000, doi: 10.1145/354756.354857.

[39] S. Hossain and D-j. Lee, “Deep Learning-Based Real-Time Multiple-Object Detection and Tracking from Aerial Imagery via a Flying Robot with GPU-Based Embedded Devices,” *Sensors*, Vol. 19, Issue. 15, 2019, doi: 10.3390/s19153371.

[40] U. Prodhan, T. Saha, R. Shaharin, T. Emon and M. Rahman, “Implementation of Low Cost Remote Primary Healthcare Services through Telemedicine: Bangladesh Perspectives,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(11), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111118.

Efficient Rain Simulation based on Constrained View Frustum

JinGi Im[1]
Dept. of Computer Engineering, Graduate School
Keimyung University, Daegu, Republic of Korea

Mankyu Sung[2]*
Dept. of Game and Mobile Engineering
Keimyung University, Daegu, Republic of Korea

***Abstract*—Realistic real-time rain streaks rendering has been treated as a very difficult problem because of various natural phenomena. Also, for creating and managing many particles in a rain streak, many resources had to be used. This paper propose am efficient real-time rain streaks simulation algorithm by generating view-dependent rain particles, which can express a large amount of rain streaks even with a small number of particles. By creating a 'constrained view frustum' depending on the camera moving in real time, particles are rendered only in that space. Accordingly, particles rendered well even if the camera keep moving or rotating rapidly. And a small number of particles are used, since the simulation is performed in a user-viewed limited space, an effect of simulation many particles can be obtained. This enables very efficient real-time simulation of rain streaks.**

***Keywords*—View-dependent rendering; realistic real-time simulation; view frustum**

I. Introduction

When a digital content is produced for a specific weather from among various weather conditions, the audience or user who encounters the medium can immerge to the content more easily. There are many types of weather conditions such as sunny, cloudy, rainy, and snowy days. There are two main methods of rendering them: off-line rendering and on-line rendering. In the case of movies, since it is a medium that does not communicate with the audience in real-time, so even if it takes a lot of time, it pursues realistic rendering results so that the results naturally melt into the filming scene. On the contrary, in content such as games, even if the factual point is relatively less important, it emphasizes real-time, so it pursues somewhat lower quality rendering results compared to that of movies.

This paper proposes a real-time simulation algorithm for rainy scene. When it rains, complicated and diverse phenomena should be considered, such as droplets, splashes, rainbows, and clouds. Although the graphics hardware has been improved drastically in recent years, real-time simulation of rains is still very difficult problem considering the various conditions and the governing physical law to simulate them. Moreover, considering the physical properties of raindrops or water, the entire simulation process becomes very complicated. Therefore, most of real-time rain simulation focuses only on specific phenomena among various phenomena in rainy weather, and many different technical methods have been proposed to obtain real-time simulation performance by approximating the required parameters. Many of them have presented for modeling realistic raindrops based on the physical properties [1]. This enables realistic rain simulation as well as raindrop modeling. Also, they could present various phenomena including collision detection between raindrops and objects but gave the disadvantage of inefficiency because calculations related to raindrops became heavy due to unnecessary invisible rendering on the screen [2]. To fix it, various studies have been proposed such as defining and rendering only specific region for efficient simulation, but when the camera is moving outside of the region, there are awkward discontinuities where raindrops are not visible or rendered outside of the space [3, 4, 5].

This paper proposes an algorithm that complements the aforementioned shortcomings by mapping rain streaks textures to particles and creating a particle system that depends on the position and FOV (field of view) of the camera. Also present a method of interaction between rain and light sources, which is a simple light scattering technique for changing color of rain streaks. In this work, the particles are rendered even when the camera is moving and rotating. Therefore, it can avoid awkward rendering where rain can be rendered only in certain spaces, which requires only a small number of particles, but appearing to be rendered in very large quantities. This enables more efficient real-time rain-streaks simulation in a 3D space.

II. Related Work

Many research methods have been proposed in realistic rain simulation, and most of them focus on certain parts of the phenomenon of rain. There are two main types of rain simulation: rendering rain streaks with translucent white quads. The other method is to map precomputed rain streak textures to quads. The rain streaks texture rendering model proposed by Garg et al. presents a vibration model for raindrops. The rain streaks texture rendering model proposed by Garg et al. presents a vibration model for raindrops. As a result, they made a database of high-quality renders for many values of the illumination parameters [1]. Then, they used simple image-based algorithms from the depth map, camera parameters, and user input for viewpoints to synthesize the final images. This method showed the performance of 10 sec/frame, which was unsuitable for real-time simulation. This study used their database but implement rain simulation in real-time. Also, for randomness of the rain streaks, hundreds of textures are used to randomly map to the particles.

Weber *et al*. focused on the relationship between raindrops and trees. In their techniques, through-fall simulation is

*Corresponding Author

analyzed phenomenologically, and rendering is done based on the amount of water stored in the tree canopy and leaves [2]. Furthermore, in the study of Nanko *et al.*, the distribution of dripping through-fall was considered temporally and spatially [6]. The through-fall was largely divided into two different water. The first water was the raindrops in the natural state. This was a free through-falling without hitting anything other objects. The second one was the water splashes stored by hitting the canopy and leaves and then re-appeared in phenomenological and hydrological condition. In their study, through-fall was implemented very similar to the actual phenomenon. But there was no explanation for the calculation of raindrops occurring outside the camera's FOV.

Rousseau et al. proposed a model representing the refraction of light occurring inside the raindrop [7]. For completion, the reflections should also have been considered when designing raindrops. However, since they thought the reflection was negligible enough, they implemented only refraction [8]. In their approach, rain streak textures extracted from the video were modified to match the camera of a scene and then blended into the image to make the artist's intention for the scene more effective. But, real-time performance was not guaranteed. In addition to this, dynamic scenes were not suitable for these methods because the texture must be transformed to fit the resulting screen.

Tariq further simplified Garg's rain textures and map them to the quads [3]. This processing was done on the GPU using DirectX, and each particle was rendered using a geometric shader at each frame over times. Tariq also made lights glow to show more realistic simulation of rain under the lights, but the relationship with the light source and rains was not described mathematically, therefore rain was not rendered properly as the camera moved or rotated.

Puig-Centelles *et al*. proposed a new real-time rain simulation technique in which a rain area was defined as an ellipse and all rain simulation was limited to a semi-cylindrical sub-volume [4]. Due to observer's movement, the update of new particle position was forced to inside the sub-volume, while their density was adapted to reduce the number of particles needed. Furthermore, they separated the close rain and far rain and added a transition area in between for a natural and realistic change [9]. The switch was made depending on whether the observer is in a rain area or not. If the camera is located in the rainy area from a very long distance, awkward scene could be rendered, with certain area raining and others not raining.

Unlike the above studies, the model proposed in this paper checks the position of the camera and then create a rain space for the camera in which particles are generated inside. This ensures that even if the camera keep moving or rotating rapidly, it makes illusion that rain is rendering over the entire scene. Other benefit is that can use only small number of particles to create heavy raining, which improves the rendering performance. On the top of that, this study presents a model for the interaction between particles and the light source, which can represent the light scattering effect.

Fig. 1 is a captured scene of the proposed model. The four red lines indicate the volume of view-frustum, and the intersection point where four lines meet together represents the actual camera position. In the Fig. 1, particles are generated only in the frustum of the FOV and when the camera rotates or moves, the location where the particles are generated is also moved according to the camera.

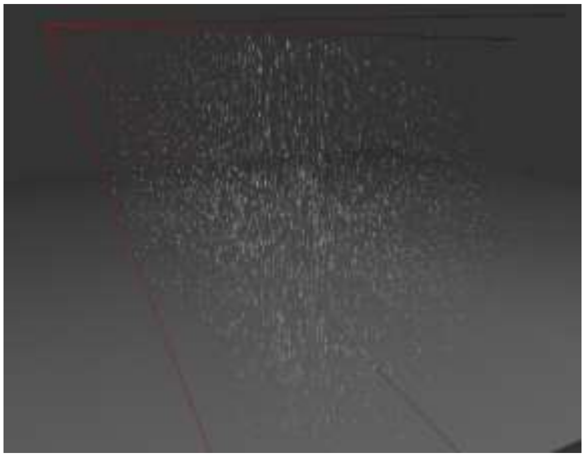


Fig. 1. Particles Rendered only in the Constrained Area.

III. ALGORITHM

In this study, we propose a real-time rain simulation method that creates a constrained rain space depending on the current camera so that particles representing the rain streaks are generated only in the camera frustum. Because of this, even when a small number of particles are used, suggested algorithm is able to synthesize seemingly heavy rain for users.

This chapter describes the detailed algorithm of the proposed model. The overall algorithm is briefly described in Fig. 2. Each procedure is as follows:

1) *Scene configuration:* Compose the overall scene such as background and model loading(street lamp, plane).

2) *Particles initialize and configuration:* In this model, the KTX(Khronos Texture) format is applied. More details in in Sec. 3.D, and the particle system was constructed using transform feedback from openGL [10, 11].

3) *Create constrained rain area:* Under the perspective projection, to make area for the rain fall, this method compute the 8 vertex positions of the truncated pyramid shape of the view frustum and the normal vectors of 6 faces. At the same time, it creates a virtual sphere in the frustum for enforcing a constraint to limit the area for generating particles.

4) *Set the initial particle positions in the rain area:* Set the initial positions of the particles in a constrained rain area. Particles are updated in position from top to bottom(-y).

5) *Set the rain streak colors by calculating the scattering of particles with the light source:* Depending on the relative position between particles and light soruce, the method of handling scattering from particles is different. In this paper, only spotlight is considered among the types of light sources.

6) *Real-time simulation:* Particles are created and rendered in constrained rain area. Calculating the rain space is handled by the CPU, but scattering and calculation of particle is handled by the shader.

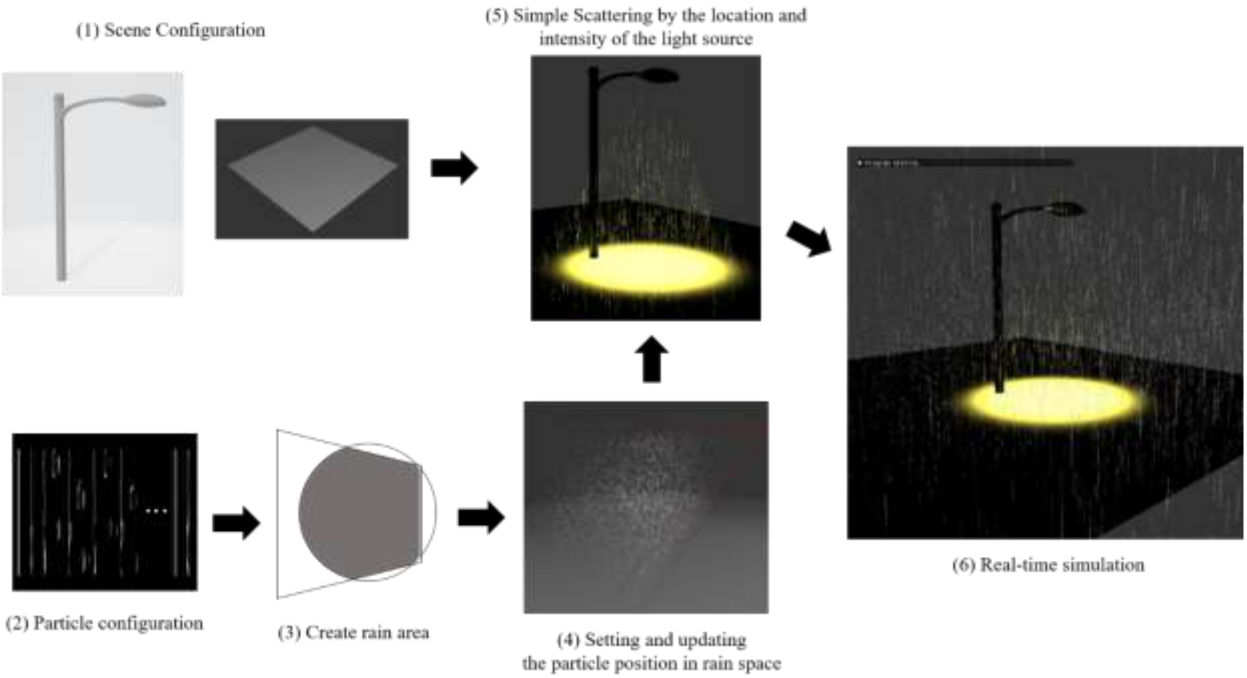

(1) Scene Configuration
(5) Simple Scattering by the location and intensity of the light source
(6) Real-time simulation
(2) Particle configuration
(3) Create rain area
(4) Setting and updating the particle position in rain space

Fig. 2. Algorithm Overview of Proposed Model.

First, explain the frustum and rain area, apply texture to particles generated only in space. After that, this paper discusses methods for light scattering in this section.

A. Create view-dependent Rain Area

To create the rain area from the camera's frustum, this method first need to calculate the height, width, center point of the far and near planes of the frustum. If the distance from the camera to the far plane increases, the size of the frustum increases as well. Therefore, the density of the particle created in the frustum is very low. In this case, it is different from what we wanted because algorithm has to make lots of particles to increase the density. Also, particles generated near to the far plane are rendered too small or almost invisible. Therefore, the distance from the camera to the far plane for a particle should be set as small as possible. In this paper, it was set to 50.0. After that, we need to find the positions of the 8 vertices that make up the frustum, which can be got from following formula:

$$f_{lr} = C_f - \{\vec{U} * (h_f * 0.5)\} + \{\vec{R} * (w_f * 0.5)\} \tag{1}$$

$$f_{ll} = C_f - \{\vec{U} * (h_f * 0.5)\} - \{\vec{R} * (w_f * 0.5)\}$$

$$f_{ur} = C_f + \{\vec{U} * (h_f * 0.5)\} + \{\vec{R} * (w_f * 0.5)\}$$

$$f_{ul} = C_f + \{\vec{U} * (h_f * 0.5)\} - \{\vec{R} * (w_f * 0.5)\}$$

$$n_{lr} = C_n - \{\vec{U} * (h_n * 0.5)\} + \{\vec{R} * (w_n * 0.5)\}$$

$$n_{ll} = C_n - \{\vec{U} * (h_n * 0.5)\} - \{\vec{R} * (w_n * 0.5)\}$$

$$n_{ur} = C_n + \{\vec{U} * (h_n * 0.5)\} + \{\vec{R} * (w_n * 0.5)\}$$

$$n_{ul} = C_n + \{\vec{U} * (h_n * 0.5)\} - \{\vec{R} * (w_n * 0.5)\}$$

where, C is the center point of far and near face, h, w is the height and width of each face. $\vec{U}$ is up vector of camera, and $\vec{V}$ is direction of camera. Also, $\vec{R}$ is a cross product and normalized vector of $\vec{U}$ and $\vec{V}$. Each vertex is described in Fig. 3.

Using the 8 vertices, this paper can find the normalized vectors for the four faces of the frustum, excluding the near and far plane. This is given subsequent formula:

$$\widehat{RP} = (n_{lr} - f_{lr}) \times (n_{lr} - n_{ur}) \tag{2}$$

$$\widehat{LP} = (n_{ll} - f_{ll}) \times (n_{ll} - n_{ul})$$

$$\widehat{BP} = (n_{ll} - n_{lr}) \times (n_{ll} - f_{ll})$$

$$\widehat{TP} = (n_{ul} - f_{ul}) \times (n_{ul} - n_{ur})$$

Where $\widehat{RP}$, $\widehat{LP}$, $\widehat{BP}$, $\widehat{TP}$ is a normal vector for the right, left, bottom and top faces, respectively, and each vector is normalized. Each face is also described in Fig. 3.

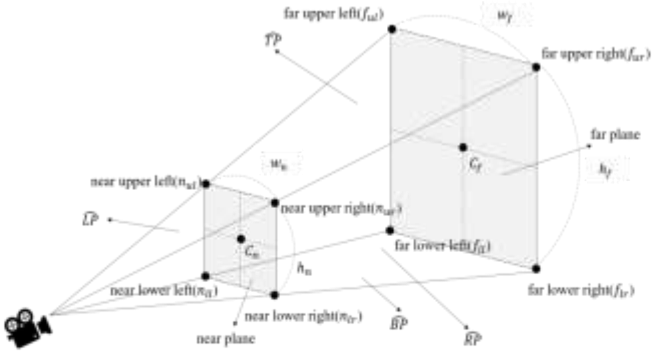


Fig. 3. Vertex and Faces of the view Frustum.

B. Virtual Sphere Setting

To constrained create space where particles will be generated, we need to create a virtual sphere that overlaps the frustum. So, we need to set the center point $C_{vs} = (x_{vs}, y_{vs}, z_{vs})$ and radius r_{vs} of the sphere. The C_{vs} of the sphere is the midpoint of the distance between the near and far faces, so that the sphere and the frustum overlap as much as possible.

As a result, as much space as possible can be defined as rain space on the frustum. Since the C_{vs} the midpoint between the near and far face, the radius of virtual sphere is defined as half the distance between two faces.

Now, algorithm need to decide the initial position $p_{par} = (x_{par}, y_{par}, z_{par})$ where a particle is created. Particle must be created inside a constrained rain space where the frustum and virtual sphere overlapped. To find this location, we first use the equation for converting from the spherical coordinate system to Cartesian coordinate system to create particles in sphere. Before this, we should get the θ and ϕ for coordinate system conversion. θ means azimuthal angle of spherical coordinate and ϕ means polar angle of spherical coordinate. This is shown in as follow:

$$\theta = 2 * PI * randmom\ seed(0,1) \tag{3}$$

$$\phi = \arccos(2 * randmom\ seed(0,1) - 1)$$

Where $\mathrm{random\ seed(a,b)}$ is a function that return on random number between a to b. And now, we can get the $\mathrm{p_{par}}$ and equation is as follow:

$$x_{par} = random\ seed(-r_{vs}, r_{vs}) \sin(\phi) \cos(\theta) + x_{vs} \tag{4}$$

$$y_{par} = random\ seed(-r_{vs}, r_{vs}) \sin(\phi) \sin(\theta) + y_{vs}$$

$$z_{par} = random\ seed(-r_{vs}, r_{vs}) \cos(\phi) + z_{vs}$$

Because of coordinate system conversion, the particles take the form of a virtual sphere. In other words, the particle is randomly set the initial position inside the virtual sphere.

So, now we need to limit the position where the particles are generated to the space where the frustum and the sphere overlap, that is, the rain area. First, select 2 of the 8 vertices of the frustum. Then, two normal vector value are got with the position of the particle as the starting point and the selected vertex as the endpoint. This normal vector is a normal vector later value to determine whether the current position of the particle is inside or outside the frustum. It is obtained as follow:

$$\widehat{PAR\_N}_{ll} = p_{par} - n_{ll} \tag{5}$$

$$\widehat{PAR\_N}_{ur} = p_{par} - n_{ur}$$

In this paper, algorithm used nll (near lower left) and nur (near upper right) vertex.

Using results of (5), calculate the dot product of two normal vectors in (5) and the four faces normal vector of the frustum besides near and far faces. Note that NLL of $\widehat{PAR\_N}_{ll}$ is the bottom left vertex of near face. So, this normal vector should be calculated with $\widehat{BP}$ and $\widehat{LP}$ vectors. Similarly, the normal vector $\widehat{PAR\_N}_{ur}$ works $\widehat{TP}$ and $\widehat{RP}$. This is shown in (5) as follow:

$$n_{ur}RP = \widehat{PAR\_N}_{ur} \cdot \widehat{RP} \tag{6}$$

$$n_{ll}LP = \widehat{PAR\_N}_{ll} \cdot \widehat{LP}$$

$$n_{ur}TP = \widehat{PAR\_N}_{ll} \cdot \widehat{BP}$$

$$n_{ll}BP = \widehat{PAR\_N}_{ur} \cdot \widehat{TP}$$

By checking whether the 4 scalar values resulting from (5) are greater than or less than 0 or not, it is possible to know that the position of the particle is inside of the frustum. If the particles are generated inside the frustum as we wish, the algorithm keeps the position unadjusted and only updates y_{par}. Conversely, if it is created outside of the frustum, then the position of the particle is moved before rendering so that it is created inside the frustum.

C. Spotlight Scattering

In order to render the rain more realistically under the various light condition, we propose a simple light scattering model between particles and lights. In our approach, we consider only spotlight because it is the type of light that affects the rain streak color significantly. For example, a spotlight such as streetlight can be found easily in real life. Other lights such as direction light and point lights are hardly seen in rainy days. Thus, this paper did not consider those lights in our study.

Fig. 4 shows three different cases when the rain streaks interact with light source. When we calculate the light scattering, the position of the particle must be decided as follows:

- Particles are located above the light source.
- Particles are under the light source but are not affected by the light.
- Particles are under the light source and are affected by the light.

Since the range that the spotlight affects has a shape of a cone, we consider particles that are inside the cone and ignore all other particles outside. To improve the physical accuracy, both 1) and 2) cases must be considered because lights may be reflected from other objects or raindrops, but since this is very insignificant and unnoticeable by the human eye, so those cases are not considered in this paper.

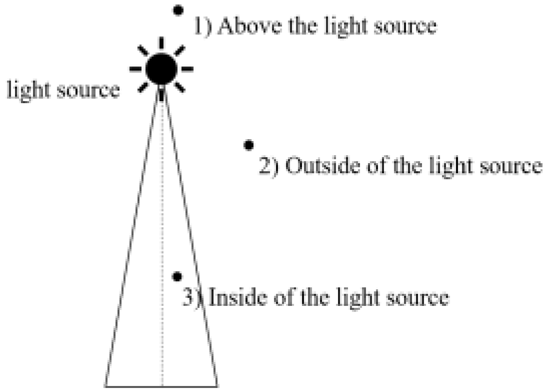

1) Above the light source
light source
2) Outside of the light source
3) Inside of the light source

Fig. 4. Conditions between Particle and Light Source.

1) Particles above the light source: Equation (7) is the equation for calculating a scalar value UD which is the dot product of the normal vector $\widehat{PL}$, which is the normalized vector from the particle to light source, and normal vector $\widehat{D}$ representing the direction of the light source.

$$ab_{light} = \widehat{PL} \cdot \widehat{D} \tag{7}$$

The ab_{light} is the value that determines whether the particle's current position is above or below the light source. If this value is higher than the height of spotlight, it means that the position of the particle is above the light source. So, the particle is not affected at all. Therefore, there is no change in particles at this case.

2) Particles under the light source: When the current particle position is under the light source, there are two cases as shown 2) and 3) in Fig. 4. Most spotlights have a cone shape. A cone is a collection of smaller or lager circles based on a point on an axis. In other words, it can be seen as a collection of circles that gradually getting smaller from the radius of the base. If the particles are in circles, they are scattered under the influence of light source. On the contrary, if particles are outside the circles, they are not affected. For this, the radius r_{circle} of the cone at the current position of the particle along the axis can be obtained using $updown$, which is the result of (7), and h_{cone}, which is height of the cone. It can be expressed as the following (8), where $base\ radius$ is the base radius of the cone:

$$r_{circle} = \left(\frac{ab_{light}}{h_{cone}}\right) * (r_{cone}) \tag{8}$$

And we can get the orthogonal distance $dist_{ortho}$ from the axis of cone to the p_{par}. $dist_{ortho}$, along with r_{circle}, is an important to know whether a particle is inside or outside the cone. To obtain $dist_{ortho}$ is expressed as (9), where $length()$ is a function that return the size of a vector as a parameter.

$$dist_{ortho} = length\{(p_{par} - h_{cone}) - ab_{light} * \widehat{D}\} \tag{9}$$

Now, we can compare r_{circle} and $dist_{ortho}$ to determine whether the p_{par} is inside or outside the cone. If r_{circle} is a larger than $dist_{ortho}$, the particle is inside the cone, which is the case as 3) in Fig. 4. And this case, the color of the particle becomes the same as the color of the light source. Also, because it is affected by light, the color of the particle appears more clearer as the intensity of the light increases.

Contrary, $dist_{ortho}$ is larger, it is the same as 2) in Fig. 4. This case, the particle does not change. The process can be expressed as the following (10) and the contents of each variable expressed in the Fig. 5.

$$if\,(r_{circle} \geq dist_{ortho}) \tag{10}$$

$$color_{par} = color_{tex} \times color_{light} \times intensity_{light}$$

$$else\ if\,(r_{circle} < dist_{ortho})$$

$$color_{par} = color_{tex} \times color_{bg} \times \alpha$$

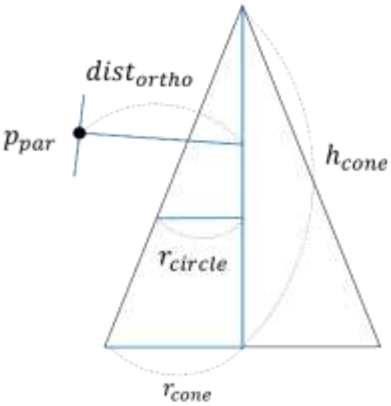


Fig. 5. Structure of Cone (Spotlight Area).

D. Texture Mapping

Garg and Nayer released their rain streaks textures as a PNG files [12]. Since this study used many textures, these files were put into one KTX file invented by the Khronos group, and then the Texture Array was used in OpenGL graphics API [13]. When we use the Texture Array, each texture corresponds to a single layer of the array. Therefore, when initializing particles, many particles are created, and they are allocated a layer for each particle. The condition for assigning a layer is random.

IV. Experiment

The proposed view-dependent rain model calculates the camera position and various parameters continuously. The experiment compares the performance of proposed model with other models after generating random numbers with a seed number. The CPU for the computer in which the experiment was conducted is Intel i7-8700, and the memory size is DDR4 16Gb * 2, a total of 32Gb. Also, the graphics card uses GTX GeForce 1080ti. All experiments were conducted in the same environment.

This study compared our proposed method with two existing models that Creus and Patow-Tariq proposed. Both models used Garg's rain streaks textures in a same way as the proposed model [3, 5]. Although the details of each algorithm may be different, it is enough to compare their FPS because three models used same rain textures. Two other models and proposed model in this paper were tested in the same environment. The changes of FPS according to the number of particles for three models are shown Fig. 7.

The proposed model in this paper, as the number of particles increased, decrease its framerate compared to the other two models. However, in the proposed model, even when a small number of particles was used, since particles were generated only in rain space within the camera's field of view, they were seemingly more than the actual number.

When we compare our result with two other method, the visual results are obvious as shown in Fig. 6. Although all three models have a fixed number of particles of 10,000, the proposed model looked to render larger number of particles than other models. This means that even with a small number of particles, we can express a large number of particles.

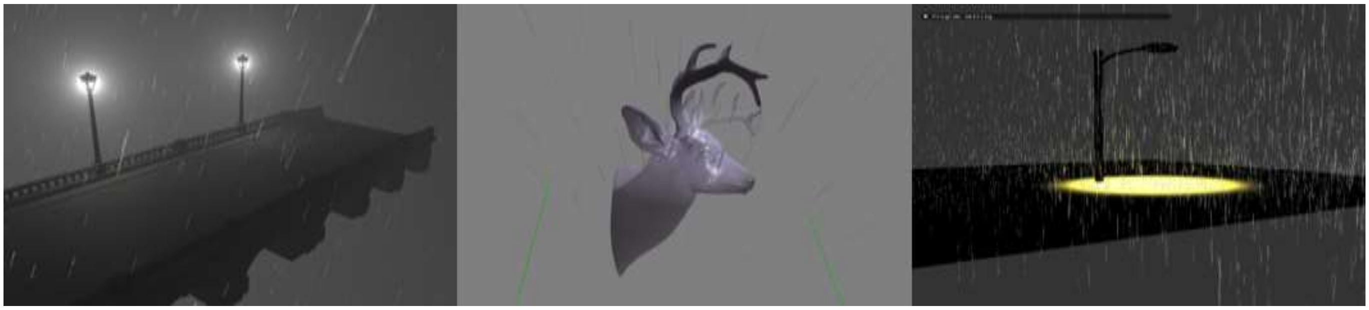


Fig. 6. (a) The Proposed Model (b) Tariq's Model (c) Crues and Patow 's Model. Three Models have the Same Number of Particles.

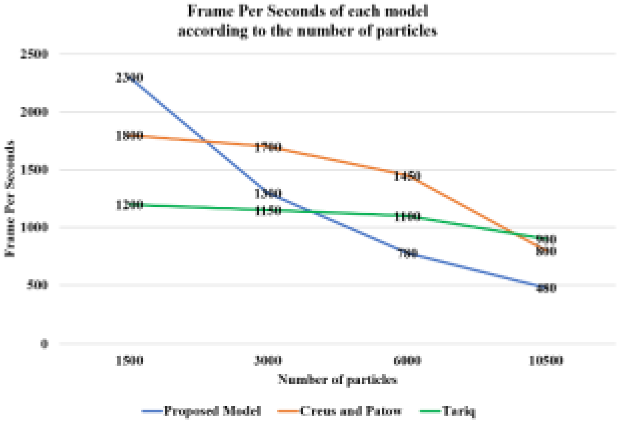


Fig. 7. Graph of Frame Change according to Particle Number.

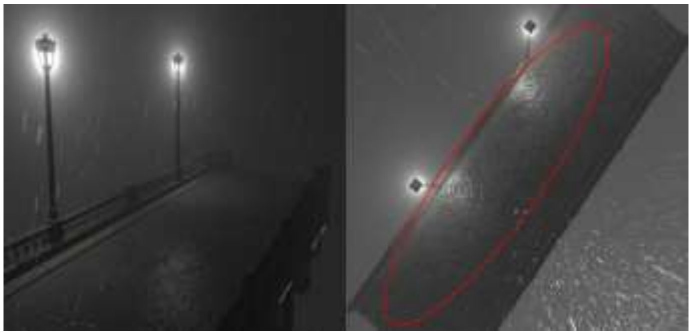
Tariq's model showed very stable performance in terms of FPS even when we increase the number of particles. In addition, a very realistic simulation result was obtained because the glow effect of the light source was considered as shown in Fig. 8(a). However, when the camera was continuously moving, at some point, particles were disappeared as shown in red circle Fig. 8(b).

The algorithm proposed by Creus and Patow, on the other hand, the FPS drops relatively in stable manner as the number of particles increases. Although not shown in the graph in Fig. 7, even when the number of particles was exceeded 10,500, real time performance was still maintained. However, as shown in Fig. 9, there were empty space in the environment where no rain was rendered when the camera is moving around. In addition, particles are keep generating and collisions are still checked even when the camera is not looking at, which degrades the overall performance.

The proposed model in this paper, as shown in the graph of Fig. 6, the FPS looks to drop higher than other two algorithms. This was caused by heavy computation on updating the position of constrained rain space, the frustum, the virtual sphere, and the particle position.

However, as shown in Fig. 10, because our algorithm makes the constrained rain space depend on the camera, even if there are a lot of changes in the camera, particles are still generating in front of the camera. This improves the visual quality of simulation.

Fig. 11 shows that the color and location of light source are fixed, and the number of particles is different. The case of (a) and (b), rendered particles are small, but it seems more than actual number. In the case of (c), rendered particles are 10,500 and it gives a feeling that it is raining quite a bit. In (d), the number of particles is the highest, 49500, and it shows that seems like it is raining a lot.



(a) Before Camera Moving. (b) After Camera Moving.

Fig. 8. Tariq's Rendering Results.

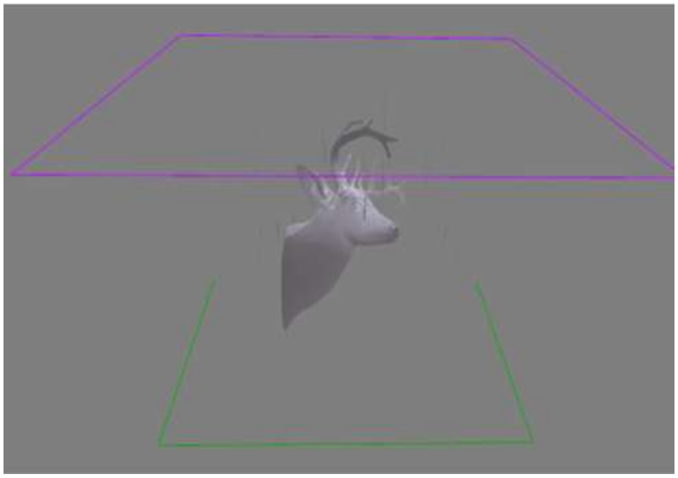
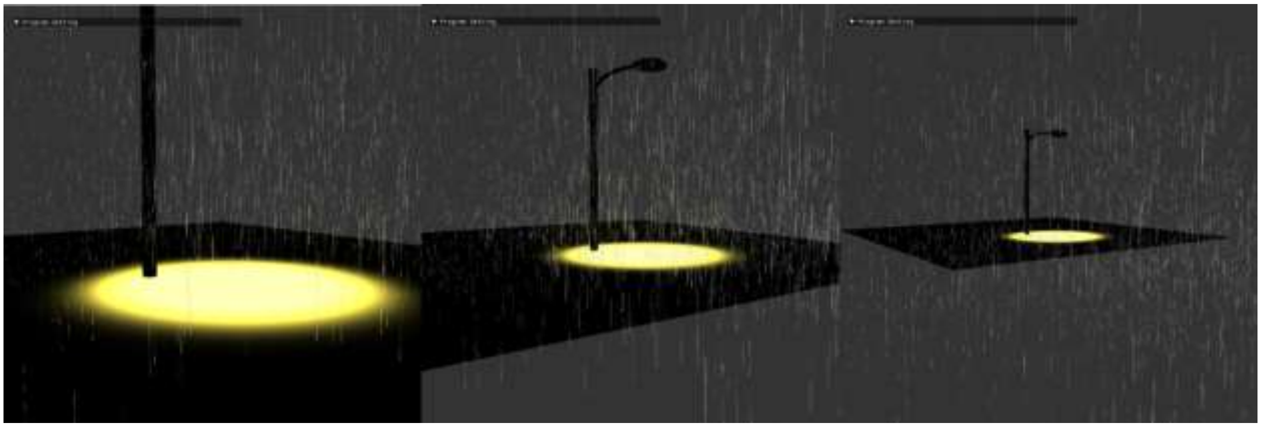
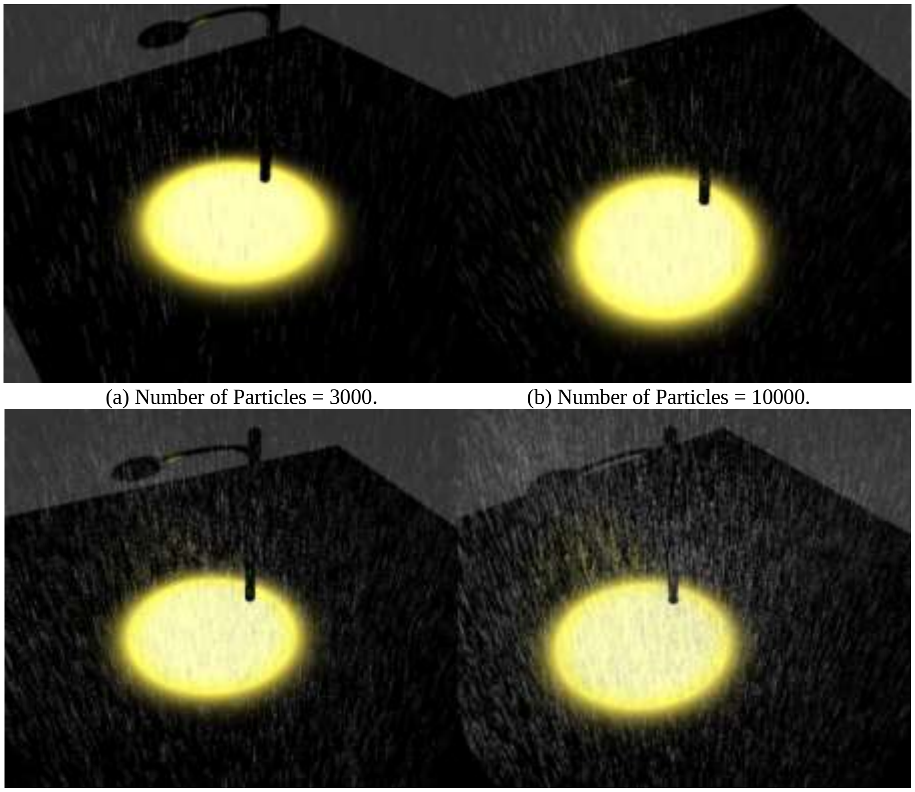


Fig. 9. Creus and Patow's Rendering Result.



(a) Camera Zoom in. (b) Camera Zoom Out. (c) Camera Zoom Out More.

Fig. 10. Proposed Model Rendering Results as Camera Zoom In and Out.



(a) Number of Particles = 3000. (b) Number of Particles = 10000.

(c) Number of Particles = 20000. (b) Number of Particles = 40000.

Fig. 11. Rendering Results of Proposed Model according to the Number of Particles.

V. Conclusion

As seen in previous experiment chapter, the proposed algorithm is somewhat inferior to other algorithms in terms of performance. However, in other researches, when the camera position is changing, the particle positions are rarely moving along the camera. Therefore, a very large number of particles are required and should be managed, thereby can waste the computer hardware resources. This study, however, creates a camera-dependent rain space that allows particles to be rendered only where the camera is rendered. In addition, it is possible to obtain the effect of making a large amount of rain falling even with a small number of particles.

VI. Future Work

Some limitations remain in our method, though. Particle system made with transform feedback is not intuitive to manage individual particles. Compute shader or GPGPU such as CUDA would provide much more flexibility in managing GPU threads [14, 15].

Another limitation is the way of using rain streaks textures. In our implementation, the textures did not choose according to the particular angle of light and camera conditions, although the texture database does have a lot of textures according to such parameters. Instead, this study randomly assigned one texture layer to one particle. Suggested algorithm ignored them because it turned out that it did not make a big different in terms of visual quality, though physical accuracy may be downgraded.

As future works, we have a plan to use GPGPU APIs to solve the problem of particle system and heavy computation [15, 16]. This allows us to take advantage of the flexibility of the GPU and improve the performance. Also, particle systems will be more intuitive and easier to manage. In addition, next study will consider the angle of light and camera conditions when chose the streak textures. We believe that rain simulation will be more physically accurated, realistic and effeicent.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(2018R1D1A1B07048414 and 2021R1A2C1012 316).

References

[1] K. Garg, S. K. Nayer, "Photorealistic Rendering of Rain Streaks," ACM Transactions on Graphics, vol. 25, no. 3, pp. 996-1002, 2006.

[2] Y. Weber, V. Jolivet, G. Gilet, K. Nanko, and D. Ghazanfarpour, "A phenomenological Model for Throughfall Rendering in Real-time," Eurographics Symposium on Rendering, vol. 35, pp. 13-23, 2016.

[3] S. Tarik, "Rain," Nvidia White Paper, 2007.

[4] A. Puig-Centelles, O. Ripolles, and M. Chover, "Creation Control of Rain in Virtual Environments," The Visual Computer, Vol. 25, no. 11, pp.1037-1052, 2009.

[5] C. Creus, G. A. Patow, "R4: Realistic Rain Rendering in Realtime," Computers & Graphics, Vol. 37, pp. 33-40, 2013.

[6] K. Nanko, Y. Onda, A. Ito, and H. Moriwaki, "Spatial Variability of Throughfall under a Single Tree: Experimental Study of Rainfall Amount, Raindrops, and Kinetic Energy," Agricultural and Forest Meteorology, 151, pp. 1173-1182, 2011.

[7] P. Rousseau, V. Jolivet, and D. Ghazanfarpour "Realistic Real-time Rain Rendering," Computer & Graphics, Vol. 30(4), pp. 507-518, 2006.

[8] L. Wang, Z. Lin, T. Fang, X. Yang, X. Yu, and S. B. Kang, "Real-Time Rendering of Realistic Rain," ACM SIGGARPH Sketches, pp. 156.

[9] A. Puig-Centelles, O. Ripolles, and M. Chover, "Creation Control of Rain in Virtual Environments," The Visual Computer, Vol. 25, no. 11, pp.1037-1052, 2009.

[10] W. T. Reeves, "Particle System – a Technique for Modeling a Class of Fuzzy Objects," ACM Transactions on Graphics, vol. 2, No. 2, pp. 91-108, 1983.

[11] Transform Feedback, Available online: https://www.khronos.org/opengl/wiki/Transform_Feedback (accessed on September 20, 2020).

[12] Rain Streaks Database, Available online: https://www1.cs.columbia.edu/CAVE/databases/rain_streak_db/rain_streak.php (accessed on August 10, 2020).

[13] OpenGL Array Texture, Availble online: https://www.khronos.org/opengl/wiki/Array_Texture (accessed on August 10, 2020).

[14] OpenGL Compute Shader, Availble online: https://www.khronos.org/opengl/wiki/Compute_Shader (accessed on December 10, 2020).

[15] CUDA Toolkit, Availble onlie: https://developer.nvidia.com/cuda-toolkit (accessed on December 10, 2020).

[16] OpenCL, Availble online: https://www.khronos.org/opencl (accessed on December 10, 2020).

A Secure Communication Process of Wireless Sensor Network Architecture for Smart Urban Environment Monitoring Applications

Rashmi S Bhaskar[1]
Research Scholar, Visvesvaraya Technological University
Karnataka, India

Dr. Veena S Chakravarthi[2]
Professor, Department of Electronics and Communication Engineering, BNMIT, Bengaluru, Karnataka, India

***Abstract*—Wireless Sensor Network has been increasingly used for remote monitoring system and its adoption in increasing exponentially for larger application too. However, there are various challenges associated with both resource management and security that roots up when the deployment scale goes massive and distributed in order. The proposed system considers a case study of smart city management where the problems associated with data transmission and security has been addressed. This is carried out using the provisioning of urban environment monitoring system that is an essential system for smart city projects to assure the citizens' better-quality well-being. The scalable and effective urban environment monitoring system requires a seamless transmission of the data from the sensor nodes to the analytics engine. The existing architectures are more designed to suit very specifically the use-cases. As a contribution, the proposed system introduces a cost-effective architecture for environmental monitoring in urban zones of smart city named as a Smart Sensor Surveillance System (4S-UEM). The core idea of the proposed system is to offer a balance between resource efficiency and resilience secure communication in large scale deployment of WSN considering smart city as deployment and assessment area. The proposed system makes use of urban geographical clustering process in order to develop an organized structure of sensor nodes. Different from any existing studies, the proposed system introduces data analytical engine followed by secure routing using gateway. The design of the proposed system is carried out using layered architecture of the communication model targeting towards a cost-effective, energy optimal, and secure data transmission to the analytics engine.**

***Keywords*—Wireless sensor network (WSN); sensors; smart city security; secure communication process**

I. Introduction

Urban environment monitoring (UEM) is an essential requirement for smart city projects [1] [2] [3] for various control mechanism applications. All these applications require a Smart Sensor Surveillance System abbreviated as '4S'. There exist multiple technologies for setting up surveillance systems for acquiring the data of importance. The analytics of these data correlates with the environmental changes [4]. Such technologies include surveillance camera systems and ambient-based monitoring technologies that suffer scalability, reliability, and the efficient correlation among spatial-temporal information [5]. The use of wireless sensor networks (WSN) as: "4S for UEM," i e., "Smart Sensor Surveillance System for Urban Environment Monitoring," aims to overcome these challenges and limitations. Therefore, this research study aims to design a standard architecture of WSN that provides a generalized architecture to comply with the "4S for UEM" requirements. The rationale behind arriving at a classic WSN for various 4S-UEM applications is overcoming the bottlenecks of the limitations of the suitability of one application-specific architecture to another. Typically, the data acquisition and analysis of many environmental parameters that include: temperature, humidity, traffic density on the road, vibration exerted on public infrastructure, pH-values in water, and many more parameters which leads to building smart control system for smart home, transportation, the safety of building & bridges, and water quality management, etc. as an actual application required for the smart city projects. Fig. 1 provides a snap-view of "4S-UEM" based applications and the intrinsic requirement.

As seen in Fig. 1, the intelligent system based on the architecture of the 4S-UEM requires four essential characteristics to be met by the suitable design of the 4S-UEM, and those characteristics are: {Lifetime, Accuracy, Coverage, Reliability}. Another additional requirement apart from these four is secure communication so that the realization of the 4S-UEM applications becomes possible in a real-time context. This paper proposes analytical modeling of the design of secure transmission through 4S-UEM to ensure accurate, reliable, and secure communication with optimal coverage.

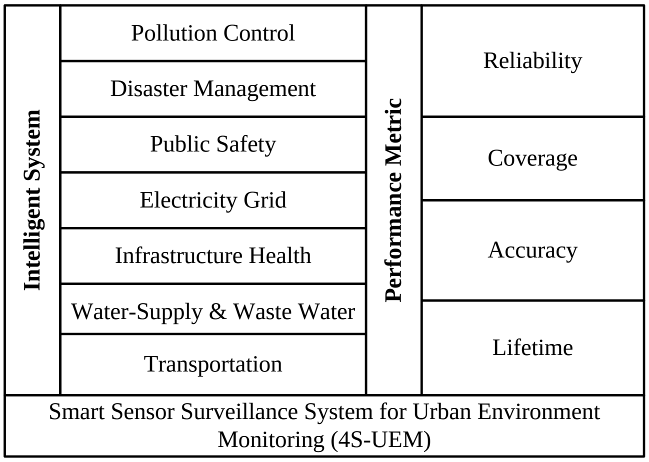


Fig. 1. 4S-UEM based Applications.

II. Review of Literature

This section describes various related work of design and development of the WSN based applications for assessing the methodological briefing of existing urban environment monitoring system (UEMS). The scope of the discussion is only limited to various resource optimization scheme which are considered as best scheme for resource management. It is because if the resource management is enhanced, eventually security scheme will be positively affected. The typical challenge in the UEMS faced is the issue of the optimal deployment of the sensor nodes that ensure maximum coverage and another is the secure communication. To provide a connecting link for transmission, a network of wireless sensor nodes is formed where a hierarchy of nodes as hop nodes, local and global sink nodes are the integral part of the WSN as static or dynamic WSN [6] depending upon the mobility of nodes and application requirements.

It is essential to consider a better network performance while designing the network deployment strategy with optimal coverage and connectivity [7]. The problem of the optimal coverage is formulated either as a greedy algorithm [8] or an integer optimization problem [9], which resembles the localization or placement problem of the facilities [10]. The sensor nodes' optimal numbers are computed using various methods like disk model, sector model, and geometry pattern model [11]. Whereas if the localization is known, then the problem of the optimal deployment in the Urban area of monitoring and deployment of WSN prefer an integer programming where popular method like i) divide-and-conquer, ii) simulated annealing, and iii) Genetic Algorithm, but these methods pose additional overhead of time complexity [12]. Irrespective of these methods, there exist some unique challenges for the deployment of the WSN for UEMS. Another important observation is that in most of the deployment, the base station's position is always fixed, which does not provide an optimal network performance [13].

There is no doubt that there are many research work being carried out towards secure routing in WSN with evolution of different methodologies and techniques. The existing methods are found to offer highly specific issue addressing scheme overlooking different challenges associated with the addressed issues. For an example, secure scheme cannot be only addressed using encryption but it equally demands resource management. Existing studies witness few direction of work aim like this integrated issue. Hence, the prominent research gap explored is that there are few studies which have linked practical energy retention with cost effective secure data aggregation in WSN considering challenging distributed and large scale area.

III. System Model for UEM

To monitor the urban environment, changes concerning temperature (T), pollution (CO2), and environmental parameters are significant for the meteorological department. The applications ranging from dairy farm monitoring to traffic management can be built on a generalized architecture: 4S-UEM, as in Fig. 2.

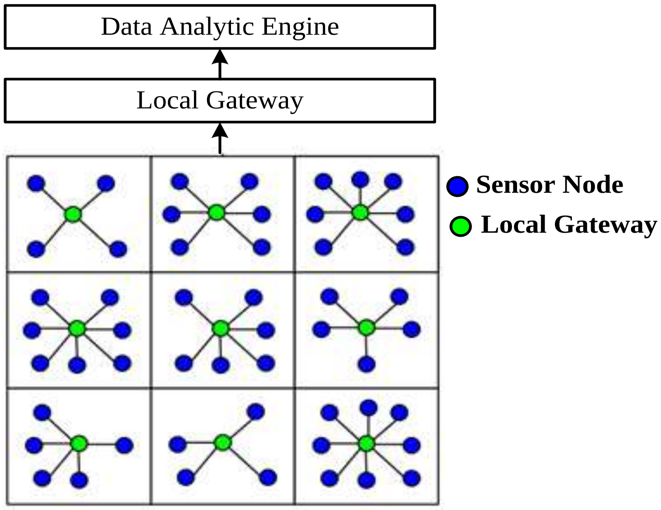


Fig. 2. Generalized Architecture 4S-UEM.

A. *Sensor Node and Data Analytics Engine Deployment*

An essential parameter in this part of implementation is basically number of sensors and simulation area. Introducing a data analytic engine is another contribution of proposed system which processes and analyzes the aggregated data unlike any existing clustering approach. The sensor nodes (Sn) are uniformly randomly distributed across a geographical area (A= L x L) of deployment in the region to monitor urban locality. An explicit algorithm-1 ensures the optimal coverage and connectivity while deploying the sensor nodes (Sn), local gateway (Lg), and data analytics engine (DaE) as intrinsic units of 4S-UEM. The Sn's deployment is modeled as a graph: G(V), where V is the vertex or a point that represents a Sensor node Sn.

Algorithm-1: Intrinsic unit deployment of 4S-UEM

```
Input: n, L,DaE
Output: G(V)
Process:
Start
for $\forall Sn_k \in n$
   [Sn.x,Sn.y]$\leftarrow f$rand(n) x L
   G(V1)$\leftarrow f$([Sn.x,Sn.y])
Initialize, {DaE.x, DaE.y}
   G(V2)$\leftarrow f$({DaE.x, DaE.y})
G(V)$\leftarrow$G(V1) $\cup$G(V2)
End
```

The WSN typically consists of 'n' number of sensor nodes (Sn) so that any sensor node ('Sn')$_k$∈n, s.t $2 \leq n$, where $n \in \mathbb{N}$ as a positive integer. The model takes 'n' and 'L'andDaEasindependent variablesto get G(V) as the dependent variable, i.e., G(V)$\leftarrow$*f*(n,L,DaE). The adjustment of connectivity and coverage of Sn and DaE for optimization takes place with variations in its localization coordinates of Sn and DaEas*(*[Sn.x,Sn.y]*)and* {DaE.x, DaE.y} respectively. The normalization of a logical layer of the architecture of the 4S-UEM is as in the Fig. 3, while the model imitates the process of deployment.

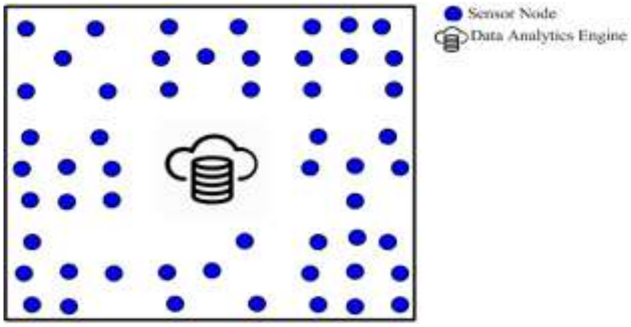


Fig. 3. Normalized Model Imitation for G(V) Containing Sensor Node and Data Analytics Engine.

B. Urban Geographical Clustering

In the smart city projects, the 4S-UEM synchronizes with the many sensor nodes deployed across the city or urban region by introducing an algorithm for Urban Geographical Clustering (UGC) to meet layer-wise communication that distance of communication is reduced. The prominent parameter of this part of implementation is local gateway which is responsible for performing translation services. The very basic requirement of UGC is to select a randomized local gateway (Lg) such that Lg ∈ {Sn}. A probabilistic approach considering the node's energy and the boundary conflict to avoid overlapping of spatial data overload is considered for the designing of UGC and selection of Lg. A straightforward process for the UGC is as in the algorithm 2. This approach establishes the communication between the lower layer of the sensor nodes or actuator with the local gateway.

Algorithm-2: Urban Clustering for 4S-UEM

```
Input: G(V)
Output: Cid, Lg
Process:
Start
Initialize, nC ← {m^2}, where m ∈ N
    $\vec{D} = |(nC - nLg)|$
    [Vmin, Id] ← $fmin(\vec{D})$
    $\vec{B}$ ← fbound(id, L)
    nC ← Id^2
    $\vec{S}$ = 1 to nC with step 1
    $\vec{S}$ ← freshape($\vec{S}$, $\sqrt{nC}$ . $\sqrt{nC}$ )
for membership of ∀ Sn into an urban-Cluster
    if
        $(Sn)_i.x \geq \vec{B}_i \wedge (Sn)_i.x < \vec{B}_{i+1} \wedge (Sn)_i.y \geq \vec{B}_i \wedge (Sn)_i.y < \vec{B}_{i+1}$
        update: Cid for ∀ Sn
    end
visualize the urban clustering
Call algorithm for selection of Lg
    Update: Lg in each communication cycle
End
```

The very basic assumption made while designing the architecture of 4S-UEM is that the number of local gateways (Lg) is as equal to the number of zones in the city or urban geographical region under monitoring i.e. Lg = nC, where nC = number of urban clustering, which is taken as a set: nC = {m^2}, where m ∈ a positive integer number. The nC is finally decided based onthenode with energy and a probabilistic approach, where the boundary constraint $\vec{B}$. To compute Bc, initially, the difference of number of local gateways (nLg) and nCas$\vec{D}$Is computed as in equation 1.

$$\vec{D} = |(nC - nLg)| \tag{1}$$

Then, the index (Id) and minimum value (Vmin) from $\vec{D}$ along with area component L goes to an explicit function to define the boundary as $\vec{B}$, which is a linearly spaced vector of L into Id+1 component, and the nC assigns as Id^2.Further, a series $\vec{S}$ as 1 to nC reshaped to the number of row and number of columns as $\sqrt{nC}$, say if nC = 16, then the $\vec{S}$ is as below:

$$\vec{S} = \begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} \tag{2}$$

To get the membership of ∀ Sn into a specific urban cluster, the localization of each $(Sn)_i \in$ n is taken as basic parameters to define its class membership with the respective urban cluster based on the condition with the boundary constraints $\vec{B}$ as in equation 3.

$$(\Sigma\nu)_\iota.\xi \ \varepsilon \ \vec{B}_\iota \perp (\Sigma\nu)_\iota.\xi < \vec{B}_{\iota+1} \perp (\Sigma\nu)_\iota.\psi \ \varepsilon \ \vec{B}_\iota \perp (\Sigma\nu)_\iota.\psi < \vec{B}_{\iota+1} \quad (1)$$

The corresponding identity to all the Sn is designated as cluster-id (Cid), with a potential of equal probability to be a Lg, and in this way, the urban clustering of all sensor nodes takes place, as seen in Fig. 4.

Section C describes the procedure for Lg selection along with the secure routing process as in Algorithm 3.

C. Secure Routing through Selected Local Gateway to the DaEin 4S-UEM

The very objective of the optimality of the architecture of the 4S-UEM is to meet the energy balance of the nodes as well as the network with the consideration of the optimal routing in the least consumption of energy as well the mitigation capacity for the data integrity against any kind of attacks which ensure the reliability of the data delivery for the analytics engine to provide the correct analysis on the real-time basis.

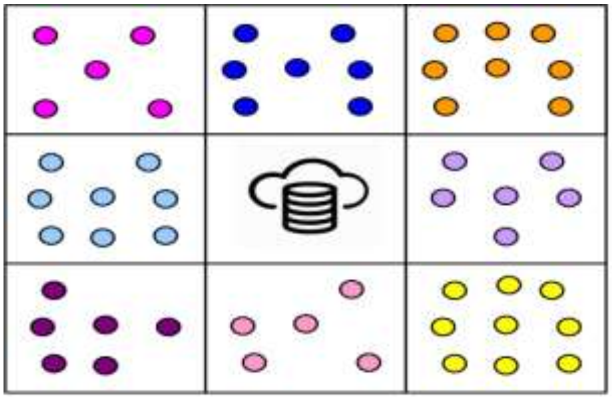


Fig. 4. Membership to each Sensor Node of 4S-UEM into an Urban Cluster.

Algorithm-3: Local Gateway selection in the Cluster of 4S-UEMand Secure Routing

```
Input:nC
Output:Lg
Process:
Start
Initialize the values of each in set:
    P: {Pl, cPl, Etx, ERx, Efs, Eamp, Ea,}
Setup process:
Initialize Pn
    Po← Rn1 x Pn
    Pkc←Rn2 x Pn
Generation of CG
    [CG1, CG2]←farbitary (Pn, ID)
Lg Selection process:
    Initialize ∀Sn  P(Sn | Lg → p %)
Update: Sn with energy in each communication cycle
        ASn←( nSn – nDSn)
        Lg←G(ASn) based on Th
        Th← $\frac{p}{1-p[Ccxmod(\frac{1}{p})}$
Call data authentication process (Algorithm -4)
    Transfer the packets after integrity authentication to DaE
    End
```

The system models take a set of parameters as independent variables, P={Pl, cPl, Etx, ERx, Efs, Eamp, Ea,}, where: Pl = size of the data packet , cPl = size of the control packet, Etx and ERx are energy required for transmission and receiving respectively. Here Efs and Eamp are the respective amplification energy in case of free space communication or dense network deployment, whichever is the case. In many of the current work, it is found that, while designing the architecture of the communication over wireless sensor network, only a basic first radio model is considered with the transmission and amplification energy into consideration [14]. Whereas the system model of 4S-UEMalso considers the energy required for the data to be aggregated or fused at the respective Lg ∈ uC(urban cluster) in the set of P, the performance resembles to the real-time constraints or context.

The characteristic of the sensor nodes (Sn) is a low capacity of computation and memory; thus, a lightweight security mechanism is proposed in the architecture of the 4S-UEM for the data authentication process in line with customized elliptical curve cryptography (ECC), while layer-wise communication. The first stage of this procedure is the setup stages as below:

- *Setup process:* In this process, the prominent parameter is a prime number (Pn) initiates to get the value of a primal Po as a product of a random number (Rn1) and the Pn and a public key center (Pkc) as Rn2 x Pn, where Rn2 is another random number. An explicit function:farbitary()is designed to take the Pnand the ID as an input argument to get the cyclic group. This operation's very basic function is to produce two cyclic groups (CG): {G1 , G2} with a single element Pn such that it produces a set of the invertible values with a single binary operation of property associative. The second stage of the system is the selection process of the Lg.

- *Local Gateway Selection Process:*IntheLg section's typical process, the Sn's energy plays a predominant role. The model validations occur with the varying probability of being Lg such that for ∀ Sn having P(Sn | Lg → p %). Further, the cluster-id (uCId) for ∀ uC and the Sn whose energy(E) < 0 is marked as a node as fault or dead node ID as (DSn) and the total number of such dead nodes = nDSn = ∑DSn. Therefore, the sensor nodes Sn in each cycle of communications to be considered as the nodes with some energy as (ASn) = nSn – nDSn is only considered to be set of the eligible Lg: Lg:→ {ASn}. The selection of the Lg from the set of {ASn} takes place based on a threshold node (Th) as in equation (4):

$$\frac{p}{1-p[Ccxmod(\frac{1}{p})} \tag{4}$$

If the Th < a random number between '0' and '1', the Sn ∈ ASn is treated as Lg from a group of G(Sn). In the equation 4, p = probability % of the ASn to be Lg, Cc = current communication iteration, G(Sn) = ∀ Sn ∈ ASn that has not been chosen as Lg in the last (1/p) communication iteration. This threshold (**Th**) is considered in the state of artwork for hierarchical communication LEACH and its variance [15][16]. Unlike any another traditional approach, the rest of the Sn receives the broadcasted cPl and decides to be part of this Lg based on the energy-based signal strength. The communication process further in the 4S-UEM communication process goes for the data authentication, aggregation then transmission to the DaE as described into the algorithm -4. Fig. 5 illustrates the current stage as Sn, DaE, and urban clustering and selection of the Lg.

It is essential to integrate a data-authentication process in communication to ensure the correctness of the data reaching the DaE, so that correct analysis of the data of the environments records provides accurate input or instructions to the control systems. However, the trade-off here is to handle the timely delivery in optimal energy way as the sensor data collected from the various uC takes some amount of energy that influence the overall network lifetime.

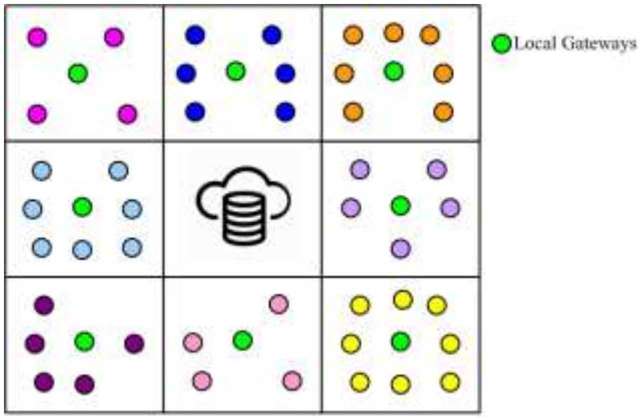


Fig. 5. Selection of Local Gateway (Lg) in the Urban Cluster(uC) of 4S-UEM Architecture.

The unique characteristics of the proposed communication process in the layer-wise approach of data from the individual sensors to the respective Lg and the aggregated or fused data transmission to the DaE in the optimal energy way even if the data gets authenticated by a very lightweight hashing mechanism of data integrity check within the sleep schedule cycle of the Sn, Lg, and DaE to avoid any kind of the collision, which is our future research direction to design an optimal tree-based routing for the channel resource allocation, which is beyond the scope of this stage of the architecture design. The details of the algorithm-4 are as below:

Algorithm-4: Data authentication, Aggregation, and transmission to the DaE as 4S-UEMSecure Routing procedure

```
Input:nC
Output:Lg
Process:
Start
Initialize, Id of Lg, and Id of ASn

    ∀ ASn ∈ uC, record messages as dP
    q ←ƒhash-1( Id-ASn, Pn)
    [q, Pk] ←ƒkeygen(Id-ASn, Rn1 )
Initialize, Signing operation
    Generate a Rn3
    P1 = Rn3 x Pn
    q1 ←ƒhash-2( P1, Id-ASN, dP)
    u1 ← $\sum[Pk,\ (Rn3 \times q1)]$
    Update: Signature
        Sig ←{P1,u1,Id-ASn, dP}
        sAm ←∑ mT x ∑ u1
        CG ∀ Lg: mark Lg → 1 based on Th
    Aggregation Verification:
    Call SPR
    End
```

In this process, the Id of ∀ Lg gets initialized, and for each uC, Initially, the Id of Sn ∈ respective Lg is used for finding the ASn, and further, ∑(ASn) participates in transmitting their data to the Lg for the aggregation process of the data packets. The secure aggregation process includes Lg selection, Sn data generation, signature verification, and routing simultaneously.

For ∀ ASn ∈ uC, the respective data packet (dP) depends upon the application used in a signing process. An explicit function ƒkg( ) takes the Id of ASn and Rn1 as used while creating the cyclic group in algorithm 3. The process of message authentication uses hashing algorithm in the key generation based on Id of ASn and Rn1, where ƒhash-1(Id-ASn) → q, where q = Pn x Id-ASn, which finally provides corresponding private key (Pk) for ASn using Pk = Rn1 x qusing an explicit function *ƒkeygen( )*. The next operation is the Signing operation, which takes {Id-ASn, dP, Pn, Pk}.

Initially, a random number Rn3 is generated in the signing process, which gives a first P1 as P1 = Rn3 x Pn, which passes into ƒhash-2( ), along with Id-ASN and dP, which gives a new hash value q1 as q1 = dP x Id-ASn x P1. Another, primal u1 is computed as in equation 5.

$$u1 = \sum[Pk,\ (Rn3 \times q1)] \tag{5}$$

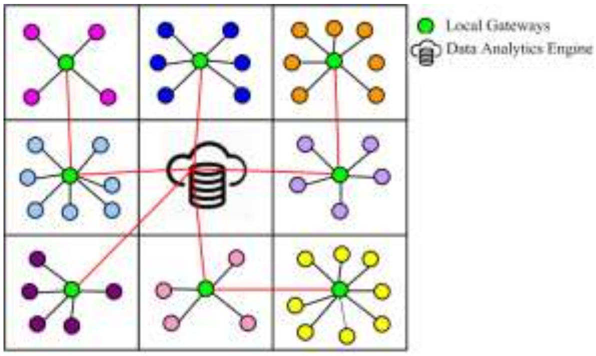


Fig. 6. The Transmission Process of Messages from uC through Lg to DaE of 4S-UEM.

Finally, the signature set includes {P1, u1, Id-ASn, dP} and each intermediate computation stores in message-term (mT) as u1 x Pkc, where Pkc comes from algorithm 3. Finally, ∑ mT x ∑ u1 assigns as a signature for aggregate messages at Lg (sAm).

The cyclic group values of for ∀ Lg which do not participate in the selection, and the flag is set to zero for all Lg, and the process continues with the comparison of a random number between 0-1 and the Th as in equation 4. Finally, the aggregation verification occurs, and the system calls for the shortest path routing either from Lg to DaE or as inter Lg routing. Fig. 6 illustrates the data transmission process of the proposed 4S-UEM.

Fig. 6 clearly illustrates two layers of data transmission as proposed in 4S-UEM. The first layer of transmission takes place between sensor nodes (Sn) to the Local gateway (Lg) and the second layer of transmission takes place between the Lg to the data analytics engine (DaE). In Section 4, observations of energy variation and the network life are described to understand the model's behavior while imitating in a sequential numerical computation platform.

IV. Performance Evaluation

There are two scenarios to validate the model, and these scenarios are 1) Traditional [6]-[13] and proposed approach of data communication without the data aggregation process and 2) Traditional and proposed approach of data communication with data aggregation process. Fig. 7 to 11 represents the graphical representation of numerical outcomes in Tables I to V with an intention towards assessing scalability. The outcomes are represented in the form of energy parameters on y-axis with increasing number of sensors to represents the influence of performance parameters. The outcome shows proposed system with better energy compared to existing approach.

TABLE I. First Node Death at Number of Urban Clusters = 4

No.of Nodes/ M	M1	M2	M3	M4
100	245	204	264	274
200	259	242	311	313
300	269	245	320	301
400	260	221	316	293
500	284	218	327	313

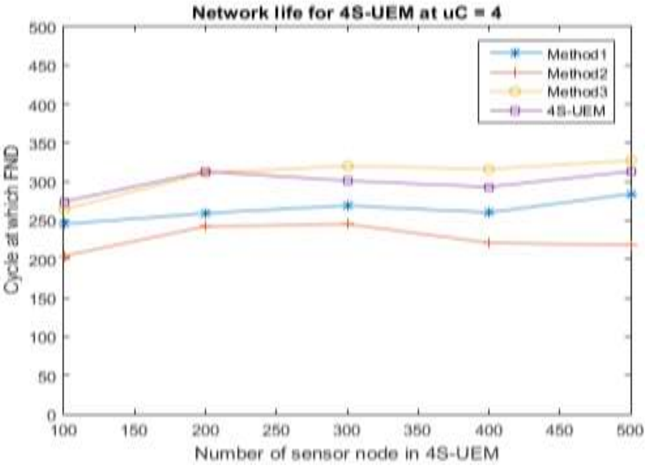


Fig. 7. First Node Death at Several Urban Clusters = 4.

TABLE II. FIRST NODE DEATH AT NUMBER OF URBAN CLUSTERS = 9

No.of Nodes/ M	M1	M2	M3	M4
100	106	123	117	170
200	240	119	259	225
300	250	142	289	224
400	262	151	309	203
500	261	142	312	209

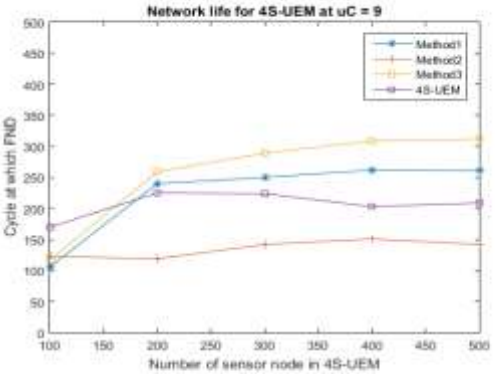


Fig. 8. First Node Death at Number of Urban Clusters = 9.

TABLE III. FIRST NODE DEATH AT NUMBER OF URBAN CLUSTERS = 16

No.of Nodes/ M	M1	M2	M3	M4
100	81	39	86	60
200	131	99	190	108
300	210	110	242	137
400	237	118	272	145
500	249	110	285	149

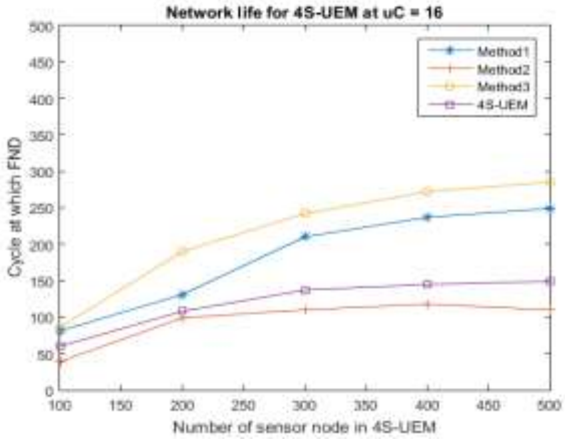


Fig. 9. First Node Death at Number of Urban Clusters = 16.

TABLE IV. FIRST NODE DEATH AT NUMBER OF URBAN CLUSTERS = 25

No.of Nodes/ M	M1	M2	M3	M4
100	80	39	80	82
200	103	99	82	69
300	85	110	150	107
400	167	118	228	100
500	210	110	246	104

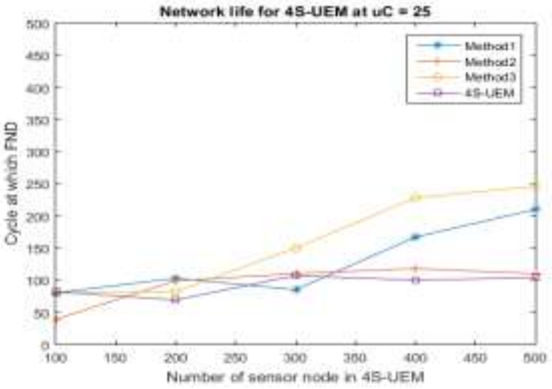


Fig. 10. First Node Death at Number of Urban Clusters = 25.

TABLE V. FIRST NODE DEATH AT NUMBER OF URBAN CLUSTERS = 36

No.of Nodes/ M	M1	M2	M3	M4
100	79	23	80	29
200	80	35	83	44
300	86	52	88	59
400	144	37	131	73
500	178	60	156	72

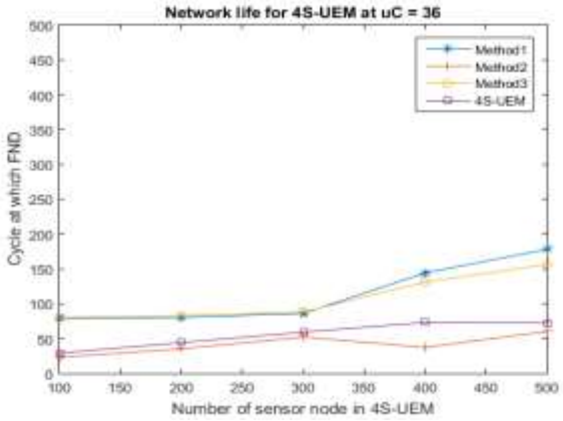


Fig. 11. First Node Death at Several Urban Clusters = 36.

V. Conclusion

The unified architecture proposed in this paper is named 4S-UEM, keeping in mind that to deploy a smart monitoring system for the urban environment monitoring purpose, the entire region of the urban zone is divided into the urban clusters where a local gateway is chosen smartly and intelligently. The local gateway collects the data and sends the data either directly or through inter-local gateway routing with the data verification provisioning. The important findings of this paper are i) proposed system offers approximately 47-58% of energy conservation over a large network of smart city, ii) The clustering model suits well with the public key encryption mechanism implemented unlike any existing methods, iii) the study offers no iterative scheme towards optimizing resources and hence it is highly cost effective. The 4S-UEM model provides a consistent result in terms of the network lifetime with a varying number of node density and the cluster. This work's future direction is to evolve a routing model so that priority-based routing takes place with further energy optimization during the channel allocation and avoidance of the collision of the packets utilizing optimal time slot scheduling of the radio.

References

[1] "The Smart City Building Blocks & Their Synergy with Smart Villages", https://www.ieeeottawa.ca/2020/08/the-smart-city-building-blocks-their-synergy-with-smart-villages/, August 2020.

[2] Z. Lv, B. Hu and H. Lv, "Infrastructure Monitoring and Operation for Smart Cities Based on IoT System," in IEEE Transactions on Industrial Informatics, vol. 16, no. 3, pp. 1957-1962, March 2020, doi: 10.1109/TII.2019.2913535.

[3] Z. Zhao, J. Wang, C. Fu, D. Liu, and B. Li, "Demo Abstract: Smart City: A Real-Time Environmental Monitoring System on Green Roof," 2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI), Orlando, FL, 2018, pp. 300-301, doi: 10.1109/IoTDI.2018.00049.

[4] R. Du, P. Santi, M. Xiao, A. V. Vasilakos and C. Fischione, "The Sensable City: A Survey on the Deployment and Management for Smart City Monitoring," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1533-1560, Secondquarter 2019, doi: 10.1109/COMST.2018.2881008.

[5] U. Gil, I. Pena, D. Guerra, D. de la Vega, P. Angueira and J. L. Ordiales, "Statistical Characterization of Medium Wave Spatial Variability Due to Urban Factors," in IEEE Transactions on Antennas and Propagation, vol. 59, no. 9, pp. 3498-3500, Sept. 2011, doi: 10.1109/TAP.2011.2161550.

[6] A. S. Alkalbani and T. Mantoro, "Security comparison between dynamic & static WSN for 5g networks," 2017 Second International Conference on Informatics and Computing (ICIC), Jayapura, 2017, pp. 1-4, doi: 10.1109/IAC.2017.8280615.

[7] J. N. Al-Karaki and A. Gawanmeh, "The Optimal Deployment, Coverage, and Connectivity Problems in Wireless Sensor Networks: Revisited," in IEEE Access, vol. 5, pp. 18051-18065, 2017, doi: 10.1109/ACCESS.2017.2740382.

[8] Saadi, N., Bounceur, A., Euler, R. et al. Maximum Lifetime Target Coverage in Wireless Sensor Networks. Wireless PersCommun 111, 1525–1543 (2020). https://doi.org/10.1007/s11277-019-06935-5.

[9] Zameni, M., Rezaei, A. &Farzinvash, L. Two-phase node deployment for target coverage in rechargeable WSNs using genetic algorithm and integer linear programming. J Supercomput (2020). https://doi.org/10.1007/s11227-020-03431-7.

[10] Wenfeng Zhou and Zhenping Li, "The multi-covering emergency service facility location problem with considering disaster losses," 11th International Symposium on Operations Research and its Applications in Engineering, Technology and Management 2013 (ISORA 2013), Huangshan, 2013, pp. 1-6, doi: 10.1049/cp.2013.2249.

[11] Deng X, Jiang Y, Yang LT, Lin M, Yi L, Wang M. Data fusion-based coverage optimization in heterogeneous sensor networks: A survey. Information Fusion. 2019 Dec 1;52:90-105.

[12] Y. Singh, J. A. Lone, P. K. Singh, Z. Polkowski, S. Tanwar, and S. Tyagi, "Deployment and Coverage in Wireless Sensor Networks: A Perspective," 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 2019, pp. 1-7, doi: 10.1109/ECAI46879.2019.9042091.

[13] Muhammed T., Mehmood R., Albeshri A. (2018) Enabling Reliable and Resilient IoT Based Smart City Applications. In: Mehmood R., Bhaduri B., Katib I., Chlamtac I. (eds) Smart Societies, Infrastructure, Technologies, and Applications. SCITA 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 224. Springer, Cham. https://doi.org/10.1007/978-3-319-94180-6_17.

[14] Mohammed SulaimanBenSaleh, Raoudha Saida, Yessine Hadj Kacem, Mohamed Abid, "Wireless Sensor Network Design Methodologies: A Survey", Journal of Sensors, vol. 2020, Article ID 9592836, 13 pages, 2020. https://doi.org/10.1155/2020/9592836.

[15] A. Yousaf, F. Ahmad, S. Hamid, and F. Khan, "Performance Comparison of Various LEACH Protocols in Wireless Sensor Networks," 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 2019, pp. 108-113, doi: 10.1109/CSPA.2019.8695973.

[16] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 2000, pp. 10 pp. vol.2-, doi: 10.1109/HICSS.2000.926982.

Customer Opinion Mining by Comments Classification using Machine Learning

Moazzam Ali[1], Farwa yasmine[2], Husnain Mushtaq[3], Abdullah Sarwar[4]
Adil Idrees[5], Sehrish Tabassum[6] , Dr.BaburHayyat[7], Khalil Ur Rehman[8]
Department of CS and IT University of Lahore
(UOL) Gujrat Campus, Gujrat, Pakistan

***Abstract*—In this era of digital and competitive market, every business entity is trying to adopt a digital marketing strategy to get global business benefits. To get such competitive advantages, it is necessary for E-commerce business organizations to understand the feelings, thinking and seasons of their customers regarding their products and services. The major objective of this study is to investigate customers' buying behavior and consumer behavior to enable the customer to evaluate an online available product in various perspectives like variety, convenience, trust and time. It performs data analysis on the E-commerce customer data which is collected through intelligent agents (automated scripts) or web scrapping techniques to enable the customers to quickly understand the product in given perspectives through other customers' opinion at a glance. This is qualitative and quantitative e-commerce content analysis in using various methods like data crawling, manual annotation, text processing, feature engineering and text classification. We have employed got manually annotated data from e-commerce experts and employed BOW and N-Gram techniques for Feature Engineering and KNN, Naïve Bays and VSM classifiers with different features extraction combinations are applied to get better results. This study also incorporates data mining and data analytics results evaluation and validation techniques like precision, recall and F1-score.**

***Keywords*—Customer comments; behavior mining; data mining; machine learning**

I. Introduction

A. Motivation

This study emphasis on need to develop some mechanism which ensures to get advantages form E-commerce users' generated data. A general approach reveals that to get opinion of other people before buying any product is common in online and offline shopping. But in this digital era, each customer has hundred or sometimes thousands of people readily available to provide valuable opinion and largely effect the decision-making process of new customers. Each customer looks for best product in lowest possible price. Actually, each customer tries to find the best commodity within his/her financial range along with surety of the justifiable quality attributes. Therefore, it is a normal practice to get neutral and genuine opinion of general public that is not generated by selling organization also not tempered by anyone else [1][2][3].

B. Consumer

A customer is an entity or person having an ability or will to buy the products and offered ventures available for purchase by advertising organization so that it may fulfil needs or want of an individual, family or a particular group of similar interests. A famous definition of consumer done by Mahatma Gandhi says, "A consumer is the most important visitor on our premises. He is not dependent on us. We are dependent on him. He is not an outsider to our business. He is part of it. We are not doing him a fulfil by serving him. He is doing us a fulfil by giving us an opportunity to do so" [4][5].

C. Consumer Behavior

Consumer behavior is characterized as "psychological, physical and social acts of potential clients as they have made their minds to access, assess, purchase and inform others about any item and its attributes". Consumer behavior is the study of single buyer, a group or organization about their course of selection, buying, utilization and disposing of commodities, services, ideas or experiences to satisfy his/her needs and also impacts of such process on the consumers and whole society. Consumer behavior varies from individual to the groups (like class students in school or college wear same uniform/dressing) and from group to firms (various groups at same place working together horizontally and vertically and decides collectively whether a product must be user by the firm or not) [6][7]. The customer opinion is often very important for advertising agency/marketer because it influences the market position as well as consumption of the product. Consumer behavior involves services and ideas as well as tangible products.

D. Internet Marketing

Internet marketing is utilization of internet as a medium to assess the showcasing and potential sale of merchandise. It has been proved highly beneficial by apply standard fundamental promotion systems on e-commerce applications [8]. Contrary to physical business strategies, online promotion and advertisement strategies have been proved far better with little hazards comparatively. Web showcasing process not only convenient for business community it also supports green solutions across the globe [9].

E. Purchase Decision

Buying decisions are defined as: "Several stages carried out by consumers before making a purchase decision on a product" [10]. According to the [11] buying behavior means activities of an individual who is involved in exchange of money for goods or services and also it involves some decision-making process to determine those activities. Consumer's decisions in buying a product involves physical and mental activities. The former refers to the direct activities for decisions making process while later involves assessment of product using some particular criteria.

F. Data Mining and E-commerce

It is a substantial undertaking to build a system which take advantages from mined knowledge. Studies revealed that some applications of data mining techniques on e-commerce data is comparatively less challenging as compared to other sort of data. For example, we can develop data mining system in E-commerce with much convenience rather than translating and correcting the data to make it suitable for data mining purpose. As data set is not collected manually or by survey but accessed electronically so it comparatively less noisy or sometimes contains no noise. Moreover, data set contains variety of vast and varied information as shown in "Fig. 1" [12] 1[3].

It has been analyzed that public information over E-commerce platforms play a vital role in success of regression models due to justifiable quantity of varied information. Therefore, E-commerce platforms provide very useful data and its inferences to produce a platform that is trustworthy for E-commerce customers [14]. They have discussed various applications of clustering and fuzzy set theory to determine issues in E-commerce platform through data mining application.

G. Consumer Behavior Mining

Consumer behavior mining is deals with web content mining which is concerned with valuable information extraction regarding users/customers opinion. Consumer behavior mining is almost a new subject in field of data mining as part of web mining and growth of E-commerce business has accelerated its growth significantly across the E-commerce applications, blogs and forums [15][16]. Consumer behavior towards online shopping instead of physical visiting markets and shops has been changed due to growth and profitability of E-commerce business. So along with online shopping habits, many people also prefer to get knowledge about public opinion regarding a particular merchandise before placing online order.

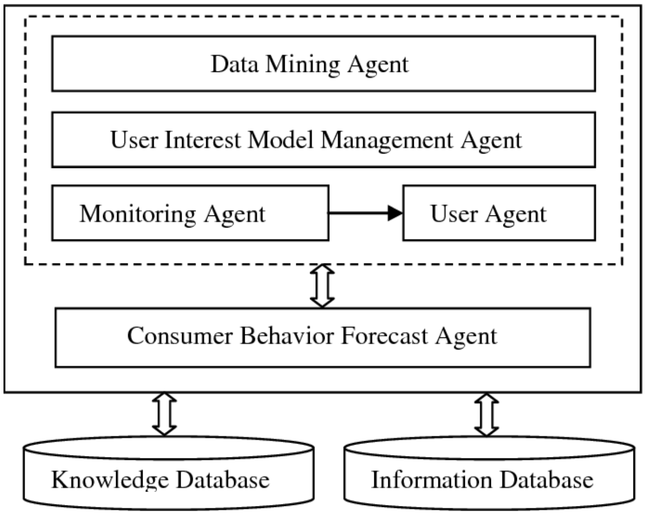


Fig. 1. Social Commerce as Sub Set of E-commerce and Social Media.

Consumer behavior is highly affected by public opinion but at same time availability of desired, accurate and quick information is a problem. Finding thousands of reviews regarding a product is not a big problem but finding summary of user opinion which is true reflection of customers' thought is a still a challenge across the E-commerce platforms [15]. Consumer opinion mining is not only useful at consumer end but also useful for manufacturing and advertisement companies the former get support about decision of transaction while the later are able to use opinions as customer feedback to improve their product accordingly while advertising agencies can easily find about individual customer opinion and post advertisements on his/her pages according to his/her choice [16].

H. Research Problem

Research objective is to identify the Consumer Behavior relevant elements related data in online E-commerce platforms. Validate Consumer Behavior relevant elements' related data with respect to customer opinions. To identity suitable text preprocessing techniques and selection of most relevant features for each Consumer Behavior relevant element. Develop a supervised machine learning based system which extract, classify into major predefined categories and preserve Consumer Behavior related knowledge in order to make this knowledge extensive, versatile, verified, easy to use and up to date.

II. Literature Review

A. Consumer Behavior

Customer experience is very important for every business organization where numerous products and vendors/brands are available. Customer experience optimizes and improves the online shopping experience of people which has been extensively increase after outbreak of pandemic Coid-19 across the world [18]. E-commerce companies and organization has experienced business prosperity due to increase in online demand of merchandise. Each business organization is string hard to improve their marketing by better understanding of customers' needs and priorities through analyzing customer buying behaviors[19][20]. Each organization trying its level best to retain their customers by offering the best shopping choices so that customers do not switch brands and shopping platforms. Organizations are also focusing to identify their potential customers segments by tracking their priorities, selections and expectations regarding products over the time [12]. It is necessary for business organizations to track, collect and organize consumer behavior possible data to develop business and analysis insight to take appropriate actions [13]. Recent studies revealed that enormous research has been conducted and a lot much more is underway to understand consumer behavior or to understand changes in customer activities over the time. Analysis of customer behavior is now an integral part of customer relations management strategies [17]. Data mining methods, tools and techniques are being incorporated to discover useful patterns using large amount of data collected by organizations. There is a variety of data mining models like clustering, association, classification, forecasting, regression, sequence discovery, visualization and machine learning models for data mining models like association rule mining, Logistic Regression, K-nearest neighbor, Neural Networks, Decision Trees, etc. E-commerce and other business organizations are collecting massive amount of data on daily basis in form of sale purchase transactions, customers profiles, cart management and product search data [21]. Such large data owner organizations are keen to unleash the potential behind this data and are also interested in mining the association among different data segments as shown in “Fig. 2”. These organization have firm believe that proper analysis of this large customer data can yield useful knowledge to get insight about consumer behavior [15]. They have proposed a customer segmentation system to discover and analyze frequent items searched by customers and their change over time. Authors employed association rule mining to discover useful and meaningful data patterns using database containing customers transactions records [22]. They also devised a strategy to automatically detect changes using customer profile and sales data of a particular period. The authors also proposed three types of changes in customer behavior: Unexpected Changes, Emerging Changes and Add/Perished Rule [19].

B. Consumer Behavior Mining

Big data technologies and their implementations are becoming cause of an immensely increasing information nowadays. Banking and insurance sector is seeking benefits of big data analytics and data mining to detect defaults and potential risks. In [23], the authors collected 22745 data samples and 14 attributes from Turkish Statistical Institution. This targeted to find the selection of best algorithm for classification to identify risk because of personal characteristics [24]. They incorporated and evaluated the several classification algorithms like Naïve Bayes, J48, Logistic Regression, Random Forest and Multilayer Perceptron were selected and their accuracies were evaluated using several evaluation techniques like Precision, Recall Roc Curve, etc. and most of them were found suitable to deal such type of data. Weka, a renown data mining tools was used to perform experimental methodologies as shown in “Fig. 3”. Data mining is extensively being used in field of medicines to predict various diseases using patient medical records especially detection and prevention of Diabetes Mellitus as it has deteriorated human, social and economic fabrics due its boosting penetration in all societies across the globe. traditional data mining techniques have been integrated with clinical research to advertise adverse effects of various diseases like Diabetes Mellitus, cancer and other diseases. These researches generally incorporate plain combinations or single classifiers. There are many comprehensive efforts are conducted to enhance the accuracy of systems with combination of multiple classifiers. [25][26] classified the diabetes mellitus in individual patients by using a set of risk factors and applied Bagging techniques of natural language processing with Adaboost, Decision Tree as a baseline experiment. The experiment was carried with Canadian Primary Care Sentinel Surveillance and three different adult teams. The Adaboost outperformed the bagging techniques an individual Decision Tree J48. Anarkali [27] have enlisted number of data mining application fields which are popular in recent times. Knowledge Driven Databases (KDD) and data mining are being employed in numerous fields to collect and analysis large chunks of information.

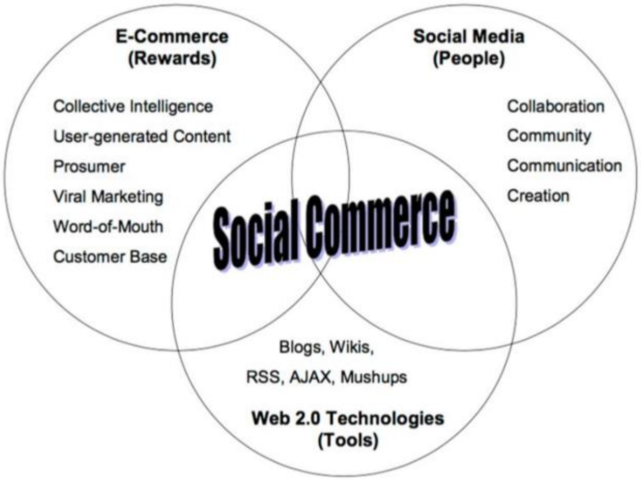


Fig. 2. Social Commerce as Sub Set of E-commerce and Social Media.

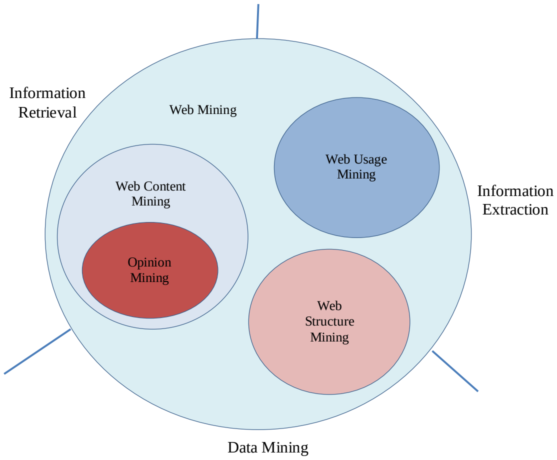


Fig. 3. Information Retrieval using Web Mining.

Integration of big data techniques in E-commerce web applications has also make it easier to collect structured, wide and large volumes of data regarding internal business process, marketing, supplier, venders and shipment. Therefore, big data analytics has open new avenues of opportunities for business companies related to E-commerce [28]. [29] presents many clustering, association and prediction techniques which are highly useful for E-commerce business. Sales forecasting, customer relationship management, customer retention management and basket analysis are common data mining models in E-commerce business. Major objective of this study is to review, implement and evaluate classification data mining techniques on user comments to classify them into different categories to assist online customer in decision making process.

C. Perceived Benefits

There are numerous benefits of internet shopping or online shopping but it there are also some risks involved which affect the consumer behavior towards online shopping [10]. The perceived benefits related to consumer satisfaction and belief that online shopping offers following benefits as compared to combinational shopping.

1) Convenience: Internet has made our life easier due to easy and quick access to the large number of desired products within seconds so we can buy almost everything with convenience. Seiders, Berry and Gresham (2000) says that convenience offers four benefits in buying process: search, access, possession and transaction as convenience is frontline attribute of online shopping which incites the consumer to go for online shopping. Furthermore, it is major predictor which prompts the consumer to go for online shopping. It also creates online buying willingness among customers. In online shopping, consumer does not need to leave his home/job or business and visit market personally for buying anything and consumer is not bound to keep cash in pocket for shopping for 24 hours [30].

2) Trust: Online shopping is characterized by time saving in terms of travelling to shops and then selection by visiting shops one by one. Therefore, online shopping saves time of customer but this statement does not stand true all the times due to late shipment of product. Generally, delivery time is often mentioned with the order and online customer confirms the order after verification of product delivery time [30].

3) Time: The time as perception of consumer that the vendor or seller will provide the best commodity in terms of price, quality, utilization and satisfactions. Most of people lack of trust in online shopping due to cheating and misconduct of e-commerce sellers. As there is lack of E-commerce cybercrime and needful legislation therefore it is sole responsibility of the seller to establish an environment of trust among people. Uncertainty in customers regarding reality of online shopping is one of the major obstacles in success of online business [30].

4) Trust: Online shopping is characterized by time saving in terms of travelling to shops and then selection by visiting shops one by one. Therefore, online shopping saves time of customer but this statement does not stand true all the times due to late shipment of product. Generally, delivery time is often mentioned with the order and online customer confirms the order after verification of product delivery time. But sometimes, due to disaster, natural calamity, power break down, strikes or transportation issues order does not arrive on time. Late delivery of product harms to trust of customer so timely arrival of product boost up customer confidence. Therefore there is strong relationship between time and online shopping but there must be some more attributes to better understand this relationship [30].

III. Methodology

This research aims to develop an E-commerce customer comments knowledge classification system based on consumer behavior attributes (listed as Convenience, Variety, Time and Trust), which depends upon tasks to be customers' comments given under any product on any E-commerce platform. We have targeted the social E-commerce platforms of online shopping to acquire data set where versatile people post variety of comments and answer each other's comments. One cannot only easily find type of his comments and discussion activity by using scrolling on web applications. By using proposed classification system, one can find the consumer behavior (Convenience, Variety, Time and Trust) relevant data and its overall status. Moreover, proposed system is not confine to classify all consumer behavior related comments, rather it is also ready to perform classification of other e-commerce customer comments under the umbrella of other classes or categories.

This study strives to build up an E-commerce customer comment data classification/categorization framework which heavily rely on the customer behavior attributes (listed as Convenience, Variety, Time and Trust). The propose system is developed by incorporating data mining and supervised machine learning text classification methods. This study employs various kinds of machine learning models like K Nearest Neighbor, Support Vector Machine, Naïve Bays, Character Based BOW and N-Gram on the state-of-the-art available data set.

A. Dataset Collection

First of all, data set is collected from E-commerce platforms using web agents or automated scripts by apply X query scripts using XML X paths. It requires up to mark professional expertise to extract exactly required and contextual data from large data dumps. E-commerce platforms contains large quantity of consumer behavior knowledge/comments and it can be utilized for consumer behavior mining (reference). Data set regarding consumer reviews about products which ranges across period from 2018 to 2020 contains almost thousands of comments about single product. But as per research requirement only 2000 total and 500 posts of each activity were included in the experiment.

B. Manual Annotation

E-commerce platforms or knowledge sharing communities and forums provide platform to fresh, experienced and professional online buyers/people to share their knowledge with people of their domain across the world and round the clock without any barrier of distance, language and level of expertise. All the members of an E-commerce platform cannot tag their post/comment with a specific category according to the chart mentioned above. If they do so then it will help the people to easily identify each comment under the umbrella of consumer behavior. By doing so, classification of each new comment will be started and data search and segregation will become easy for people who want to learn about opinions of other customers who have already utilized the product from the same platform. To visualize the comments data and better understanding of data set, it is categorized into 4 major categories and each category contain associated attributes related to consumer behavior. Every comment in selected data set is manually annotated and verified by E-commerce experts.

C. Attributes Associated with Each Class

1) Convenience: Convenience, Best, Appropriate, Easy, Great, Perfect, Useful, Functionality, Effective, Desire.

2) Variety: Variety, Forever, Compatible, Collection, Intended, Different, Diversity, Pack 3-1, Specific, Warranty.

3) Time: Recommend, Comfortable, Reliable, Excellent, Complains, Described, Quality, Satisfaction, Honest.

4) Trust: Time, Period, Delivered, Received, Pair, Come up with, Duration, Deal, Successfully.

D. Text Preprocessing

Customer comments classification process is actually started with data collection from E-commerce platform and then the most important function of data mining is performed that is called preprocessing of data. Data preprocessing is important because machine learning models are very sensitive to the features in each document. So pure and more relevant feature can be extracted by feeding the data set to some programming application. Before providing data to machine learning models, it is preprocessed and in case if data set is obtained from social media or E-commerce platform then it contains useless and noisy data which is mandatory to remove to get good results from machine learning classification models. This noisy data includes hashtags, symbols, domain terminologies, trends, smilies, web link, social media promotions, etc. so such noisy material is mandatory to remove to make the training data set clean. Preprocessing of E-commerce data set that has been manually annotated is made free form punctuations, semicolons, quotes, notations and the above said useless elements of language [31][32].

1) Tokenization: Tokenization process splits the data set text into the single elements or tokens by using a specify delimiter like space as shown in Table I. In this study we have used space in words as delimiter. As a text document is composed of linguistics elements and structures called as sentences and each sentence is further made up of grammatical units which are separated from each using space, full stop, hyphen, comma, slash. etc. Tokenization is first step in text preprocessing in which grammatical unit is converted into the tokens using some delimiter and output of this process is tokenized document [33].

2) Stop words removal: Stop words removal is next step to the tokenization. When all the text data set has been converted into individual tokens then unimportant words are removed because in a text data set, all the words are not important with respect to the document or class context. There are numerous words in the input text document which are less or least informative for the machine learning model [34]. Many words occurs multiple times in the document and such repeating words are also included in the noisy data. These less informative and repeating word are call stop words. Stop words removal is an important step in text preprocessing because stop words affect the features set extraction and ultimately efficiency of machine learning model.

As mentioned above, stop words gives little meaning of a word or context and are not helpful to the machine learning models. Stop words removal reduces the data set volume by removing useless or less important token or elements. Ultimately it reduces the computations coat and time. This technique is simple in use and helpful in increasing classification accuracy. Example of these words are ‘The’, ‘is’, ‘also’ etc which are frequently used in textual data. We remove all stop words in English language by using standard list of stop-words.

3) Identification of hash tags: Hash tags are also part and parcel of social media and E-commerce comments data set. Hash tags are commonly known a trends and it has specialty that it is stated with special symbol and it does not contains any kind of separator like space, comma or full stop. Hash tags are not dictionary words and also do not have any particular meaning so it is a good approach to remove Hash Tags to get useful and meaningful words which are easily recognized by dictionary or Word Net in Natural Language Processing.

4) Spell checking and correction: E-commerce text data set in form of comments or discussion in natural language contains multiple words which are not part of language dictionary but are understandable by the people belong to some specific domain. These words are not recognized by search engine optimization and also natural language processing techniques are confused of such words. Majority of such words is set of words which are not properly written and contain spelling mistakes. This problem of such words can be resolved up to some extent using regular expressions or mutually annotated data.

5) Stemming: Words are used in various forms in the English language text. It base might have different forms or behavior when it is used in different sentence structures like present, past, future tenses, singular, plural, adjectives, pronouns, etc [35][36]. English language has different laws for all these situations of a root word. No doubt each words comes from its root word of dictionary but it is used in different styles in the sentences. If we take an example of a word “Go” then we will come to know that it is used as “go”, “went”, “gone”, “going”, “goes”, etc. But dictionary has a single root word for all these words that is “Go”. These words have different posters or shapes but have same meaning as base word. To check the real frequency of each word in the document, stemming is applied. Stemming converts each word to its root word and so actual frequency of each word is calculated as shown in Table II. Stemming is a natural language processing based approach that is used in search engine optimization and information retrieval system. There are many stemming algorithms which can be used but in our study as E-commerce process activities are reported in English language by filling summary and description fields, we use Porter stemming technique to converts all tokens to their base stems.

TABLE I. EXAMPLES OF E-COMMERCE COMMERNTS

E-commerce Customers Comments
I got product in given time period
They promised me to send this in a duration of 2 months and they sent me

TABLE II. COMMENTS DATA AFTER STEMMING

Comments After Stemming
Get Product give time period
Promise send during month send

E. Feature Engineering

Feature selection and extraction that is collectively known as feature engineering employs multiple techniques to extract useful features from a given text document [37]. This study incorporates Bag of Words (BOW) and N-Gram (1 - 4) for the purpose of feature selection.

1) N-Gram: A text document is a connected sequence of n number of words/token/items/elements [40]. These items refer to some symbols, letters, and pairs of words so n-grams are combination of words patterns. Sequence of words to make a sense in a document is called N-gram where N is number of words in a pattern. Uni-gram describes single word, Bi-gram represents two, Tri-gram shows a sequence of three words and so on. Let's consider example of E-commerce comment: "this is a beautiful camera". Uni-grams of this text are 'this', 'is', 'a', 'beautiful', 'camera'. Bi-grams produced from this comment are 'this is', 'is a, 'a beautiful', 'beautiful camera' post. Tri-grams of this text piece are 'this is a, 'a beautiful camera as shown in Table III.

TABLE III. EXAMPLE OF CONVERTING TOKENS TO BI-GRAM, TRI-GRAM AND QUAD-GRAM

Tokens	I got product in given time period
Bi-grams	I got, got product, product in, in given, given time, time period
Tri-grams	I got product, got product in, product in given, in given time, given time period
Quad-grams	I got product in, got product in given, product in given time, in given time period

2) Bog of words: The Bag of Words (BOW) is also a features engineering model that counts all the useful feature without giving them weight with respect to document as well class corpora [38][39]. It counts the word number times it occur in the document and also does not accounts the sequence and order of words in the document. Each word in the Bag of Word model is independent of the next and previous words. Let's consider an example to understand the Bog of Words feature selection and extraction technique of Natural Language Processing. The cat is better than the dog and: The weather is better than yesterday.

F. Experimental Setup

All the above mentioned experiment are conducted on the same platform and IDE. Same text preprocessing techniques are applied on the whole data set prior experiment execution. Each experiment took 2400 customer comments as training input data set and 1600 customer comments as testing data et. Four classes (convenience, Variety, Time and Trust) were used to label the data set. Each class comprises of 1000 E-commerce customer comments which is further divided into 600 training and 400 testing instances/document/comment.

Preprocessing techniques are applied on the 70% of the data set and also have employed multiple preprocessing methods like tokenization, stop-words removal, word completion and spell checking, stemming are applied to each document of training data.

Initially, our data set was in form of raw comments which were obtained from E-commerce online shopping platforms. We applied tokenization as first step of preprocessing using natural language processing. To implement tokenization, we used 'space' between two words a delimiter. Following the tokenization, removal of stop words and noisy data is removed to sanitize our data set. At end of stop words removal, stemming is applied to get root words of each lingual element. For stemming, standard stemmers are used because each comment contains valid words after former preprocessing steps.

Machine learning model development and implementation steps are followed by preprocessing steps. Supervised machine learning algorithms are trained on the training data set which has undergone from text preprocessing steps. Feature selection and extraction is an important step which is characterized by the selection of most important and meaningful elements from each document with respect to each class. As we have mentioned in literature review and other sections that in supervised machine learning, algorithms are mostly learned on the probabilities. Unique words are selected and extracted from preprocessed training data etc. Frequency of each word is calculated using BOW model. This probability matrix measures the probability of each unique word in any class. Probability for each word is then calculated from training data. Square root of each probability is computed by calculate square root of each value.

1) Experiment 1: The first experiment is performed using Bag of Word Model feature extraction technique. The major goal behind performing the experiment using BOW model approach to illustrate the classification efficiency and accuracy of supervised machine learning model. The experiment is carried out using preprocessed training data set from which features are extracted Bow approach and algorithms are trained on BOW features. The hash map produced from this experiment is given in the following Table IV.

TABLE IV. RESULT OF EXPERIMENT 1 WITH BOW MODEL

Features	No. of Features	Naïve Bayes	SVN	KNN
Bag of Words	2317	71.52	**72.23**	68.95

2) Experiment 2: Experiment 2 is extension of experiment 1 and it yields better results by combining the token by N where N >= 1 & N = <4. The major objective of this experiment is to clearly demonstrate the difference of accuracy of supervise machine learning models from the experiment 1 by incorporating N-Gram technique with BOW approach. This experiment brings into use same data set as used in the previous experiment. In this experiment, the whole programming environment remain same as in previous experiment. It also brings into use preprocessed data set as discussed in previous section. Words with highest frequency carry less information for a class as compared to least frequency. Supervised machine learning model is trained using training data set of 2400 comments and on the basis of this training, unlabeled and unseen comment (testing data set) is classified. Same features along with threshold value are applied as in last experiment. In feature extraction phase, different N-gram patterns are applied. We used N-Gram where value of N ranges from 1 to 4.

- Unigram pattern: consist of one word for extracting semantic information.
- Bigram pattern: consist of two word for extracting semantic information.
- Trigram pattern: consist of three word for extracting semantic information.
- Quad gram pattern: consist of four word for extracting semantic information.

Following the preparation of N-gram pattern matrix is used to compute the score of each individual class which is computed by total number of words in corpus divided by their individual frequencies as shown in Table V. Training model is generated and each classifier used in the experiment.

TABLE V. RESULT OF EXPERIMENT 2 WITH BOW AND N-GRAM MODEL

Features	**No. of Features**	**Naïve Bayes**	**SVN**	**KNN**
Unigram	2317	64.33	71.11	62.02
Bigram	13957	73.68	**82.32**	71.25
Trigram	11695	57.12	53.78	53.39
Quad Gram	17658	43.98	40.35	48.61

G. Comparison of Accuracies of Data Mining Techniques Followed in Experiment

Performance of the proposed classification system in terms of accuracy measures is depicted in Table V. It is showing accuracy measures of all the classifiers along with their N-Gram values. There are three classifiers employed with different N-Gram values from N=1 to N=4. Results reveal that Support Vector Machine (SVM) has outperformed with the best accuracy. SVM gives 71.11% and 82.32% accuracy with uni-gram and bi-gram respectively while KNN is better with tri-gram whereas the over performance of KNN is less than SVM. Naïve Bays stands between SVM and KNN in performance as shown in Tables VI, VII, and VIII.

H. Comparison of Precision and Recall Scores

Precision, recall and F1-score measures for all given algorithms using Uni-gram, Bi-gram, Tri-gram and Quad-gram. All three algorithms are applied one by one on the same data set to get accuracy, precision, recall and F1-score.

We have also performed K-fold (where k = 10) cross validation mechanism to validate the confusion matrix and data set authentications. We have divided our dataset into 10 folds (f1, f2, f3 . . . . f10) of equal size. We trained all the classifiers one by one to f1 to f9 folds and then from f1 to f8 and tested for f9 fold and so on.

The overall performance of all three algorithms are compared and their comparison report is given in the graph below. The Support Vector Machine which has been proved the best algorithm in our text classification system and it has better values for confusion matrix as compared to the KNN and Naïve Bays as shown in Table IX.

TABLE VI. RESULTS EVALUATION FOR NAÏVE BAYS ALGORITHM

Class – Naïve Bayes	**Precision**	**Recall**	**f1-score**
Analysis	0.76	0.75	0.76
Synthesis	0.81	0.66	0.75
Evaluation	0.72	0.57	0.82
Implementation	0.70	0.80	0.40

TABLE VII. RESULTS EVALUATION FOR SVM ALGORITHM

Class – SVM	**Precision**	**Recall**	**f1-score**
Analysis	0.61	0.76	0.69
Synthesis	0.74	0.71	0.10
Evaluation	0.75	0.69	0.51
Implementation	0.62	0.80	0.82

TABLE VIII. RESULTS EVALUATION FOR KNN ALGORITHM

Class – KNN	**Precision**	**Recall**	**f1-score**
Analysis	0.78	0.73	0.75
Synthesis	0.81	0.76	0.74
Evaluation	0.80	0.65	0.83
Implementation	0.72	0.68	0.39

TABLE IX. RESULTS EVALUATION FOR N-GRAM MODEL

Features	**Class – Naïve Bayes**	**Precision**	**Recall**	**f1-score**
Unigram	***Naïve Bayes***	0.71	0.72	0.71
	SVM	0.82	0.66	0.72
	KNN	0.75	0.68	0.64
Bigram	***Naïve Bayes***	0.88	0.82	0.84
	SVM	0.67	0.69	0.68
	KNN	0.68	0.68	0.62
Trigram	***Naïve Bayes***	0.54	0.66	0.60
	SVM	0.48	0.54	0.51
	KNN	0.46	0.48	0.53

IV. Discussion

This study incorporate data mining and machine learning technqiues along with NLP to develop an automated system to categrorize the products w.r.t given classes. We have used N-grams, BOW and TF-IDF technqies for features extratcion. There could be feature engineering tecniques which might improve the system relaibality. Our proposed system is good with proper grammar and well spelled words but in case of slangs in data set might confuse the system. The most important thig to discuss is that this is an initial approach to classify consumer comments under the given classes to assist both, seller and buyer. Therefore, we don't have any benchmark to compare our results.

V. Conclusion

This study has evaluated the proposed machine learning model with various data analytics techniques as mentioned in literature like accuracy, precision, recall and F1-score. Briefly, SVM algorithm along BOW and bi-gram features engineering techniques is proved winning classifier in the proposed E-commerce customer comment classifier. This work opens new avenues to E-commerce customers and sellers to get a quick status of customer opinion in four important contexts which helps many customers to decide the about purchase of product. At the same it assist sellers to improve their product or services in the given four context (convenience, variety, time and trust) using data mining and machine learning and Natural Language Processing. This wok not only demonstrates the usefulness of machine learning and data mining in E-commerce business development and customer assistance but also identify preprocessing techniques and the important features engineering methods.

VI. Future Work

This work is concerned to classification of E-commerce comments data using supervised machine learning model by incorporating BOW and N-gram feature engineering methods. Currently we have selected/preferred those words/features with highest value to its respective class. In future work, we shall employ semantic and syntactic features engineering techniques to select features with contextual relevance. In this way, we will get the improved percentage accuracy i.e. to consider different dimensions of vectors in E-commerce customer's comments classification and other aspects of E-commerce related text classification.

References

[1] Chen, X., Duan, S., & Wang, L. (2020, June). Comments Prediction Model on Emotional Analysis Based on Bayes Classification. In Journal of Physics: Conference Series (Vol. 1575, No. 1, p. 012020). IOP Publishing.

[2] Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: a study of consumer perceptions. Electronic commerce research and applications, 2(3), 203-215.

[3] Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., & Bezbradica, M. (2019, September). Customer purchase behavior prediction in e-commerce: a conceptual framework and research agenda. In International Workshop on New Frontiers in Mining Complex Patterns (pp. 119-136). Springer, Cham.

[4] Raorane, A., & Kulkarni, R. V. (2011). Data mining techniques: A source for consumer behavior analysis. arXiv preprint arXiv:1109.1202.

[5] Voinea, L., & Filip, A. (2011). Analyzing the main changes in new consumer buying behavior during economic crisis. International Journal of Economic Practices and Theories, 1(1), 14-19.

[6] Nayyar, T. (2019). Analyzing Customer Buying Behavior.

[7] Saeed, R., Lodhi, R. N., Rauf, A., Rana, M. I., Mahmood, Z., & Ahmed, N. (2013). Impact of Labelling on Customer Buying Behavior in Sahiwal, Pakistan. World Applied Sciences Journal, 24(9), 1250-1254.

[8] Pahwa, B., Taruna, S., & Kasliwal, N. (2017). Role of Data mining in analyzing consumer's online buying behavior. International Journal of Business and Management Invention, 6(11), 45-51.

[9] Familmaleki, M., Aghighi, A., & Hamidi, K. (2015). Analyzing the influence of sales promotion on customer purchasing behavior. International Journal of Economics & management sciences, 4(4), 1-6.

[10] Singh, M., Jyani, L., Verma, R., Rajpurohit, L., & Goswami, P. Analysis of Consumer Behavior on SCM Related Factors Using Data Mining: A Case Study of the Indian E-Commerce Industry.

[11] Altunan, B., Arslan, E. D., Seyis, M., Birer, M., & Üney-Yüksektepe, F. (2018, August). A data mining approach to predict E-Commerce customer behaviour. In The International Symposium for Production Research (pp. 29-43). Springer, Cham.

[12] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[13] Anggoro, M. A., & Purba, M. I. (2020). The Impact of Attractiveness of Ads and Customer Comments Against to Purchase Decision of Customer Products on the User of Online Shop Applications in the City of Medan. Jurnal Ilmiah Bina Manajemen, 3(1), 1-9.

[14] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[15] Mattosinho, F. J. A. P. (2010). Mining Product Opinions and Reviews on the Web. Technische Universitat Dresden.

[16] Gowtamreddy, P. (2014). Opinion mining of online customer reviews (Doctoral dissertation).

[17] Martin, M. (2017). Predicting ratings of amazon reviews-techniques for imbalanced datasets.

[18] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[19] Chen, X., Duan, S., & Wang, L. (2020, June). Comments Prediction Model on Emotional Analysis Based on Bayes Classification. In Journal of Physics: Conference Series (Vol. 1575, No. 1, p. 012020). IOP Publishing.

[20] Mahmud, B. U., Bose, S. S., Majumder, M. M. R., Arefin, M. S., & Sharmin, A. Ecommerce Product Rating System Based on Senti-Lexicon Analysis.

[21] Belém, F. M., Silva, R. M., de Andrade, C. M., Person, G., Mingote, F., Ballet, R., ... & Gonçalves, M. A. (2020). "Fixing the curse of the bad product descriptions"–Search-boosted tag recommendation for E-commerce products. Information Processing & Management, 57(5), 102289.

[22] Sahib, S. M. Customers Buying Behaviour towards Online Shopping-A Study of Rural People in Southern Western Region of Punjab Dr. Monica Bansal.

[23] Kunjithapatham, K. A., & Santhanakannan, A. A Study on Consumer Behaviour towards Online Shopping in Kanchipuram Town.

[24] Sahu, M. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh. Journal of Xi'an University of Architecture & Technology Issn No, 1006, 7930.

[25] Vishwakarma, A., Ojha, T., & Mohanty, D. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh.

[26] Prabhakumari, K., & Silviya, M. T. ANALYSING CONSUMER ATTITUDE AND BEHAVIOUR TOWARDS ONLINE SHOPPING IN COIMBATORE CITY.

[27] Anggoro, M. A., & Purba, M. I. (2020). The Impact of Attractiveness of Ads and Customer Comments Against to Purchase Decision of Customer Products on the User of Online Shop Applications in the City of Medan. Jurnal Ilmiah Bina Manajemen, 3(1), 1-9.

[28] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[29] Bhatti, A., & Rehman, S. U. (2020). Perceived benefits and perceived risks effect on online shopping behavior with the mediating role of consumer purchase intention in Pakistan. International Journal of Management Studies, 26(1), 33-54.

[30] Altunan, B., Arslan, E. D., Seyis, M., Birer, M., & Üney-Yüksektepe, F. (2018, August). A data mining approach to predict E-Commerce customer behaviour. In The International Symposium for Production Research (pp. 29-43). Springer, Cham.

[31] CLARK, M., RUTHVEN, I., HOLT, P., SONG, D., & WATT, S. (2012). OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[32] RANDERSON, K., BETTINELLIB, C., FAYOLLE, A., & ANDERSON, A. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[33] JENNINGS, B., TSATTALIOS, K., & CHAKRAVARTHI, R. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University. Scientific Reports, 6, 20504.

[34] Johnson, I. M., & Copeland, S. M. (2008). OpenAIR: The Development of the Institutional Repository at the Robert Gordon University. Library Hi Tech News.

[35] SACHSE, S., SILVA, F., IRFAN, A., ZHU, H., PIELICHOWSKI, K., LESZCZYNSKA, A., ... & KUZMENKO, O. OpenAIR@ RGU The Open Access Institutional Repository at Robert Gordon University.

[36] Basani, Y., Sibuea, H. V., Sianipar, S. I. P., & Samosir, J. P. (2019, March). Application of Sentiment Analysis on Product Review E-Commerce. In Journal of Physics: Conference Series (Vol. 1175, No. 1, p. 012103). IOP Publishing.

[37] Li, N., & Zhang, P. (2002). Consumer online shopping attitudes and behavior: An assessment of research. AMCIS 2002 proceedings, 74.

[38] Nagra, G. K., & Gopal, R. (2014). Consumer Online Shopping Attitudes and Behavior: An Assessment towards Product Category. International Journal of Marketing and Technology, 4(5), 54.

[39] Sahu, M. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh. Journal of Xi'an University of Architecture & Technology Issn No, 1006, 7930.

[40] Vishwakarma, A., Ojha, T., & Mohanty, D. Factors Affecting Online Buying Behaviour in Youth with Special Reference to Chhattisgarh.

Spoken Language Identification on Local Language using MFCC, Random Forest, KNN, and GMM

Vincentius Satria Wicaksana[1], Amalia Zahra, S.Kom, Ph. D.[2]
Computer Science Department, Bina Nusantara University, Jakarta, Indonesia

***Abstract*—Spoken language identification is a field of research that is already being done by many people. There are many techniques proposed for doing speech processing, such as Support Vector Machines, Gaussian Mixture Models, Decision Trees, and others. This paper will use the system using the Mel-Frequency Cepstral Coefficient (MFCC) features of speech input signal, use Random Forest (RF), Gaussian Mixture Model (GMM), and K-Nearest Neighbor (KNN) as a classifier, use the 3s, 10s, and 30s as scoring method, and use dataset that consists of Javanese, Sundanese, and Minang languages which are traditional languages from Indonesia. K-Nearest Neighbor has 98.88% of accuracy for 30s of speech and followed by Random Forest that has 95.55% of accuracy for 30s of speech, GMM has 82.24% of accuracy.**

Keywords—Gaussian mixture model; random forest; K-Nearest Neighbor; spoken language recognition; MFCC; GMM; KNN

I. Introduction

Indonesia is an archipelago in the Southeast Asia region. Indonesia consists of large islands and small islands spread from Sabang to Merauke, so that the Indonesian State is dubbed the Archipelago State. Indonesia is recorded as having 17.504 islands, therefore, the State of Indonesia has a variety of ethnicities, races, religions and cultures. Because of this diversity, Indonesia has a wide variety of languages, ranging from Javanese, Sundanese, Bahasa Batak, and many more. Therefore, some of regional languages in Indonesia are also extinct because the language is not widely used in the regions anymore. To prevent it from extinction, by collecting the dataset of regional languages to be studied, it can help to prevent extinction of regional languages, because when building a classification technique, a large scale of dataset is needed and by developing the SLI, the application can be used as a leading component of applications such as translators used to classify regional languages, which later can be used in speech-based information systems, speech-based translate, and others.

Referring to the problems above, the need for information technology solutions in the field of Spoken Language Identification is getting higher. Due to the Spoken Language Identification technology, Indonesian citizens who do not understand regional languages when visiting other areas or when tourists come to an area where residents do not understand Indonesian, can be helped by this technology.

In spoken language identification it takes several steps to identify a language, starting from the sound extraction such as MFCC [8] method to techniques for classifying language. Several techniques are used to classify languages, including deep neural networks [1], Gaussian mixture models [2][5][8], support vector machines [3], Random Forest [3], and others. Random Forest, KNN and GMM are technique that is quite widely used for classification, this technique has some parameter that can be tuned, which is very useful to increase accuracy. There has been a lot of research on spoken language identification, but no one has done research on spoken language identification that uses segmented speech. In this study, Random Forest, KNN and GMM will be used for classification techniques, the accuracy will be obtained from segmented speech in 3 seconds, 10 seconds, and 30 seconds. This study will examine spoken language identification using the techniques mentioned above and using the GMM technique which is often used in spoken language identification which is segmented at 3 seconds, 10 seconds, and 30 seconds as the baseline.

II. Literature Review

Spoken Language Identification (LID) [4] is a process for determining the identity of the language spoken. LID [4] is based on the linguistic properties of language obtained from the results of speech extraction. The performance of an LID system depends on the amount and reliability of information and how efficiently it is integrated into the system.

The sound structure of a language can be categorized into acoustic-phonetic, phonotactic, and prosodic. Acoustic-phonetic is one of the structures of the sound which is related to the analysis of the physical properties of the sound being spoken. While phonotactic is a sound structure related to the syllable structure of a language, for example Languages such as Dutch, English, and German allow a large number of consonants at the beginning and at the end of the syllable. In contrast, Maori, which is spoken in New Zealand, only allows syllables consisting of a vowel, two vowels, or a consonant plus a vowel. Prosody is a structure of sound related to rhythm, intonation, and stress of sound, for example, Mandarin has the same letter but has a different intonation, for example the word “ma” with high intonation means mother, while “ma” word with intonation drops later to ride means horse.

In [6], the authors discussed about spoken language identification using Shifted Delta Coefficient and Shifted Delta MLP as feature extraction and using Gaussian Mixture Model and Support vector Machine as classification technique.

In [2], the authors discussed about spoken language identification using Mel-frequency cepstral coefficients as feature extraction and using Gaussian Mixture Model as classification technique. In his research, to improve the performance of the Gaussian mixture model in his research, the total mixture was gradually added to get optimal results. In his research it was also explained that Tamil and Telugu languages have good performance by using mixture values between 128 to 512. Starting with 32 mixtures which produced low accuracy, namely 70% for Tamil and 85% for Telugu. Then the mixture components are increased little by little to get the desired results. When the mixture was increased to 128, the resulting accuracy was almost 100% for Tamil, while for Telugu, it got 100% accuracy. Then the mixture component is increased again to 512 which is the best point for the classification of the two languages, when the accuracy rate of both reaches 100%. If seen from the results above, it can be concluded that by gradually increasing the mixture component, it can improve performance in language identification. Despite of that, research that conducted by [7], discussed that by increasing the mixture component, the performance of the technique will increase, but the higher the mixture component will increase the computation cost or increase the time in computing.

In [3], the classification methods used are Support Vector Machine and Random Forest and for the feature extraction used are MFCC, LPC, and a combination of the two techniques, to find out which technique provides the best accuracy. To conduct an evaluation, [3] used the IIIT-H dataset which contains 5000 samples, consisting of 6 languages, namely Hindi, Telugu, Bengali, Marathi, Tamil, and Malayalam. Then 300 samples were taken randomly from the IIIT-H dataset using a 16kHz sound signal. The evaluation was carried out in 2 phases, the first phase was carried out using the MFCC feature with Support Vector Machine and Random Forest. Meanwhile, for the Random Forest technique, the resulting accuracy is 75.9% and 74.3% for the SVM technique. The second phase is carried out using the LPC feature with the same classification technique. The Random Forest technique used produces an accuracy of 61.5% and 67.16% for the SVM technique. In [3] is not stated why total trees of 300 has the best performance compared to the lower total of trees.

In [9], the study was conducted using a Gaussian mixture model as a technique for classifying languages which will be used to compare feature extraction. For feature extraction, MFCC, SDC, and a combination of both are used. In his research, the database used is the Arunachali Language Speech Database (ALS-DB), which consists of 6 languages, namely Adi, Apatani, Galo, Nyishi, Hindi and English. The experiment was carried out using the GMM with total of 1024 mixture component, MFCC features with 12 cepstral coefficient numbers.

Research conducted by Gupta et al. (2017) using Support Vector Machine and Random Forest. In his research, it is stated that by combining Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) will increase the accuracy of the language identification. In [3], it is not explained why Gupta et al choose total trees of 300 compared to the lower total trees. In [3], only the total accuracy of the features extracted from the frame is explained, it is not explained how the accuracy from classification using 3s of speech, 10s of speech, and 30s of speech. Therefore, this paper will use MFCC and Random Forest to see the effect of the total trees on the accuracy and expanding the testing method by segment the duration of the test from frame to 3s speech, 10s speech, 30s speech and briefly discusses the performance of computation time when conducting model training using three traditional language from Indonesia. This paper will use KNN as a rarely use technique in language identification to see if its fits as classifier for spoken language identification and use GMM that widely use in language identification as baseline.

III. Proposed Method

A. Dataset

At the data collection stage, speech data collection will be collected both from the internet and from native speakers of the local language. In order to get the correct acoustic of the language, native speakers who are fluent in the regional language are needed. The dataset obtained will be divided into 2 datasets, the first is for the training dataset and the second is the test dataset. With a ratio of 70:15:15 , that is, 70% of the dataset will be used for training the dataset, 15% of the dataset will be used for validation dataset, and the other 15% of the dataset will be used for test dataset. The distribution of the dataset can be seen in Table I below.

The Javanese and Sundanese dataset will be obtained from *openslr* and Minang dataset language will be collected from *youtube* and recorded Speech.

B. Pre-Processing

After the data is collected, the Javanese and Sundanese language dataset will be sorted again. After that, the speech that obtained from YouTube will be processed again.

For Javanese and Sundanese dataset, each dataset will be combined into 80 minutes long. As for dataset that obtained from *youtube*, the first step is to remove the noise, song, background song, and unnecessary item from the recording. Same as Javanese and Sundanese dataset, Minang dataset will be combined or cut to achieve 80 minutes long of training dataset.

For test dataset, each language will have 20 minutes of speech data, the speech data will be divided into 3s, 10s, and 30s of speech data. After the recording is cut, the recorded file will be changed to wav format, because the compressed data produced by wav has a sound quality that is almost the same as the original sound. The sound will be resampled to 44.1 kHz and use a bit rate of 32 kbps.

TABLE I. Total Duration for each Language

Language	Total Duration	Training Dataset Duration	Validation Dataset Duration	Testing Dataset Duration
Sundanese	200 Minutes	140 minutes	30 minutes	30 minutes
Minang	200 Minutes	140 minutes	30 minutes	30 minutes
Javanese	200 Minutes	140 minutes	30 minutes	30 minutes

C. Model Development

Every speech has its own characteristics, to get the characteristics of the speech, feature extraction will be executed. This study will use MFCC feature, the MFCC feature will be extracted from the pre-processed input signal. The MFCC feature will be presented using vector c, which a set of vector C has the value of C_1, C_2, C_3, C_4, C_5, C_6…C_n. In vector C, n represent the total coefficient will be extracted from the speech every frame. This study will use *python* and library provided from *librosa* to extract the MFCC feature. Total coefficient that will be used in the experiment is 13 and the total length of the frame is 25 milliseconds.

After extracting the features from the testing data and training data, a model development will be carried out. The model will be developed using random forest, Gaussian mixture model, and K-Nearest Neighbor. Random forest is a classification algorithm consisting of many decisions' trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. This study will use the random forest, Gaussian mixture model, and K-Nearest Neighbor and the library is provided by *sklearn*. Before using this model, the best parameter was determined for each method.

The Random Forest technique, there are several parameters will be tuned, such as n_estimator, criterion, max_depth, and max_sample_leaf. N_estimator is used to determine the total trees to be used, the random forest that used in [3] will be used as baseline. This total tress parameter will be used to compare the best n_estimator for this dataset. Criterion is used to measure the impurity of a node. Max_depth represents the depth of each tree in the forest. The deeper the tree, the more splits it has, and the more it captures more information about the data. The parameter for KNN will be tuned are the type of weight, total leaf size and number of neighbors and the last classifier is GMM that widely used for LID [2], [7], and [9].

D. Evaluation

In this study, the total percentage of accuracy will be measured. After the experimental process is complete, the evaluation results will be entered into a table for further observation. From table below, method column used to list the method that used, and the duration row is used to classify the average accuracy based on the duration of the speech.

The experiment flow can be seen from Fig. 1 below.

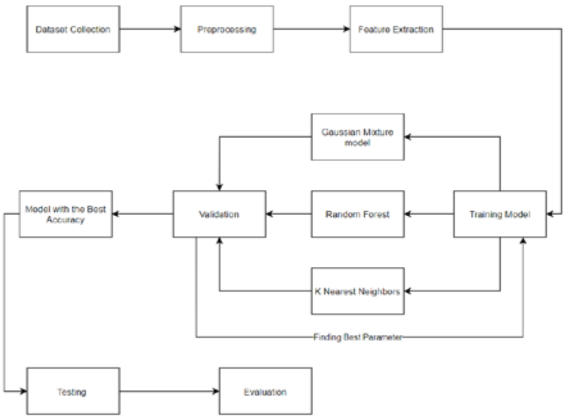


Fig. 1. Experiment Flow.

IV. Result and Discussion

The experiment is focused on comparing performance between three classifier techniques on three segmented duration. Classification was performed on three different language. Once the models are trained and the feature are extracted, the classifier was used to classify the dataset. The accuracy score of classification between three languages are reported in terms of percentage of accuracy. Tuning the min sample leaf parameter for random forest using validation dataset are recorded on Table II.

From Table II, the result shows us that there is no significant accuracy difference between each parameter, so min_sameple leaf will be set to the default value, the default value is 5.

From Table III, the result shows us that the higher max_depth number, the better accuray it has. The accuracy increases periodically as total max depth value increased and reach its peak at total max depth of 50, but from 50 to 100, there is no significant increase between 50 and 100, the score almost similar, so total max_depth of 50 will be used for this parameter.

From Table IV, the result shows us that there is no significant difference between criterion gini and entropy, but Entropy is more computationally heavy than gini, so it will increase the time computation, so gini will be used for this parameter.

TABLE II. Finding Best Parameter for MIN_Sample_Leaf.

Parameters			**Accuracy**		
Min sample leaf	**N_estimator**	**N_jobs**	**3 s**	**10 s**	**30 s**
5	100	16	87.44%	90.18%	94.72%
10	100	16	87.11%	89.90%	94.44%
15	100	16	86.78%	89.35%	94.16%
25	100	16	86.47%	89.07%	93.88%

TABLE III. Finding Best Parameter for Max_Depth

Parameters			**Accuracy**		
Max Depth	**N_estimator**	**N_jobs**	**3 s**	**10 s**	**30 s**
10	100	16	83.47%	83.99%	84.44%
15	100	16	85.16%	86.11%	88.61%
25	100	16	86.44%	88.25%	92.49%
50	**100**	**16**	**88.05%**	**91.29%**	**96.11%**
100	100	16	87.88%	91.57%	96.38%

TABLE IV. Finding Best Parameter for Criterion

Parameters			**Accuracy**		
Criterion	**N_estimator**	**N_jobs**	**3 s**	**10 s**	**30 s**
Gini	100	16	88.00%	91.48%	96.38%
Entropy	100	16	87.78%	90.83%	96.11%

From Table V, the result shows us that there is no significant difference between 100 to 250 totals of trees, so total trees of 100 will be used in this experiment, gini as criterion parameter, min_sample_leaf of 5, and max_depth of 50 for this technique; fFor GMM, the search of the best parameter for number and mixture and covariance type.

In Table VI, full covariance type has the best score rather than the other covariance type. Next, the best parameter for GMM total mixture was tuned. [2] GMM accuracy was used as baseline.

In Table VII, by increasing total number of mixtures, the accuracy increased that is stated in [2]. Despite of that, there is a significant decrease from 128 to 256. In order to confirm that there are no errors in the code, the test was run 3 times and still has the same score, so the test is stopped at 256 and is decided that the total mixture of 128 has the best score for this dataset. Next, KNN is widely used on machine learning, but it is rarely used in language identification, so in this paper, KNN technique was used to determine if KNN is suitable for language identification or not. There are several parameters that will be used, such as K, weight, and size of leaf.

TABLE V. FINDING BEST PARAMETER FOR N_ESTIMATOR

Parameters				**Accuracy**		
N_estimator	**Min_sam-ple_leaf**	**Max Depth**	**criterion**	**3 s**	**10 s**	**30 s**
100	5	50	Gini	87.52%	90.92%	95.27%
150	5	50	Gini	87.63%	90.27%	94.72%
250	5	50	Gini	87.62%	90.09%	94.72%

TABLE VI. FINDING BEST PARAMETER FOR COVARIANCE TYPE

Parameters			**Accuracy**		
Covariance Type	**GMM Mixture**	**N_jobs**	**3 s**	**10 s**	**30 s**
Full	128	16	71.77%	77.77%	82.77%
Tied	128	16	61.86%	66.20%	70.00%
Diag	128	16	64.33%	70.92%	80.27%
spherical	128	16	62.19%	64.90%	64.16%

TABLE VII. FINDING BEST PARAMETER FOR NUMBER OF MIXTURE

Parameters			**Accuracy**		
Covariance Type	**GMM Mixture**	**N_jobs**	**3 s**	**10 s**	**30 s**
Full	16	16	45.45%	44.35%	57.77%
Full	32	16	56.25%	53.05%	63.88%
Full	64	16	75.61%	79.72%	75.55%
Full	**128**	**16**	**76.55%**	**80.09%**	**81.38%**
Full	256	16	46.27%	45.09%	48.33%

From Tables VIII and IX, there is not any significant difference from each parameter, each parameter has similar score, so the default value will be used for type of weight from the library which is uniform and total leaf of 20 because total leaf of 20 has the best better accuracy than 30 and 40, and it has less computation time. Next, the best parameter for K will be tuned for this dataset; the result can be seen from Table X.

From the table above, it can be concluded that by increasing the total of K, it will increase the score. Total k of 10 and 20 has the best accuracy, there is a slight difference between two of them, but the computation time and complexity must be considered, because by increasing the total K, the computation time and complexity will be increased. So, for this technique, total K of 10 will be used to get the accuracy from testing dataset. After getting the best parameter for each technique, each technique will be tested using dataset to get the accuracy result for each technique with tuned parameter.

In Table XI represents performance of feature with different classifier. The KNN classifier gives the highest score from 3 second of speech until 30 second of speech. RF gives almost similar score to KNN. GMM gives the lowest score in language identification from 3 sec, 10 sec, and 30 sec.

TABLE VIII. FINDING BEST PARAMETER FOR WEIGHT

Parameters			**Accuracy**		
Type of Weight	**K**	**N_jobs**	**3 s**	**10 s**	**30 s**
uniform	5	16	86.91%	91.01%	96.11%
distance	5	16	86.63%	90.64%	95.83%

TABLE IX. FINDING BEST PARAMETER FOR SIZE OF LEAF

Parameters			**Accuracy**		
Size of Leaf	**K**	**N_jobs**	**3 s**	**10 s**	**30 s**
20	5	16	86.91%	91.01%	96.11%
30	5	16	88.91%	90.01%	94.11%
40	5	16	87.91%	92.01%	93.11%

TABLE X. FINDING BEST PARAMETER FOR K

Parameters				**Accuracy**		
Size of Leaf	**Type of Weight**	**Total K**	**N_jobs**	**3 s**	**10 s**	**30 s**
30	uniform	5	16	86.36%	90.74%	95.83%
30	uniform	10	16	88.18%	93.61%	98.88%
30	uniform	15	16	85.38%	88.51%	91.38%
30	uniform	20	16	87.83%	91.12%	96.11%

TABLE XI. MODELS ACCURACY

Technique	**Accuracy**		
	3 s	**10 s**	**30 s**
MFCC + KNN	88.19%	93.61%	98.88%
MFCC + GMM	72.35%	80.59%	82.24%
MFCC + RF	87.66%	90.64%	95.55%

V. Conclusion

In this paper, this paper compares the widely use technique in language identification which is GMM and rarely use technique, which is KNN, and another technique called random forest to see if its good in segmentation speech or not.

From Table XI, KNN has the highest accuracy in each segment, with a score of 88.19% for 3s, 93.61% for 10s, and 98.88% for 30s, then followed by RF which has an accuracy score of 87.66% for 3s, 90.64% for 10s, and 95.55% for 30s. And GMM has the lowest score for each segmentation. However, when doing training and testing model for each technique, KNN use longer computation time when compared to Random Forest because KNN is called the lazy learner.

It can be concluded that KNN and RF is better than GMM and has the best accuracy for Javanese, Sundanese, and Minang.

Suggestion for the future research is to get more Minang dataset variation, such as high pitch speech, low pitch speech, and the other and using another feature extraction technique to see if there is a better feature extraction technique for KNN and RF.

Acknowledgment

Thanks to Amalia Zahra, S.Kom., Ph. D. from Bina Nusantara University for the useful discussions, guidance, and assistance in this research.

References

[1] Heracleous, P., Takai, K., Yasuda, K., Mohammad, Y., & Yoneyama, A. (2018). Comparative study on spoken language identification based on deep learning. *European Signal Processing Conference*, *2018-Septe*, 2265–2269. https://doi.org/10.23919/EUSIPCO.2018.8553347.

[2] Athiyaa, N., Jacob, G., Science, C., Anna, R., College, G., & Phil, M. (2019). Spoken Language Identification System using MFCC features and Gaussian Mixture Model for Tamil and Telugu Languages. 4243–4248.

[3] Gupta, M., Bharti, S. S., & Agarwal, S. (2017). Implicit language identification system based on random forest and support vector machine for speech. *2017 4th International Conference on Power, Control and Embedded Systems, ICPCES 2017*, *2017-Janua*, 1–6. https://doi.org/10.1109/ICPCES.2017.8117624.

[4] Lee, C. H. (2008). Principles of Spoken Language Recognition. *Springer Handbooks*, 785–796. https://doi.org/10.1007/978-3-540-49127-9_39.

[5] Chellappa, R., Veeraraghavan, A., Ramanathan, N., Yam, C.-Y., Nixon, M. S., Elgammal, A., … Reynolds, D. (2009). Gaussian Mixture Models. Encyclopedia of Biometrics, 659–663. doi:10.1007/978-0-387-73003-5_196.

[6] Wang, H., Leung, C. C., Lee, T., Ma, B., & Li, H. (2013). Shifted-delta MLP features for spoken language recognition. *IEEE Signal Processing Letters*, *20*(1), 15–18. https://doi.org/10.1109/LSP.2012.2227312.

[7] Kumar, A., Hemani, H., Sakthivel, N., & Chaturvedi, S. (2015). Effective preprocessing of speech and acoustic features extraction for spoken language identification. *2015 International Conference on.*

[8] *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings*, *May*, 81–88. https://doi.org/10.1109/ICSTM.2015.7225394.

[9] Sarmah, K., & Bhattacharjee, U. (2014). GMM based Language Identification using MFCC and SDC Features. *International Journal of Computer Applications*, *85*(5), 36–42. https://doi.org/10.5120/14840-3103.

Conceptualizing Smart Sustainable Cities: Crossing Visions and Utilizing Resources in Africa

Dr. Ahmed Al-Gindy[1]
Dr. Ziad Elkhatib[4]
Faculty of Engineering
Canadian University Dubai
Dubai, United Arab Emirates

Eng. Aya Al-Chikh Omar[2]
Faculty of Engineering and Technology
Aldar University College
Dubai, United Arab Emirates

Eng. Mariam Aerabe[3]
Faculty of Engineering and Technology
Aldar University College
Dubai, United Arab Emirates

***Abstract*—Recent advancements in technologies enabled the development of smart cities to be more effective and possible. Smart cities depend on intelligent systems, artificial intelligence, the internet of things, control system, and many more advanced technologies. Sustainability challenges and problems worldwide, with smart and sustainability concepts, reflect almost mutual goals. It includes improving and providing the essential life services for all people efficiently while depending on sustainable, clean, and renewable energy with considerations of different economic, educational, health, social and environmental aspects in the city. In this research, a cost analysis process has been implemented to ease the implementation and resource utilization of smart and sustainable cities in Africa. The challenges and difficulties of those implementations are summarized.**

Keywords*—Smart cities; sustainable energy; renewable energy; internet of things; artificial intelligence

I. Introduction

Now-a-days, over three-quarters of the world’s population lives in urban areas. It is noticed that the population growth in urban areas is higher than in rural areas. As such, cities are struggling with an overpopulation crisis triggering a shortage in capital resulting in issues across the society due to social and economic imbalance [1]. As technology is evolving, the concept of developing smart communities is becoming more desirable as the use of technologies like Artificial Intelligence (AI) and the Internet of Things (IoT) can help in solving a variety of problems in coexisting societies. For instance, IoT provides new ways for cities to utilize data to control traffic, reduce emissions, and allow better resources to provide people with a safe and clean smart environment. Smartness is a dynamic term that can be expressed as 'green technology.' The term “Smart City” is not just referring to a digital city; it also highlights the need for talented people and residents who are attentive to the priorities of having a smart, prosperous, and resilient city. It also refers to combining intelligence, political strategy, and infrastructure to create holistic urban services. Several aspects can aid in developing a Smart City, some of these aspects are: “Smart Mobility”, “Smart Living”, “Smart Economy” and many more that, when combined and strategized, can translate the practical meaning of having environmentally friendly, sustainable and smart city [2]. In general, no definitive or conclusive method can determine whether a city is smart or not. The significant primary definition relies on Information and Communication Technology (ICT) and the advanced immerging technologies like AI, IoT, etc.

On the other hand, implementing these technologies in underdeveloped countries is heading towards urbanization and industrialization. This implementation requires an understanding of smart city elements and how they are applied across various industries in the cities. This paper introduces a proposed framework and architecture in underdeveloped countries that can emphasize the use of new technologies and their performance to improve the quality of life in these inhabitants.

II. Literature Review

During the 2000s, cities worldwide radically changed, where the concentration turned to the vast development in digital technologies, knowledge, and the increase in environmental awareness and concern. Along with this development, new concept and model for the cities have been introduced. Over the last decades, these connected issues related to digital development started to initiate a new heading towards Smart Sustainable Cities.

According to [2], smart cities arise due to the intelligent use of technologies in different ar, such as the education field, health sector, transportation, and energy use.

Sustainability and sustainable city's urban development concept brings awareness of the production and use of available resources required in industrial areas, transportation, education, health, business, and investments.

Sustainable urban concept collaborates in increasing environmental awareness in naturally available resources in smart cities [3].

The author in [4] demonstrates that evaluating a smart city should be based on different factors like encompassing sustainability and quality of life and the collaboration of technological elements.

The concept of sustainable cities became popular in the 1990s [5]. It refers to the relationship between the conversion of resources exploitation to be consistent with the present and future requirements and the development of countries' economic and environmental sustainability aspects.

Considering all the aspects, sustainable urban development can be defined as the combination of urban modernization to be smart with the use and protection of available sustainable resources.

The energy system for smart cities requires a significant share of renewable energy in the different aspects of the city. The energy system becomes more diverse, and it could be integrated into other areas with a crucial impact on the cities' development process.

There are several examples of smart and sustainable cities worldwide on different continents, as illustrated in Fig. 1 [6].

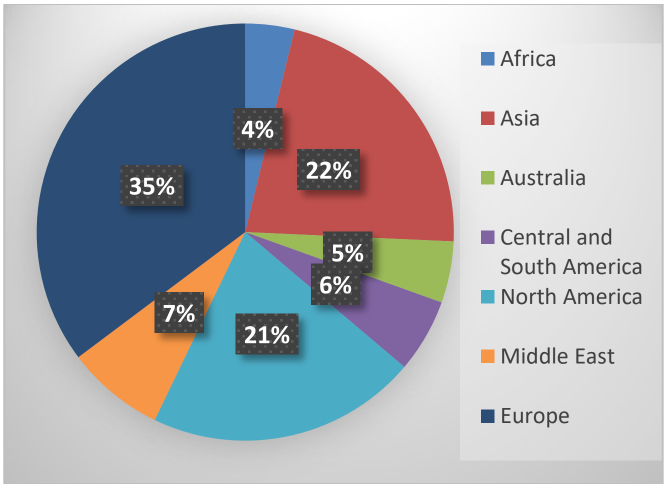


Fig. 1. The Total Sustainable and Smart Cities from Arcadis's.

Although Europe seems to lead the transformation of traditional cities to smart and sustainable cities, around 35% of the total cities engaged in sustainable and smart cities, 65% of the cities are moving forward steps to convoy the technology advancement and development. For example, a city like Dubai in the Middle East has a problem with the dramatic increase of the population, leading to a significant increase in environmental pollution. That is why Dubai is integrating an energy strategy action plan to depend on renewable energy and reduce energy demand and waste by 30% by 2030. But what about the underdeveloped countries in Asia and Africa?. Do they have the ability to be transformed into smart cities? Could technological advancement provide optimal solutions to develop the cities' urban?. Do underdeveloped countries modernize their infrastructure and work based on an intelligent system?.

Building cities' innovative transformation requires applying the most effective, advanced, and developed information and communication technologies. In addition, artificial intelligence and the internet of things are also needed. The underdeveloped overpopulated countries have significant challenges when providing innovative transformation technologies.

A future smart city in the population urban in India is dramatically increasing. Around 50% of the population is expected to be in cities by 2030. This has a significant challenge for the government because of the wasted resources and the weak infrastructure. Based on global studies in [7], one in eight people lives in slums, while in India, one in six people live in slums where they lack access to safe drinking water, food, and homes. The Indian government aims to build 100 smart cities to provide essential services for inhabitants.

Moreover, many cities worldwide cannot supply the city's infrastructure with artificial intelligence and internet of things technologies. The cities aim to be innovative by providing essential services to the people depending on the available resources.

III. Material and Methods

Building smart and sustainable cities in underdeveloped countries require significant effort to extract and set a strategic plan for efficient use of available resources so that these resources can be used to develop intelligent systems that provide the essential services for people.

The sustainable resources are different in each country. For example, Africa holds 65 percent of the world's arable land, and over 70 percent of people depend on forests and woodlands. Thus Africa can use the wood with the help of solar energy as a renewable resource to build a sustainable and intelligent system. Examples of countries in Africa are Egypt and Sudan, located geographically on the AL-Nile River, which is the longest river in Africa. It could be an efficient resource to transform the cities to be smart.

Several countries in Africa lack electricity and internet services, and electricity is only provided for essential life services. Thus those countries cannot provide electrical energy to build intelligent systems or implement Internet of things technologies in the cities, as using these technologies requires a considerable budget. These countries can benefit from utilizing their different available resources to provide power for intelligent systems.

Let's set one of Egypt's cities as an intelligent city target. Egypt's population is over 100 million in 2020, where this considerable population percentage indicates the availability of various resources, a high ratio of traffic jams and accidents exists. Taking these indications, Egypt can utilize its resources to build a smart city efficiently.

For example, Egypt can utilize the water in the Al-Neil River to generate electrical energy. The requirements are a hydraulic turbine to convert flowing water into mechanical energy and a hydroelectric generator to convert the mechanical energy into electrical energy. Still, the pressure on the available water in the Al-Neil river is severed, that's why we need to utilize more resources to produce electricity. Now the question is, how can Egyptian cities become smarter with the help of the population?. The human body contains a massive quantity of energy. This energy works as a fuel for our everyday activities, but how can we generate electrical energy from the human body's movement?. The movement of human bodies generates kinetic energy that can be converted to power that can be later used to power electrical devices. The conversion of energy can be achieved via using piezoelectric sensors, electromagnetic, and electrostatic effects [8], thus adding these sensors in different ways to crowded places like streets, airports, malls, gyms, hospitals, schools, etc., can help in producing a sufficient amount of electricity.

Moreover, solar energy cannot be ignored where it is proofed that the sun is a powerful, sustainable resource to generate power.

Integrating Internet-of-things and artificial intelligence technologies in cities means that the city infrastructure, including many different objects, will be connected. Artificial and machine learning techniques, including advanced sensors and controllers, are utilized in smart cities to create smart and connected buildings, autonomous vehicles, connected vehicles, smart education, smart health framework, and smart industries [9].

IV. Proposed Framework for Smart and Sustainable City

A. Smart and Sustainable Transportation System

The intelligent transportation system in smart cities focuses on autonomous vehicles and connected vehicles that are extensive research. Many big companies built autonomous cars with different features and implemented advanced technologies. Autonomous vehicles have many significant features, such as mobility for disabled and older people to drive safely for long distances. Also, autonomous vehicles are charged electrically, which makes them eco-friendly. Accordingly, implementing an innovative, sustainable, and connected transportation system requires strong internet connections and sufficient electrical energy to be charged. Thus, sustainable resources play a vital role in converting the transportation system to a smart system. But here, two questions arise, how to generate electricity to charge the cars and power the internet connectivity? And how to implement an intelligent transportation system in Egypt?

There are many ways to charge cars and build a connected transportation system. Let us assume that all the vehicles in Egypt, whether public or private cars, taxis, or buses, are transformed into smart autonomous and electrical vehicles. They can be charged by adding a solar panel above the vehicle connected to the battery. It can help in reducing the need to set it in an electrical charging station. Still, we need to provide different options for charging, for example, adding a piezoelectric sensor which is a sensor that converts the pressure of the cars on the street to electrical energy. This electrical energy can be consumed in recharging the electric vehicles, turning street lights on or supplying the internet service provider to strengthen internet connectivity. Additionally, the connectivity between vehicles that are usually implemented via the internet connection, the connection method could be changed by making the vehicles that are close to each other for a specific distance connected, and they can communicate with each other via Bluetooth connectivity. The connected vehicles can communicate in case of an emergency situation, car accidents, bad weather conditions, or any different unexpected situation. In addition, sharing between vehicles can help exchange information about the other vehicle positions, speeds, routes, stopping, and decided to change lanes. It can be implemented using advanced machine learning algorithms and the internet of things technologies. Still, to ensure the efficiency of the connectivity, the communication also could be implemented through a mobile application installed in each driver phone; in this case, the energy consumption to provide robust internet connectivity is reduced [10].

B. Smart Connected and Sustainable Infrastructure

Smart and connected buildings are a new component introduced in smart cities where smart buildings contain many advanced and various embedded devices for proper control. Smart building concept means providing a safe and comfortable environment for people. Thus smart buildings will have different components and objects that maintain the comfort level for people who require less time and effort, including efficient and interconnected heat, ventilation, and cooling systems for individual floors in the buildings or other areas in the building and the smart metering of electricity, gas, and water, occupancy monitoring systems and hybrid vehicle charging technology.

Moreover, automation and wireless technology are essential in smart buildings. People communicate with the building’s components like doors, windows, lights, machines via different control options like mobile application, voice, and radio frequency identification, which can help in reducing the percentage of diseases that could be transmitted by touch.

A question arises here, how to supply these buildings with electricity and internet connectivity consistently?

According to previous research [11], global energy consumption in commercial and residential buildings has steadily increased between 20% and 40% in developed countries. Moreover, as the population grows, the higher the percentage of power consumption, governments, especially in undeveloped countries, should follow different strategies to provide electrical energy for buildings.

There are many different ways for generating consistently electrical energy in buildings, such as adding solar panels on the roof of the facility connected with batteries to be used later. In addition, depending on the kinetic energy generated from human body movement to be converted to electrical energy could be a good solution. Generating electrical energy from a human could be implemented in several ways. For example, they are adding a piezoelectric sensor in the shoes, where shoes now can convert the kinetic energy into electrical energy stored in a portable battery to be used later in charging the smart devices like phones or watches. Accordingly, this will reduce the power consumptions used in charging smart devices. Furthermore, a piezoelectric sensor could be added to mats in the entrance and the exit buildings where this generated energy could be used later in controlling the intelligent doors, windows, lights, and so on.

In addition, some buildings can specify space for people to do their sports. This space can be utilized sufficiently if a piezoelectric sensor provided the specific space to produce electricity while people are running or using doing their daily workouts. Also, there is an additional energy source, where a generator can be connected to bicycles to gain benefits from the mechanical energy generated by humans to be converted to electrical energy that can be used in the smart buildings [12]. Moreover, a simple, sustainable, intelligent system could be implemented in Egypt, connecting solar panels and light sensors to the streets' light. Thus the light will be charged from the solar panels, and it will be turned on and off based on the sunlight availability so that the lights will be off during the morning and afternoon time, and it will be on at the evening time.

All these options can be implemented to power the intelligent systems that will control the essential services in the building, whether the building is a mall, gym, park, school, hospital, or home.

C. Smart and Sustainable Health Framework

Smart hospitals are applying extensive use of new intelligent systems and technologies to improve healthcare quality at less cost.

Concentrating on and implementing new technologies like artificial intelligence, robotics, 3d printing, augmented and virtual reality, and telemedicine plays a vital role in immediate requirements like reducing cost and high efficiency and long-term goals like greater precisions, fewer errors, and better outcomes.

The healthcare framework is changing due to the increasing number of today's patients because of vital viruses who require healthcare services.

The change in healthcare framework to be smart includes embedding many new technologies into hospitals' design and operations to improve patients' experience, embedding new technologies to improve healthcare, and create an interconnected system between all hospitals. A question arises here: how to integrate new technologies in Egypt where electricity is not consistently available, thus using renewable and sustainable resources and green practices could be challenging.

Several suggestions that could convert the hospital nowadays in Egypt to smart and sustainable hospitals such as hospitals can implement the same intelligent system in smart homes: the HVAC system and patient record to control room's temperature. When the room in the hospital is empty, the system is adjusted to the minimum ventilation settings. When the patient is in the room, he can control the temperature via remote control or voice. Although hospitals' HVAC systems help in electricity consumption reduction, at the same time, this system helps in reducing the injuries percentage by patients who are in need to leave the bed to adjust the room temperature.

Moreover, as the hospitals and medical centers are usually crowded places, thus converting the kinetic energy of people who are entering the hospitals to electrical energy could help in reducing the electrical energy consumption in the hospitals. Moreover, hospital buildings can include the same feature as smart buildings, solar panels, fuel cells, and an underground water system.

But what about the intelligent healthcare system?. In addition to the use of generated electricity from different resources to power the smart devices used in intelligent systems, a part of the healthcare system can be converted to a virtual system and gain benefit from telemedicine technology, where there will be an artificial intelligence virtual health assistant persist in a mobile application that can talk and respond to people questions and situations, if it is a simple health problem that the virtual assistant could solve, then the problem could be easily solved. Otherwise, a human assistant will be presented to aid in suggesting the patient's problem. Furthermore, in medicine, artificial intelligence-based analytics for supporting the decision-making process is required, that is why an example of an artificially intelligent algorithm recently developed to rapidly detect Covid-19 by combining patient's chest scan with clinical information like age, blood report and contact information, the system can help in evaluating the infected patients [13]. Moreover, intelligent systems can assist in work management like scheduling and planning clinical staff working time and allowing the clerk desk laborers to reduce their intensive tasks by replacing their work with automation and artificial intelligence algorithms.

D. Smart Education System

Intelligent systems and the internet of things can provide significant communication between virtual and physical objects, thus implementing artificial intelligence and internet of things technologies in the education system plays an effective role in converting the traditional education systems into a smart system for both teaching and learning process.

Transforming the education system means integrating new technologies like artificial intelligence, simulation, and virtual and augmented reality. A smart education system enables customized learning, which means students get a learning plan based on their skills, strengths, and interests to upgrade students' learning capacity for knowledge. Smart learning system includes many software and hardware tool like online resources, analytical tools, smart devices, interactive whiteboard, e-books, and e-bag.

Additionally, a smart education framework means converting the learning environment to be smart, including classrooms, schools, universities, etc.

The author in [14] explains some applied examples of smart learning, such as smart classrooms, including electronic and smart devices like internet connectivity, smart whiteboard, tablets, e-books, and a projector. While smart laboratories should have special smart devices like virtual reality glasses and computers to simulate different subject topics and requirements, for example, simulating the global wars in history subject could be in a very interesting way in virtual reality simulator, or simulating human heart with all its parts using augmented reality. Also, in university, virtual reality and simulation can positively affect university students by converting all the theory parts to either simulated or practically applied in labs which assist in a better understanding of their major. Using these technologies in schools, students will be more satisfied with understanding lectures, doing presentations, and conversing with others. At the same time, teachers will be more confident and comfortable while giving the lessons and explaining the new concepts and ideas. Moreover, some applications transform the education system to smart such as:

- Smart surveillance system: this system could be applied in schools or universities to collect information from daily classes by using a camera, microphone and artificial intelligence algorithm to extract the information from the retrieved records from cameras to be then stored in the cloud for future use and for absents students to attend the classes through his/her smart device.
- Smart attendance system: collecting the attendance information in school or university requires time, especially in big classrooms or universities, that is why a smart attendance was developed that work based on RFID (Radio Frequency Identification) system where an RFID reader will be in each classroom and an RFID card with each student, thus the students need to pass their cards on the RFID reader to be recorded as attending. Also, the attendance system can be implemented through NFC (Near Field Communication), where NFC will be embedded in each student's phone, and an NFC scanner is placed in each classroom. Thus the attendance is recorded by scanning the NFC in each phone to store the information later in a specific server.

Implementing these systems requires a consistent amount of electricity to power the systems. The school building can be provided with the same sustainable techniques and ways to use smart buildings like solar panels and human energy harvesting. Converting students' energy to electricity would generate a good amount of energy.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll-down window on the MS Word Formatting toolbar's left.

E. Smart Industry

The industry sector in smart cities had an incredible revolution that made changes in manufacturing. Many new concepts have been introduced, such as industry 4.0, which is the outcome of the shift of digitalization and automation in the manufactory sector. These concepts enabled self-optimization, cognition, and customization in the industry field, where communication with machines and computers is possible. It is implemented through the internet of things, machine learning, autonomous robots, the internet of services, big data management, and cloud computing.

The internet of things is an essential component in the industry field in smart cities. It enables systems connectivity, communication between different objects and machines and can be accomplished by controlling other parts in the industry like controlling temperature, light, devices, and remote control. In contrast, the internet of service (IoS) is presented as the "service vendors" which provide digital services through the internet according to the business type.

Moreover, big data management and analysis play a vital role in the revolution of the industry sector, where a vast amount of data and information are produced daily in the industry. Big data management and analysis are used through utilizing new technologies to make analyses that can be further used in the development process of the produced products.

Furthermore, robotics automation is the technology that has dramatically changed the world, where robots, machines, and automation reduced the employers' number, effort, and time with high proficiency and accuracy.

On the other hand, developing the industry field in Egypt with implementing these technologies requires a good amount of energy consistently. It is something that is not easy to be provided. That is why a new adaptive way of generating electricity is required, as solar panels or human energy harvesting is not enough to power on giant machines during the manufacturing process!

Commonly, in countries worldwide, the manufacturers are built far away from the residential areas. Thus in Egypt, the populations are the most near to Al-Nile river where the government is producing the majority of electrical power from the water of Al-Neil, but how the manufactories will be provided with a sustainable resource of electricity far away from the Al-Nile river?

As the significant area of Egypt is desert, a good amount of electricity could be produced from the sand. According to [15], and is used to generate and store energy in Italy, where the plant is supposed to be as a concentrated solar panel that focuses sunlight on heating sound. The hot sand produces around 500 kilowatts of energy with one turbine added which is enough to power a small village, and more turbines can be added to increase the power, in this case, a reduction is made in the unsustainable electricity consumption, where the city can develop the smart industry sector without huge consumption of electricity.

V. Results and Discussions

After introducing the five essential parts of smart cities and the way, they are developed in underdeveloped countries, and create a relationship between developing these parts and sustainability. The author in [16] explains the smart city architecture as illustrated in Fig. 2.

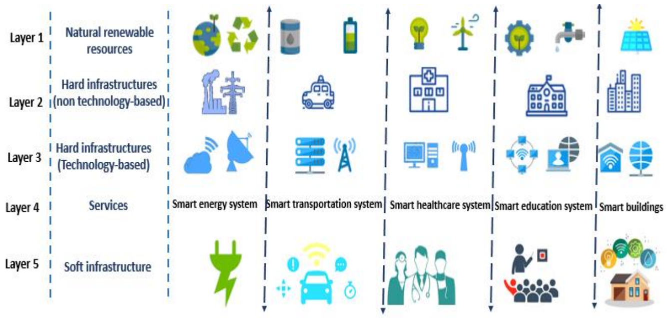


Fig. 2. The Smart City Architecture.

The smart city architecture includes four different layers:

Layer 1: Natural resources in the environment include the renewable resources available in the city, such as water, sunlight, fuel, and wind.

Layer 2: Physical infrastructures and services that are existed in the city. These physical infrastructures do not depend entirely on technology in their daily operations, such as transportation modes like cars, trains and buses or buildings, schools, offices, hospitals, etc.

Layer 3: Physical infrastructures and services that are technology-based and depend on technology to perform daily operations, such as communication networks, sensors, security cameras, etc.

Layer 4: The new smart services and applications developed based on the latest and advanced technologies and are used daily, such as smart transportation systems, smart health systems, smart energy systems, etc.

Layer 5: Soft infrastructures are individuals and communities living in urban areas such as stakeholders, companies, and other people utilizing smart applications and services.

Developing smart cities and implementing the smart city architecture with smart applications and services require performing a cost-benefit analysis (CBA) that is used to estimate the value of the cost and the benefits of developing and implementing a project. We can finalize whether the project is worthwhile or not.

Cost-benefit analysis is a systematic quantitative method of assessing the desirability of new projects or policies that are important for the long term and might have future effects on society.

The cost-benefit analysis is essential for stakeholders as it provides a comprehensive overview of the cost and benefit of a project. The cost-benefit research includes the following:

A. *Project Lifetime*

The project lifetime varies for different projects and depends on whether the project is private or public. The project's lifetime usually changes if the project is continually improved and modified, such as cars, buildings, highways, etc.

When the cost-benefit analysis is applied to investments in transportation systems, the project lifetime assumptions should consider that the transportation system project should have an infinite lifetime [17].

There are some examples shown in Table I of public projects' lifetime.

As illustrated in Table I, the energy and environment sectors have more life than the industry sector projects. The energy and environment sectors depend on renewable resources; thus, it can last for a long time, but for the industries that rely on systems or machines like telecommunication and industry sectors, the improvements for the project can aid in increasing the project lifetime limited period.

Thus, we need to consider the lifetime for different sectors for implementing a smart transportation system as the transportation system is related to the seven sectors mentioned in Table I.

B. *Cost-Benefit Analysis Process*

The cost-benefit analysis concept is used to evaluate the efficiency of the proposed solutions within the smart cities. The CBA is a general platform that constructs many questions and answers to provide the basic information for the construction and formulation of smart solutions. Fig. 3 illustrates the primary analytical framework of a simple CBA model designed to determine the efficiency of the solutions.

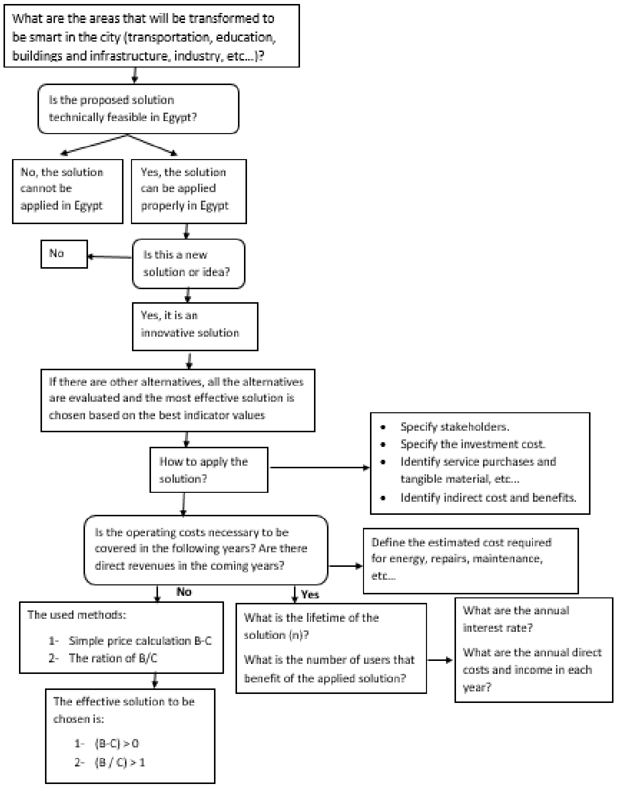

What are the areas that will be transformed to be smart in the city (transportation, education, buildings and infrastructure, industry, etc…)?
Is the proposed solution technically feasible in Egypt?
No, the solution cannot be applied in Egypt
Yes, the solution can be applied properly in Egypt
No
Is this a new solution or idea?
Yes, it is an innovative solution
If there are other alternatives, all the alternatives are evaluated and the most effective solution is chosen based on the best indicator values
How to apply the solution?
• Specify stakeholders.
• Specify the investment cost.
• Identify service purchases and tangible material, etc…
• Identify indirect cost and benefits.
Is the operating costs necessary to be covered in the following years? Are there direct revenues in the coming years?
Define the estimated cost required for energy, repairs, maintenance, etc…
No
Yes
The used methods:
1- Simple price calculation B-C
2- The ration of B/C
What is the lifetime of the solution (n)?
What is the number of users that benefit of the applied solution?
What are the annual interest rate?
What are the annual direct costs and income in each year?
The effective solution to be chosen is:
1- (B-C) > 0
2- (B / C) > 1

Fig. 3. The Primary Analytical Framework of Cost-benefit Analysis for a Smart Solution in a Smart City.

TABLE I. THE PROJECT LIFETIME FOR DIFFERENT INFRASTRUCTURE SECTORS

Infrastructure sector	**Project lifetime (years)**
Energy	25
Environment	30
Railways	30
Roads	25
Ports and airports	25
Telecommunications	15
Industry	10

Using the primary analysis of the CBA, several questions arise such as:

1) What are the areas that the smart solutions will be applied in?

2) Is the solutions technical could be applied in Egypt?

3) Are there any other alternatives for the solution to be considered?

4) What are the possible ways of implementing the solution? Specify all the input requirements, costs, and stakeholders.

5) Will it be necessary to cover the operating costs in the upcoming years? And what are the direct benefits in the coming years? If yes, what are the lifetime (years) of the solution and the number of users that the services will concern?

6) What are the annual interest rate and the annual indirect, direct cost yearly?

For example, let us suppose that a smart transportation system will be applied in a smart city in Egypt; what are the cost and benefits of the project, and how to evaluate the efficiency of the project?

According to [18], Egypt is planning to build smart cities on the cusp of the nation's sustainable development strategy in 2030.

Suppose we suppose that a smart city will be placed in Great Cairo city, the largest urban city in Africa, the Middle East, and the Arab world. A plan has been set to develop the urban in the city where there is a study performed to illustrate the major indicators of the Greater Cairo Region in 2022, shown in Table II.

TABLE II. MAJOR INDICATORS OF GREATER CAIRO REGION IN 2022

Socio-Economic Indicators	**2022**
Population (Million)	20.7
Motorization (Million Vehicles)	2.50
Trip Generation (Million Trips)	25.1

The development plan includes:

- Improvement in the road network and urban expressway network.
- For the public transport: extending the metro lines, satellite cities corridors, super tram system, network improvements, and optimized route structure.

To evaluate the development plan, the variables in implementing the project should be specified: the cost and the benefits of a smart transportation system.

The cost of the smart transportation system includes the following:

1) Capital cost: the capital cost means the initial cost of planning, designing, and implementing a project. The capital cost includes the design cost for roads, cars and systems, construction cost, electrical and electronic engineering cost, programmer cost, labor cost, and equipment purchasing cost in the smart transportation system.

2) Maintenance cost: Maintaining the technical parts for proper working conditions is required.

3) Operating cost: it is the cost of the advertisements, rent payments, and license fees needed annually.

4) Other costs: in addition to the operation and maintenance costs, more costs are to be considered. Specifically, there is additional cost related to smart solutions, technologies, cybersecurity, control and intelligent systems, and power generation systems.

That means the total cost of a smart transportation system in Egypt = $C_{capital} + C_{maintainance} + C_{other}$

The benefit is the revenue of the investments in a project, many benefits come from the smart transportation system in a smart city such as:

Travel time and cost reduction: The goal of any project is to gain the maximum benefit and pay less cost and efforts, thus the smart transportation system help in reducing the cost of traveling by lowering the payment amount for fuel consumptions, at the same time, the time of travel is reduced due to the connected cars' systems that are designed to reduce the traffic congestion time and reduce the traffic accidents as much as possible. According to the cost-benefit analysis, the financial value of the travel time reduction for different transportation modes can be expressed as:

$$V_{Time} = \sum_i (\Delta t_i \ x \ w_{time\ i}) \qquad (1)$$

Where Δt_i is the reduction in the travel time of each mode i.

Fuel and energy consumption: The primary challenge of developing a smart city in Egypt is providing the required energy to power the proposed intelligent systems. Thus we want to generate the maximum amount of renewable energy and reduce the consumed amount of fuel which is something difficult to be implemented especially for the transportation systems that depends on fuel for its daily operation, but converting the transportation system in Egypt to be sustainable system and replacing the mechanical cars with smart electrical cars means the fuel consumptions should be reduced, especially if a renewable resources generate the electrical energy.

For the smart transportation systems, the annual cost of the reduction in energy and fuel consumption is illustrated in this equation:

$$V_{Fuel} = \Delta G_{Fuel} \ x \ w_{Fuel} \qquad (2)$$

Where ΔG_{Fuel} is the amount reduced of fuel every year and the w_{Fuel} is the average price of fuel.

Reduce the gas emission: According to [19] the annual percentage of the energy-related carbon emission will be 10% higher than the percentage of emissions in 2014, but with the advancement in hybrid and electrical cars, the percentage will be decreased in 2030, as there are more concerns about the energy efficiency spreads and carbon reduction policies.

The author in [20] evaluates the development plan in the transportation system in Cairo by comparing cost-benefit analysis of the transportation system in 2001, committed project 2022, and the optimum result in 2022. It has been found that the development plan will help in recovering the average trip speed to around 18 kilometers per hour with the increase in the transport demand in 2022, with road congestion around 1.0. Moreover, the passenger of public transport in 2022 will be around 20.3 million per day compared to 18.2 million per day in 2001. In comparison, the carbon emission will be reduced by 15% compared to the percentage of emission in 2001.

The development plan and the CBA in Table III in Cairo transportation system does not cover the intelligent systems and the internet of things techniques and the smart energy generator systems, also the CBA ignores the future prices, population, requirements and improvements.

Where the smart and intelligent transportation system determines to install advanced and modern cameras, sensors, solar panels and internet connectivity.

Table IV shows the actual price of sensors and cameras used in smart cars and smart transportation systems collected from different online shops.

To reduce the cost of implementing an intelligent transportation system in Egypt, we must consider that not all the citizens can replace their car with a new smart electrical car which cost around 10,000 USD to 20,000 USD for small cars such as TESLA cars according to their prices online, but if the government announces clearly that, all citizens who live or work in new Cairo smart city must drive an electrical smart car, the citizens will either buy a new smart electrical car or transform their cars by installing an electric motor which cost around 6000 USD and adding large batteries about 1000 USD – 3000 USD, then installing sensors and cameras are required to be a smart car as well which costs around 1000 USD to 1500 USD based on the prices in Table III, thus the cost of transforming the cars or buying a new smart electrical car will cost almost the same.

Thus, based on the number of motorization in new Cairo in Table II, and based on the cost of transforming or buying new smart electrical cars, the cost of the system will be around 26 billion dollars, which is the capital cost for an intelligent transportation system in the high population new Cairo city. Also, the development of the city cost will be added to the cost of the intelligent transportation system according to Table III is 59.8 billion.

Furthermore, the cost of renewable energy resources which will be used in the cars, roads and in generating electricity from water or wind to be used later in charging the electrical cars such as the solar panels, wind turbine generator, Hydro generators and piezo electric sensor are shown in Table V.

According to Tables III, IV and V, the capital cost of developing smart sustainable transportation system in new Cairo will exceed 100 billion USD. Comparing the capital cost of the intelligent transportation system in Egypt 2022 with the capital cost for a smart transportation system in United States with lower population in Newark city 2015 [21] which is a developed city, where the capital cost was 100 billion without depending on the sustainable and smart energy systems, thus we can figure out that, the result of the capital cost calculation in Egypt is reliable.

TABLE III. THE COST ANALYSIS FOR THE DEVELOPMENT PLAN OF GREATER CAIRO

Scenario	**The transportation system in 2001**	**Committed transportation system project 2022**	**The optimum scenario of the developed transportation system**
Cost (LE billion)	-	18.2	59.8
Benefit-Cost Ratio	-	-	1.77
Trip speed (Km/hr)	19.0	11.6	18.0
Sharing percentage of public transport	70.9	61.7	57.9
Number of public transport passengers (million / day)	13.3	18.2	20.3
Vehicle-kilometer (million pcu-km/day)	62.8	127.3	139.7
Volume /Capacity Ratio	0.74	1.11	1.00
Population within 800m of Public transport (million)	2.04	3.09	8.20
Employment within 800m of public transport (million)	1.11	1.70	4.20
Students within 800m of public transport (million)	0.74	1.08	2.70
Carbon emission (million ton)	12.2	15.9	13.6

TABLE IV. THE ACTUAL PRICE OF SENSORS AND CAMERAS USED IN SMART CARS AND SMART TRANSPORTATION SYSTEMS

Sensors	**Actual price**
Ultrasonic sensor x4	16 USD to 40 USD
Automotive mono-camera	100 USD to 1000 USD

TABLE V. PRICES OF RENEWABLE POWER GENERATORS

Power generators	**Actual price**
Solar panels	200 USD to 300 USD
Hydroelectric generator	3,000 USD to 80,000 USD
Wind turbine generator	2,000 USD to 4,000 USD
Piezo electric sensor	40 USD to 60 USD

VI. Challenges for Smart and Sustainable Cities

It is widely known that to implement smart city projects, we have to combine information technology with internet-connected technologies. This case is especially true for underdeveloped countries that this paper focuses on. Smart city projects combine several technologies and applications that help improve management like waste management and smart grids [22]. Suppose smart cities became the 'modern standard'. In that case, decision-makers might need to navigate through different fields like connectivity, infrastructure, data, and security as cost and development are challenging. With technology and standards continually changing and quickly evolving, municipalities will need to avoid being trapped in using integrated solutions from a single provider, which ultimately leads to information and resources silos, making it even harder to exchange data with other municipalities. Thus, understanding that building a smart city goes beyond automating internal systems and introducing new applications. Implementing the concept of smart and sustainable cities in a country comes with various obstacles, including specific city criteria, variation of technologies, connecting infrastructures and conserving resources.

A. *Infrastructure Modernization*

The term "city infrastructure" includes housing, sanitation, water, and sewer supply, waste management, transport, energy supply, and distribution. The difference between smart and conventional infrastructure is the ability of smart infrastructure to adapt intelligently to changes in the environment without disregarding user needs to achieve improved performance. Smart city infrastructure provides the basis for all smart city themes, including transportation, economy, government, environment, and citizens. The smart infrastructure design depends on how modernized or developed a country is.

For a city in an underdeveloped country, developing a smart infrastructure will require considering cost, space and other factors to promote better management and development of resources. For example, to create smart transportation infrastructure, a massive amount of data must be gathered to redesign the transportation network and build new applications. Thus, requiring a large amount of capital from a country that is moving to develop and could not afford high costs. These countries can maintain the city's current infrastructure and focus on optimizing existing infrastructure resources and monitor performance. For example, establishing renewable energy infrastructure to the existing one to manage natural resources like water and utilize wind and solar energy to improve waste management. However, in both developing and developed countries, the primary reason behind smart infrastructure applications is adapting to society's demands for sustainable growth. Table VI summarizes some of the sustainable challenges and smart infrastructure solutions proposed for modernization.

B. *Security and Privacy*

Data is a core component of Smart Cities. Sensor networks, smart meters, cell phones and IoT devices all produce massive quantities of data. These generated data could be converted by the city government to create new insights that could be monetized and shared with various stakeholders. Smart urban infrastructure components pose risks and safety issues, as smart infrastructure may be vulnerable to hacking and unauthorized access. The problem of protecting people's privacy is also a significant concern in this regard. Many of the widely mentioned risks include stealing users' data and implementing ransomware. Of course, the hazard level grows as IoT systems become more controllable and more autonomous. In these latter situations, cybercriminals can exploit vulnerabilities to remotely control IoT devices to alter sensor or system actions, sabotage these devices, or even cause physical damage to the surrounding environment [23]. Such activities result in cyber hackers having more access to confidential data that expose customer behavior trends. We might also see more mobile device robberies, as more can now have physical access to homes and workplaces. As IoT ecosystems are built and often include several stakeholders and vendors, and customers, it is necessary to be aware of possible compliance problems if either confidential data is hacked or these sensors and devices are themselves regulated by cyber criminals [23]. Organizations and municipalities must identify and explicitly recognize each other's positions, obligations, and standards in terms of data protection and even in the event of data breaches within IoT ecosystems. To expand current cybersecurity strategies, a risk control policy should be developed so that both organizations in the public sector obtaining these systems and in the private sector developing them will need to be take into consideration the ramification of developing and obtaining these technologies and the data protection and AI threats that will need to be resolved by effective governance steps [24].

TABLE VI. Challenges with Infrastructure Modernization

Sustainable Challenges	**Proposed Smart Infrastructure**	**Description**
Improving Energy infrastructure	Smart Meter	Monitor behavior regarding electric energy and measure consumption
	Smart Grids	Manage variable renewable power supply
Affordable, good quality connectivity	High-Speed Internet	Cellular and Low Power Wide Area (LPWA) technologies
Environmental Performance	Environment Sensor Network	Collect data regarding environmental condition and level of pollutants
Health and Education	Telemedicine, Remote healthcare, and Online Education	Services and products to access education and health services remotely

C. *Internet Connectivity*

Introducing smart city technologies requires a robust and efficient broadband network. This shows the need to continue bridging the digital gap to utilize and take advantage of smart applications (apps). Mobile broadband networks play a key role in a shortage of fixed networks. Smartphones revolutionized smart cities as several "apps" were developed to ease users in handling several facilities like road and transportations, health and fitness, and electricity and water.

Due to increased sensor and data used, robust and high connectivity is needed to maintain and accommodate it. This is where 5G shines and provides the essential structure, capability and service that provide cases underpinned by technological features like high bandwidth, absolute reliability, wide availability, and responsive connectivity. Such utilization of technology is limited to a city's budget, which is the case with underdeveloped countries [25].

D. Funding and Financing

Conceptualizing smart cities provides a digital layer on any capital asset, opening the way to unparalleled possibilities, technological improvements, and increased profitability. One significant factor that can make a smart city design productive is the consistent path to steady revenue (its revenue model). A reliable funding source is particularly important if the project is pursuing private financing. When planning the project, it is important to be mindful of the difference between the terms "Financing" and "Funding". The government provides particular money for specific reasons with free charges and no interests in terms of funding. In comparison, the term financing refers to financial/private institutes that grant a project a certain amount of "funding" as equity to be returned later with interests. For underdeveloped countries in Africa, it is crucial to think about revenue and cash flow, how funds can be recovered from the overall revenue, the value created, and how to monetize it. That's why it is important to strategize a plan to secure funds and finances from different investors, as some might focus on infrastructures like roads and transportation while others might focus on different sectors and themes.

E. Energy Management

Cities that have high population densities in a country like Egypt generally require a huge power source. Once these cities are developed as smart cities, renewable energy resources will be integrated with the city's electrical infrastructure to support the economy and life-quality growth. Smart cities can help sustain energy and environmental challenges that produce a proportion of carbon emissions. It is generally known that cities play an increasingly decisive role in addressing climate and environmental challenges. Cities are naturally able to grow to face the world's many energy and environmental problems. Human and intellectual capital, along with economic and political influence, can fuel the increased use of renewable energy. The primary constraints that might be faced make these cities hubs for technical and social creativity to drive change globally. Energy and environmental considerations are key that drives (and limit) all facets of urban planning [26].

VII. Conclusion

This research claim that implementing the smart city model could provide an opportunity for underdeveloped countries to handle the effects of rapid urbanization, including their economic and environmental implications and the decrease in quality of life. Achievement of sustainability and enhancement of quality of life in underdeveloped cities varies from one country to the other. According to the findings, no standard criteria determine a city's smartness, and analysis must be performed before settling on smart projects and systems. This research analyzed the cost and benefits of making New Cairo, Egypt into a smart city. Also, this research discussed how utilizing the rich resources in Africa and investing in smart technologies makes the underdeveloped countries across the continent leap urban revolution. We finalized that in smart city projects, sustainability becomes interwoven with various objectives and priorities, all of which interfere and influence each other, generating potential feedback loops and unforeseeable results.

References

[1] B. O. Bureau, "Why we need smart cities," 19 February 2019.

[2] M. Deakin, "From intelligent to smart cities," Intelligent Buildings International, 2011.

[3] T. Yigitcanlar, "Smart cities: an effective urban development and management model?," Autralian planner, 2015.

[4] J. c.-L. a. J. M.-F. Marsal-Lluisa Marsal-Llacuna, "Lessons in urban monitoring taken from sustainable and livable cities to better address the smrt cities initiative," Technological Forecasting and Social Change, 2015.

[5] M. Roy, "planning for sustainable and urbanisation in fast growing cities: Mitigation and adaption issues addressed in Dhaka, Bangladesh," Habitat international, 2009.

[6] G. Match, "From Sustainable to Smart Cities," The Future of Modern Cities, 06 November 2018.

[7] S. Kumar, "Aimed at ending poverty, smart city project ends up hurting the poor itself," 26 December 2018.

[8] Q. H. J. W. Z. Z. &. X. L. Keli Li, "Wearable energy harvesters generating electricity from low-frequency human limb movement," Microsystems & Nanoengineering, 2018.

[9] J. M. G. G. H. N. M. A. Hossein Shahinzadeh, "IoT Architecture for Smart Grids," in International Conference on Protection and Automation of Power System (IPAPS), 2019.

[10] I. R. Pawl Gora, "Traffic models for self-driving connected cars," Elsevier, 2016.

[11] J. C. LuisPérez-Lombard, "A review on buildings energy consumption information," ScienceDirect, 2008.

[12] E. Cosgrove, "Can you really power your home with a bicycle generator?," 4 September 2020.

[13] D. R. H. V. Kumar Chebrolu, "Smart use of artificial intelligence in health care," 22 October 2020.

[14] D. Mohanty, "Smart learning Using IoT," International Research Journal of Engineering and Technology (IRJET), 06 June 2019.

[15] R. E. World, "Italian Company Uses Sun-Heated Sand to Produce Energy," 21 May 2015.

[16] Anthopoulos, "Understanding smart cities: A tool for smart government or an industrial Trick?," Springer, 2017.

[17] D. Lee Jr, "Fundamental of life-cycle analysis. Transportation research record," Journal of the transportation research board , 2002.

[18] A. G. o. b. o. U. N. I. C. –. Cairo, "Smart Cities: Egypt "Theory and Application – Towards an Egyptian Smart Cities Code"," 09 December 2019.

[19] ExxonMobil, "Innovating energy solutions: Research and development highlights," Energy and innovation , 2014.

[20] J. I. C. A. (JICA), "Public-private partnership for Cairo urban toll expressway network development," Katahire & Engineers International , 2006.

[21] X. Xiong, "Cost-Benefit analysis of smart cities," Xiangyuan Xiong, 2018.

[22] N. A. a. K. Hasley, "Smart cities face challenges and opportunities," 25 July 2018.

[23] N. D. Evans, "The IoT meets the Internet of Behaviors | CIO," 2014 June 2014.

[24] N. G. a. M. Sobiecki, "Artificial intelligence in smart cities," 10 August 2020.

[25] Telecoms, "the challenges of 5G and smart cities," 15 June 2020.

[26] P. L. Thiez, "Smart City: Energy Challenges Facing Sustainable Cities," 01 October 2018.

Monophonic Guitar Synthesizer via Mobile App

Edgar García Leyva[1]
Instituto Politécnico Nacional. SEPI-ESCOM
Mexico City, Mexico

Elena Fabiola Ruiz Ledesma[2]
Instituto Politécnico Nacional.
ESCOM, UPIICSA
Mexico City, Mexico

Rosaura Palma Orozco[3]
Lorena Chavarría Báez[4]
Instituto Politécnico Nacional.
ESCOM, Mexico City, Mexico

Abstract—**In the guild of guitarists, it is common to work with guitar synthesizers because the emulation of a great variety of sounds that are produced by different musical instruments, starting from just playing the guitar, which means, a piece of music is played with a guitar, but other musical instruments are actually heard such as, a saxophone, a violin, a piano or percussions, depending on the instrument that has been selected. The problem that arises in this article is that synthesizers are expensive and due to their size, the transportation of the equipment is often impractical. As mentioned, the development of a mobile application that has the function of a monophonic synthesizer is proposed as a solution. In this way, the cost is greatly reduced, and additionally, the user is able to install the application on a mobile device with Android operating system and connect it to an electric or electro-acoustic guitar through an audio interface; obtaining as a result, a functional technological instrument by offering guitarists an alternative with respect to conventional synthesizers. The construction of this application used the Fast Fourier Transform Radix-2 as a signal recognition algorithm, which allowed obtaining the fundamental frequencies generated by the guitar, which were transformed into MIDI notation and later used in sound emulation.**

Keywords—Monophonic synthesizer; guitar; sound emulation; mobile application

I. Introduction

The advent of tablets and smart cell phones has opened a range of possibilities in all areas of knowledge. In the field of music, technological development has occupied a privileged place by having mobile applications that help in tuning musical instruments, in measuring time to practice, in recording and editing music, among other aspects.

The rise of technology has allowed the creation of tools in order to support musicians in a vast diversity of ways. The tool that is interesting to highlight in this article is related to the emulation of monophonic sounds of musical instruments or other types of sounds, which is obtained through the use of synthesizers. The Royal Academy of the Spanish Language defines synthesizer as: "Electronic musical instrument capable of producing sounds of any frequency and intensity and combining them with harmonics, thus providing sounds of any known instrument, or sound effects that do not correspond to any conventional instrument" [1]. Synthesizers are very useful devices because when connected to the guitar, are able to provide a great variety of sounds, increasing the possibilities of musical interpretation of the guitarist.

Two of the problems that arise are its portability and its high cost; Due to the mentioned, based on computer science and mathematics, it was decided to develop a monophonic synthesizer, using a mobile application, which allows the guitarist to use a synthesizer through a mobile device such as a smart cell phone or a tablet, which can be transported easily, without having to make an additional expense.

Music can be classified into two main categories: monophonic and polyphonic. Monophonic music is made up of a single melodic line, which means, only one musical note sounds at a certain time, while polyphonic music is made up of more than one melodic line, which means, two or more musical notes sound at the same time [2]. The present study focuses on monophonic music.

The guitar is a musical instrument that allows its player to express musical notes in different ways. The guitar, like other instruments such as the piano or the violin, generates analog sounds, which need to be converted to digital sounds in order to be read by the computer. From this digitization, it is possible to apply signal recognition techniques for different purposes and in this case, to detect the fundamental frequencies generated by the guitar. A transformation was applied to these frequencies using the notation mentioned in the Musical Instrument Digital Interface (MIDI), to obtain a discrete representation of the musical notes, which are defined in an interval that goes from 0 to 127, where each number corresponds to a musical note.

The overall objective of this study was to develop a portable tool that allows the guitarist to emulate monophonic sounds of other musical instruments or other types of sounds using the guitar, all this through computational and mathematical techniques. For this purpose, the following specific objectives were proposed:

- Recognize frequencies generated by guitars making use of hardware and software resources of mobile devices.
- Transform frequencies generated by guitars to MIDI notation.
- Emulate monophonic sounds of musical instruments or other sounds based on the obtained MIDI notes.

This article is divided into 4 sections. The second section shows some synthesizers available on the market, as well as some application programming interfaces that have been developed in order to recognize signals and to support the execution of MIDI sounds. The third section deals with the theoretical references about the recognition of monophonic sounds and the transformation of frequency to MIDI notation. The fourth section shows the methodology used for the development of the monophonic synthesizer following the stages of the incremental Software Engineering model. Subsequently, the results obtained from the tests carried out are shown and finalized with the conclusions.

II. Related Work

At present, there are independent synthesizers for guitar, which means, mounted on electronic circuits, some of them are: MEL9 [3], SY-300 [4], GR-55 [5], among others; However, when it comes to mobile applications that perform the function of a guitar synthesizer on the Android operating system, there are no formal alternatives to it.

On the other hand, there are some application programming interfaces (APIs) aimed at the Android operating system, which can be useful for the construction of a synthesizer, for example, those that serve to perform frequency recognition and those that serve to execute MIDI sounds, some of these interfaces are: TarsosDSP [6] and MIDI Driver [7] respectively.

In the present mobile application, it was chosen the frequency recognition through the Fast Fourier Transform Radix-2 algorithm, according to [8], making use of the native Java development kit, while for the execution of MIDI sounds, the model that was done in [7] was retaken.

III. Theoretical Aspects

In this section, reference is made to the techniques used to recognize monophonic sounds and, on the other hand, the digital interface of musical instruments (MIDI) is presented, which shows a notation that serves to discretize frequencies.

A. *Monophonic Sound Recognition*

Due to in this article the monophonic sounds generated by the guitar are taken as a basis, some algorithms that can be used for the recognition of monophonic sounds are specifically mentioned.

There are mainly two approaches that are used to perform monophonic sound recognition. One of them consists in analyzing the signal samples in the time domain, and the other in analyzing them in the frequency domain. A widely used method in the time domain is autocorrelation, which compares a signal with delayed versions of itself at successive intervals to find the highest amplitudes within the signal and measure the distances between them. Through these distances the period of the wave can be inferred, and with it, the present frequency of the monophonic sound can be detected. On the other hand, there is the analysis of signals in the frequency domain, where algorithmic implementations of the Discrete Fourier Transform are used, with which a set of frequency intervals and their amplitudes are obtained. A simple way to detect the frequency of the monophonic sound present quickly is to select the frequency with the greatest amplitude [9].

In this article, the signals in the frequency domain are analyzed, so some algorithms that can be used to recognize the monophonic sounds generated by the guitar in that domain are specifically mentioned.

B. *Fast Fourier Transform (FFT)*

Fast Fourier Transform is an efficient mathematical implementation of the Discrete Fourier Transform (DFT), which is a particular case of the Fourier Transform for sequences of finite length in which the spectrum is evaluated only in a few specific frequencies, and therefore, a discrete spectrum is obtained [10, 11].

Over time, several FFT algorithms have been developed such as: prime factor, split radix, vector radix, split vector radix, Winograd Fourier transform, etc. [12].

The FFT algorithm used to develop the mobile application of the monophonic synthesizer is the Radix-2. The Radix-2 algorithm is considered the most used for the FFT calculation, it works when the number of data samples is a power of 2, in case the number of samples does not satisfy this criterion, the missing spaces are filled with value 0, this does not alter the calculated frequency spectrum. The input and output of an FFT are expressed in complex numbers, in this case, for its implementation, two arrangements are accepted to store the real and imaginary components in them, when using the recording tools of mobile devices, all the bytes of audio information are obtained, which are used within the real components, while the imaginary components are always filled with zeros, the output of the algorithm is contained in two other arrays, one corresponding to each type of component, where the frequency spectra are stored, since only the actual samples are used for input, only the first half of the components need to be analyzed. Each component of the frequency spectrum is related to the previous component, since it is the sampling frequency divided by the number of samples of the FFT [8].

In order to recognize monophonic sounds, the power or amplitude spectrum (Xp) is analyzed, selecting the frequency f with the greatest amplitude, which is calculated through the sum of the squares of its real (Xreal) and imaginary (Ximag) components, as shown in (1),

$$Xp = Xreal(f)^2 + Ximag(f)^2 \tag{1}$$

The FFT algorithm used to develop the mobile application is Radix-2, which means that it must work on a group of samples whose number is a power of two [8].

C. *MIDI*

Musical Instrument Digital Interface (MIDI) is a music notation system that allows computers to communicate with musical synthesizers. MIDI files contain instructions to create the pitch, volume and duration of the notes, this based on a sequence of events called: note_on and note_off [13].

Musical notes are not encoded by their names, instead numbers from 0 to 127 are assigned as shown in Table I. For example, the number 57 corresponds to a musical note A with a frequency of 220 Hertz (Hz).

Equation (2) shows the transformation of frequency in Hz to MIDI note,

$$MIDINote = round(69 + 12 \times \log_2 (f / 440)) \tag{2}$$

TABLE I. MIDI NOTES ASSOCIATED WITH ITS NAME AND FREQUENCY IN HZ [15]

Name	MIDI note	Frequency (Hz)
D	38	73.42
D♯/E ♭	39	77.78
E	40	82.41
F	41	87.31
F♯/G ♭	42	92.50
G	43	98.00
G♯/A ♭	44	103.83
A	45	110.00
A♯/B ♭	46	116.54
B	47	123.47
C	48	130.81
C♯/D ♭	49	138.59
D	50	146.83
D♯/E ♭	51	155.56
E	52	164.81
F	53	174.61
F♯/G ♭	54	185.00
G	55	196.00
G♯/A ♭	56	207.65
A	57	220.00
A♯/B ♭	58	233.08
B	59	246.94
C	60	261.63
C♯/D ♭	61	277.18
D	62	293.66
D♯/E ♭	63	311.13
E	64	329.63
F	65	349.23
F♯/G ♭	66	369.99
G	67	392.00
G♯/A ♭	68	415.30
A	69	440.00
A♯/B ♭	70	466.16
B	71	493.88
C	72	523.25
C♯/D ♭	73	554.37
D	74	587.33
D♯/E ♭	75	622.25
E	76	659.26
F	77	698.46
F♯/G ♭	78	739.99
G	79	783.99
G♯/A ♭	80	830.61
A	81	880.00
A♯/B ♭	82	932.33
B	83	987.77
C	84	1046.50
C♯/D ♭	85	1108.73
D	86	1174.66
D♯/E ♭	87	1244.51
E	88	1318.51
F	89	1396.91
F♯/G ♭	90	1479.98
G	91	1567.98
G♯/A ♭	92	1661.22
A	93	1760.00

Where round is a function of rounding to one digit, factor 12 is the resulting linear pitch space per octave, factor 69 is note A (440 Hz), which is taken as reference, $\log_2$ is used according to the logarithmic pitch perception in humans and the variable f is the input frequency that will be converted to a MIDI note [14].

General MIDI is a standardized specification for electronic musical instruments that respond to MIDI messages. General MIDI was developed by the American MIDI Manufacturers Association (MMA) and the Japan MIDI Standards Committee (JMSC) and first published in 1991 [16].

Within the general MIDI specification 128 sounds of musical instruments or other types of sounds are included, and these are divided in sections such as: Piano, Chromatic Percussion, Organ, Guitar, Bass, Strings, Ensemble, Brass, Reed, Pipe, Synth Lead, Synth Pad, Synth Effects, Ethnic, Percussive and Sound Effects which are used in this mobile application and are shown in Table II.

TABLE II. SOUNDS OF THE GENERAL MIDI SPECIFICATION [16]

00 - Acoustic Grand Piano	32 - Acoustic Bass	64 - Soprano Sax	96 - FX 1 (rain)
01 - Bright Acoustic Piano	33 - Electric Bass (finger)	65 - Alto Sax	97 - FX 2 (soundtrack)
02 - Electric Grand Piano	34 - Electric Bass (pick)	66 - Tenor Sax	98 - FX 3 (crystal)
03 - Honky-tonk Piano	35 - Fretless Bass	67 - Baritone Sax	99 - FX 4 (atmosphere)
04 - Electric Piano 1	36 - Slap Bass 1	68 – Oboe	100 - FX 5 (brightness)
05 - Electric Piano 2	37 - Slap Bass 2	69 - English Horn	101 - FX 6 (goblins)
06 - Harpsichord	38 - Synth Bass 1	70 – Bassoon	102 - FX 7 (echoes)
07 - Clavi	39 - Synth Bass 2	71 – Clarinet	103 - FX 8 (sci-fi)
08 - Celesta	40 - Violin	72 – Piccolo	104 - Sitar
09 - Glockenspiel	41 - Viola	73 – Flute	105 - Banjo
10 - Music Box	42 - Cello	74 – Recorder	106 - Shamisen
11 - Vibraphone	43 - Contrabass	75 - Pan Flute	107 - Koto
12 - Marimba	44 - Tremolo Strings	76 - Blown Bottle	108 - Kalimba
13 - Xylophone	45 - Pizzicato Strings	77 – Shakuhachi	109 - Bag pipe
14 - Tubular Bells	46 - Orchestral Harp	78 – Whistle	110 - Fiddle
15 - Dulcimer	47 - Timpani	79 – Ocarina	111 - Shanai
16 - Drawbar Organ	48 - String Ensemble 1	80 - Lead 1 (square)	112 - Tinkle Bell
17 - Percussive Organ	49 - String Ensemble 2	81 - Lead 2 (sawtooth)	113 - Agogô
18 - Rock Organ	50 - Synth Strings 1	82 - Lead 3 (calliope)	114 - Steel Drums
19 - Church Organ	51 - Synth Strings 2	83 - Lead 4 (chiff)	115 - Woodblock
20 - Reed Organ	52 - Choir Aahs	84 - Lead 5 (charang)	116 - Taiko Drum
21 - Accordion	53 - Voice Oohs	85 - Lead 6 (voice)	117 - Melodic Tom
22 - Harmonica	54 - Synth Voice	86 - Lead 7 (fifths)	118 - Synth Drum
23 - Tango Accordion	55 - Orchestra Hit	87 - Lead 8 (bass + lead)	119 - Reverse Cymbal
24 - Acoustic Guitar (nylon)	56 - Trumpet	88 - Pad 1 (new age)	120 - Guitar Fret Noise
25 - Acoustic Guitar (steel)	57 - Trombone	89 - Pad 2 (warm)	121 - Breath Noise
26 - Electric Guitar (jazz)	58 – Tuba	90 - Pad 3 (polysynth)	122 - Seashore
27 - Electric Guitar (clean)	59 - Muted Trumpet	91 - Pad 4 (choir)	123 - Bird Tweet
28 - Electric Guitar (muted)	60 - French Horn	92 - Pad 5 (bowed)	124 - Telephone Ring
29 - Overdriven Guitar	61 - Brass Section	93 - Pad 6 (metallic)	125 - Helicopter
30 - Distortion Guitar	62 - Synth Brass 1	94 - Pad 7 (halo)	126 - Applause
31 - Guitar harmonics	63 - Synth Brass 2	95 - Pad 8 (sweep)	127 - Gunshot

IV. METHODOLOGY

The mobile application was developed using the phases of the incremental Software Engineering model, which applies linear sequences in a staggered manner as the calendar of activities progresses. Each linear sequence produces deliverable software increments [17]. The diagram in Fig. 1 describes the stages used for this development.

According to what Mall [18] points out, first of all a simple system is built and delivered which implements only a few basic characteristics. During a few successive iterations, improved versions are deployed and delivered, until the desired system is finally realized.

The software requirements are first divided into several modules or features that can be built and delivered incrementally. This is graphically represented in Fig. 2.

Returning to what Pressman and Mall [17, 18] point out, 3 modules were created as part of the development of the mobile application.

In the first module, the recognition of the frequencies generated by the guitar was carried out, for which the Fast Fourier Transform Radix-2 algorithm was applied, making use of the hardware and software resources of mobile devices. When playing a musical note with the guitar, the fundamental frequency corresponding to said musical note was obtained in real time, showing it on the mobile application interface.

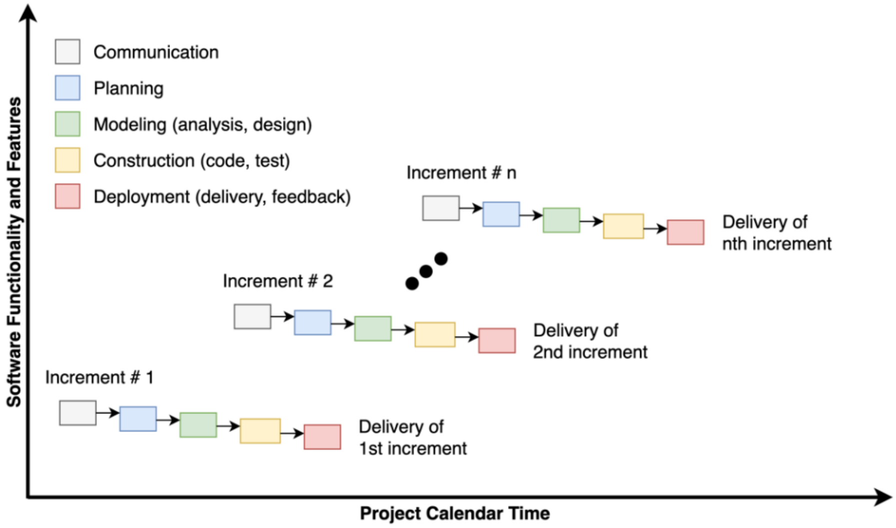

Communication
Planning
Modeling (analysis, design)
Construction (code, test)
Deployment (delivery, feedback)
Increment # n
Delivery of nth increment
Increment # 2
Delivery of 2nd increment
Increment # 1
Delivery of 1st increment
Software Functionality and Features
Project Calendar Time

Fig. 1. Diagram of the Incremental Model.

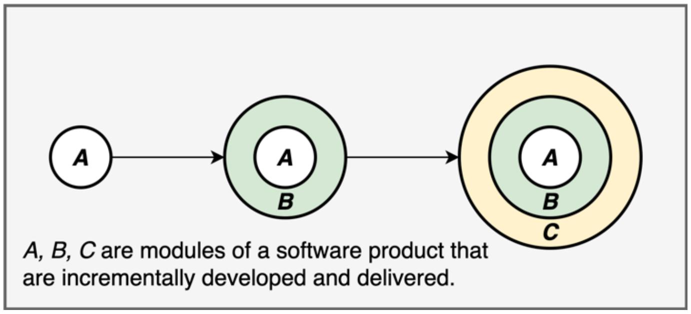

A
A
B
A
B
C
A, B, C are modules of a software product that are incrementally developed and delivered.

Fig. 2. Incremental Software Development.

In the second module, the transformation of frequencies to MIDI notes was carried out, for which equation 1 shown in the MIDI section was applied, obtaining as a result, discrete values of frequencies within a range from 0 to 127, showing the aforementioned transformation together with at the fundamental frequency in the mobile application interface.

The third module allowed to emulate the monophonic sounds of musical instruments or other types of sounds, based on the MIDI notes obtained. The note_on and note_off events of the digital interface of musical instruments were used to execute and stop the MIDI notes. Consequently, the moment the user plays a musical note with the guitar, the note_on event is activated, with which the sound chosen by the user within the 128 included in the mobile application must be heard; for example, a saxophone, a trumpet, a violin, among others; whereas when there is an absence of sound, the note_off event was activated to keep the mobile application silent.

For the construction of all these modules, the Java programming language was used together with the Android Development Kit (SDK).

A. *Logical Block Diagram of the Structure of the Mobile Application*

Fig. 3 shows the block diagram of the mobile application. As can be seen, an audio input is required, which goes through a sampling process to obtain the discretized signal, making use of the audio recording tools offered by the Android operating system. Subsequently, the filter allows to eliminate the peaks of the signal, which means, it eliminates the unwanted noise, and then to apply the Fast Fourier Transform and thus obtain the fundamental frequencies. From these fundamental frequencies, the transformation to MIDI notation is carried out. From this moment, the sound chosen by the user can be emulated through the execution of the note_on and note_off events.

B. *Operation of the Mobile Application*

Fig. 4 represents the operation of the mobile application. In order to use it, it is necessary to connect it to an electric or electroacoustic guitar through an audio interface, the last one is connected to an audio output device such as an amplifier.

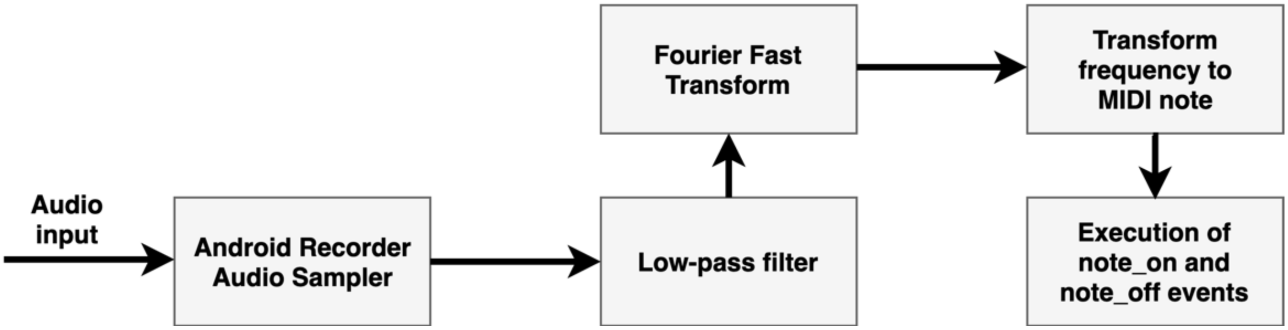

Audio input
Android Recorder Audio Sampler
Low-pass filter
Fourier Fast Transform
Transform frequency to MIDI note
Execution of note_on and note_off events

Fig. 3. Block Diagram of the Mobile Application.

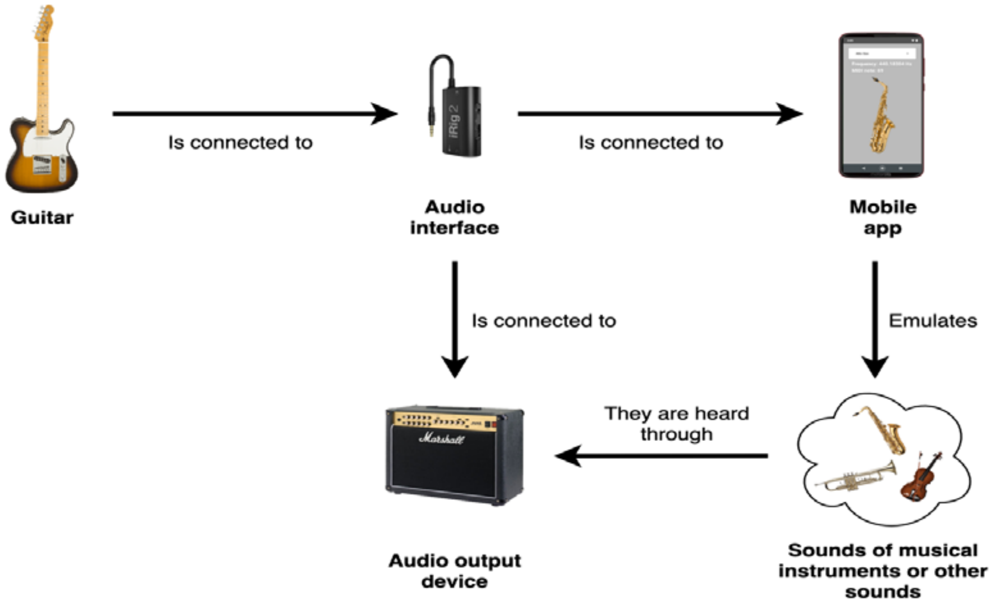


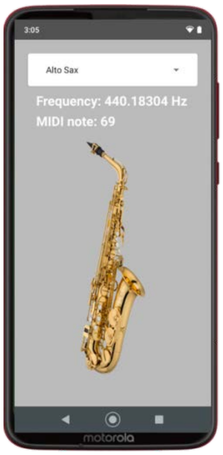
Fig. 4. Mobile Application Operation.

V. Results

The mobile application has been developed using the Java programming language in order to run on mobile devices with the Android operating system. The tests were carried out on a device from the Motorola brand, a Moto G7 Plus model with 64 GB of storage and 4 GB of RAM, with the Android 10 (Android Q) operating system installed.

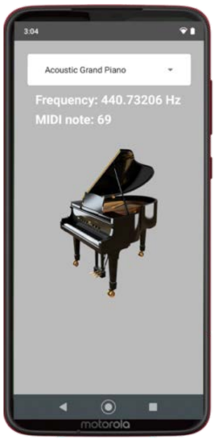
The work was carried out with a sample of three guitarists to carry out the tests of the operation of the mobile application. Two of the guitarists used an electric guitar and the remaining guitarist used an electro-acoustic guitar. The musical instruments were connected to the mobile application through an iRig 2 audio interface. The guitarists selected in the list of the main interface of the mobile application, several of the 128 sounds that were available to be emulated and they played the guitar obtaining the sound expected. Additionally, they were able to observe the frequency of the note played, as well as its transformation to a MIDI note in real time.

The mobile application was subjectively evaluated with the feedback obtained from the guitarists, who pointed out the great usefulness of this mobile application because it broadens their possibilities of interpretation with their musical instrument, without requiring more than their mobile device and an audio interface. On the other hand, they mentioned that they noticed a slight latency between the moment they played the guitar and the emulation of the chosen sound, which could be reduced using a low-level programming language. Two examples of the mobile application in operation are shown in Fig. 5 and Fig. 6.



Frequency: 440.18304 Hz
MIDI note: 69

Fig. 5. Example of the Execution of a Musical Note a 440 Hz with the Guitar using the Alto Saxophone Sound through the Mobile Application.



Frequency: 440.73206 Hz
MIDI note: 69

Fig. 6. Example of Playing a Musical Note a 440 Hz with the Guitar using the Acoustic Grand Piano Sound through the Mobile Application.

VI. Discussion

The mobile application was developed with the Java programming language. Due to Java uses a virtual machine (Java Virtual Machine, JVM), which processes the instructions before being executed [19], a considerable latency was obtained in the results. To those who want to return to this article for their research, they are advised to use a lower-level programming language such as C or C++, with which the latency would be reduced considerably and obtain better results, because the instructions are executed directly. Guitar effects pedals commonly use recommended programming languages [20].

VII. Conclusion

The development of mobile applications has acquired great relevance due to the variety of areas where they can be used. As mentioned, there are applications that allow communication between people, planning travel routes, requesting food delivery and entertainment, to name a few examples. Additionally, the applications have allowed people to count with tools that provide the opportunity to explore, grow and develop in other areas such as music, which contributes to their comprehensive training. The mobile application presented in this work benefits this aspect because the user has a synthesizer in the palm of his hand that is capable of emulating monophonic sounds of different musical instruments or other types of sounds with a guitar. This synthesizer, unlike the ones available on the market, is affordable, easily transportable and usable anytime, anywhere. Applications as the one described above allow the practice of music to be accessible to huge number of people.

It has been considered as future work, to make this mobile application a polyphonic synthesizer, where to obtain the polyphony generated by the guitar, the frequency spectrum obtained by the FFT Radix-2 algorithm will be taken up, and search and decision mechanisms will be used. It is also planned to adapt the mobile application to a lower-level programming language like C++ for lower latency.

Acknowledgment

This work was supported by Instituto Politécnico Nacional, COFAA, EDD and SIP. (Project SIP20200832).

References

[1] REAL ACADEMIA ESPAÑOLA, “Diccionario de la lengua española”, 23.ª ed. [Online]. Available: https://dle.rae.es. [Accessed 6 April 2021].

[2] R. Bennett, “Léxico de música”, Madrid: Ediciones Akal S.A., 2003.

[3] Electro-Harmonix, “MEL9”, 2016. [Online]. Available: https://www.ehx.com/products/mel9. [Accessed 25 March 2021].

[4] BOSS, “SY-1000”, 2019. [Online]. Available: https://www.boss.info/mx/products/sy-1000/. [Accessed 25 March 2021].

[5] ROLAND, “GR-55”, 2011. [Online]. Available: https://www.roland.com/mx/products/gr-55/. [Accessed 25 March 2021].

[6] University College Ghent, “TarsosDSP”, 2019. [Online]. Available: https://github.com/JorenSix/TarsosDSP. [Accessed 8 March 2021].

[7] B. Farmer, “Midi Driver”, 2021. [Online]. Available: https://github.com/billthefarmer/mididriver/. [Accessed 8 March 2021].

[8] R. Neuenfeld, M. Fonseca y E. Costa, “Design of optimized radix-2 and radix-4 butterflies from FFT with decimation in time”, 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS), pp. 171-174, 2016.

[9] J. Strawn, C. Abbott, J. Gordon and P. Greenspun, “The Computer Music Tutorial”, London: The MIT Press, 1996.

[10] R. W. Heath, “Introduction to Wireless Digital Communication”, United States of America: PRENTICE HALL, 2017.

[11] V. Montero, “Software para identificación de música”, Sevilla: Universidad de Sevilla, 2020.

[12] D. Takahashi, “Fast Fourier Transform Algorithms for Parallel Computers”, Japan: Springer, 2019.

[13] J. Jamrich, “New Perspectives on Computer Concepts”, United States of America: Cengage Learning, 2018.

[14] P. Blanchard and D. Volchenkov, “Random Walks and Diffusions on Graphs and Databases An Introduction”, Germany: Springer, 2011.

[15] R. Izhaki, “Mixing audio concepts, practices and tools”, Great Britain: Routledge, 2017.

[16] M. Association, “General MIDI”, [Online]. Available: https://www.midi.org/specifications-old/item/general-midi. [Accessed 6 April 2021].

[17] R. S. Pressman and B. R. Maxim, “Software Engineering: A Practitioner's Approach”, New York: Mc Graw Hill Education, 2019.

[18] R. Mall, “Fundamentals of Software Engineering”, Sonepat: PHI Learning Private Limited, 2018.

[19] Oracle, "Java Virtual Machine Technology", [Online]. Available: https://docs.oracle.com/javase/8/docs/technotes/guides/vm/index.html. [Accessed 6 April 2021].

[20] B. Holmes, “Guitar Effects-Pedal Emulation and Identification”, Belfast: Queen’s University Belfast, 2019.

Implementation of Artificial Neural Network in Forecasting Sales Volume in Tokopedia Indonesia

Meiryani[1]
Accounting Department, Faculty of Economics and
Communication, Bina Nusantara University
Jakarta, Indonesia 11480

Dezie Leonarda Warganegara[2]
Doctoral Program, BINUS Business School
Bina Nusantara University
Jakarta Indonesia

***Abstract*—Predicting sales is one way to get company profits. Tokopedia Indonesia is one of the marketplaces that is included in the type of e-commerce customer to customer (C2C). This research was conducted in order to help sellers in the Tokopedia Indonesia marketplace to predict the sales of their merchandise, so that sellers can prepare or stock items that are predicted to increase in sales by implementing Artificial Neural Networks. Artificial neural networks can help predict future sales values. The data is divided into training data and testing data. The results of the analysis of this study indicate that the network model obtained reaches an accuracy rate of 95%.**

***Keywords*—Forecasting; e-commerce; backpropagation; artificial neural network**

I. Introduction

Sales data and information are very important to company to plan sales to be come, for example: customer data, number of vehicles, price cars, spare parts, types of vehicles and those that are not inferior its importance is deep government policy provide vehicle taxes as well as fuel subsidies vehicle [1,2]. The report of the research company said that in early 2017, it increased 51% from 2016 internet users which recorded around 132.7 million internet users [3]. From this data it is known that 24.74 million internet users have shopped online. It is recorded that from 2016 to 2017 internet users spent around IDR 74.6 trillion to shop on various e-commerce (Aditya, 2017). Rebecca (2016) there are six types of e-commerce with different characteristics, namely business to business (B2B), Business to customer (B2C), Customer to Customer (C2C), customer to business (C2B), Business to administration (B2A), and online to offline (OZO) [4]. The C2C e-commerce type model will be the author's focus in this study. Rebecca (2016) state that Tokopedia Indonesia is one of the marketplaces that is included in the C2C type of e-commerce that allows anyone and anywhere to be a seller or a buyer. Tokopedia Indonesia provides various tradable items such as electronic devices, baby equipment, men's and women's clothing, cellphone accessories, laptops, computers, and others. Based on www.alexa.com, it is known that Tokopedia is in the first place for this type of marketplace after the 8th Top Sites on the Alexa Rank [5].

In the Tokopedia marketplace, there is information on the number of items that have been sold, the number of people who have seen the item's page, how long it will take for the goods to be sent to the courier, the number of customers, the description of the goods, the number of people who favor the item and so on [6]. The prediction of the number of sales is an important factor that determines the smooth running business of a company. This prediction is very useful for determining how much goods to be ordered in the following month. Common problems faced by a company is how to predict or forecast sales of goods in the future based on previous sales data. The prediction is very influential to determine sales targets that must be achieved. Planning that effective both in the long term and in the short term depending on the forecast demand for products to be sold [7].

Forecasting techniques are widely used for the planning process and decision making, a prediction trying to predict what will happens and that will be needed. There are artificial neural networks forecasting technique that is often used namely Backpropagation. This technique usually used in multilayer networks with the aim of minimizing error in the output generated by network [8]. Not all items that are seen or favored by many people will experience high demand, but sometimes high enthusiasts are not only seen or favored by many people so that there is an out of stock of goods. To solve this problem, information is needed to those who sell goods and services on Tokopedia by making an analysis that can provide information related to the sales volume for a product or service being promoted. So that with this information, parties who sell goods and services on Tokopedia can provide stock that matches the predicted interest of an item being promoted. One method of overcoming this problem is by designing an artificial neural network architecture or commonly referred to as an artificial neural network. One type of algorithm for this artificial neural network is backpropagation [9]. Backpropagation is an artificial neural network model that is often used and in great demand as a multi-year learning algorithm related to identification, prediction, pattern recognition and so on. The backpropagation algorithm is a type of supervised learning algorithm where the output of the network is compared with the target output so that an error is obtained. Then the error will be propagated back to modify or improve the weight of a network in order to minimize errors. Based on this background, the authors are interested in applying the artificial neural network method with the backpropagation algorithm to predict sales volume on Tokopedia [10].

II. Theoretical Framework

The research entitled analysis of the backpropagation method and radial basis function to predict rainfall with artificial neural networks was carried out by Vincent Rinda Resi (2014). This study discusses the very high rainfall prediction model. The prediction model will be used for various things, one of which is flood prevention. The results obtained from the two methods found that the backpropagation method was able to provide better accuracy, namely 99% than the radial basis function method [11]. [12,13] in their research discusses the prediction of members' interest in a cooperative product. Today's cooperatives, especially the PTPN VII Musi Landas Cooperative, are very much needed by the community because they play an important role in their daily life. The constraints faced by the cooperative are in determining the products that are of interest to its members. If the cooperative can predict this, it will reduce losses and will increase sales which will have an impact on cooperative income. So this research applies a data mining which can predict the interest of members in a product. The data mining technique applied is classification using the decision tree method with the C4.5 and DTREG algorithms. Based on this research, several conclusions were obtained, one of which was to produce information about product categories that were of interest to the members of the PTPN VII Musi Landas Cooperative.

Further research related to artificial neural networks with backpropagation algorithms is a study conducted by [14,15] discuss the backpropagation application discussed to predict the movement pattern of earthquake points in Indonesia from January 2015 to April 2015. The data used is daily data from the Meteorology, Climatology and Geophysics Agency. The conclusion is that the network with momentum and adaptive learning of 0.9813 shows pretty good results with an MSE value of 0.047735 on the 104th iteration/epoch with a maximum of epoch = 10000, learning rate = 0.3, and mc = 0.8. The results of the mapping of the predicted position of the earthquake point are at latitude 0.0405 LU, longitude 124.4015 LS and magnitude 3.26 SR which are in one zone with earthquakes that occur on the same day and date, namely latitude 0.84, LU longitude 126.28 latitude and magnitude 4.8. [16,17] Discusses how to predict the stock prices of Bank Central Asia, Gudang Garam and Indofood. In this study, it was concluded that the artificial neural network parameters obtained MSE with the smallest value obtained with window size 10, hidden neurons 10, and 10000 iterations for all test cases data PT Bank Central Asia Tbk, PT Gudang Garam Tbk and PT Indofood Sukses Makmur Tbk. The results of each test produce MSE 0.002708159 for BBCA data test cases, 0.001074818 for GGRM data test cases and 0.002440852 for INDOFOOD data test cases [22]. Further research is related to the comparison of methods conducted by [18,19] discuss the performance of the three methods in predicting the price of gold because gold is an item that can be used for investment, so an understanding of the shifting of gold prices is needed in order to be able to get a profit. Of the three methods, backpropagation is the best algorithm for predicting gold prices with an accuracy of 95% [20].

III. Research Methodology

The target population used in this study is all data on the Tokopedia Indonesia website. Then the researchers took randomly from one of the categories of goods in Tokopedia Indonesia, namely the computer accessories category. From the data obtained, three types of computer accessories categories will be analyzed by researchers, namely mouse, speaker & sound, and bag & case. The data is obtained by opening the goods web page one by one, then by means of scraping, the researcher selects the features that are considered to affect the variable (Y), namely the number of items that have been sold. The selected features will be used as the supporting variables (X) needed, namely the type of item, the price of the item, the number of people who saw the item, the time of delivery of the goods to the courier, customers from the goods shop, the number of people who have favored the rating. This scraped data is an item that has been advertised on Tokopedia until February 11, 2021. The data analysis method used in this study is an Artificial Neural Network with a backpropagation algorithm to predict buyer interest patterns on the Tokopedia Indonesia website, especially in the computer accessories category. Software used by researchers to analyze the data using Microsoft Excel and RStudio.

IV. Results and Discussion

Preprocessing data is the first step in an analysis to check and correct when there are missing values before starting the learning process [21]. When there is information that is not available on one or more object variables or certain cases, data correction will be carried out. The researcher examined the missing data, can be seen in Table I, as follow:

In Table I, it can be seen that all variables do not have missing data, which means that the next steps can be taken, namely data transformation and data sharing.

Data sharing is intended to divide data into two parts, namely training data and testing data which have their respective functions. The training data is used to train the learning algorithm during the training process. This data distribution is not divided equally, but the percentage for training data is greater than the test data. The following Table II is related to the percentage of data sharing used.

TABLE I. Checking Missing Data

Variable	Valid	Missing	Percentage Valid
Sold	281	0	100%
Type	281	0	100%
Price	281	0	100%
Seen	281	0	100%
Time Send	281	0	100%
Customer	281	0	100%
Favorite	281	0	100%

TABLE II. Data Partition

	Percentage	Total
Training data	80%	224
Test data	20%	57
Total	100%	281

Source: processed data

Table II above can be seen that the percentage of training data is greater than the test data because the learning algorithm while carrying out the training process works optimally.

The determination of input and output patterns is based on the formulation of this research problem. So that there are seven variables as input that are considered influencing the target (output). Determining Network Architecture and Parameters Artificial Neural Network has an architecture consisting of the number of layers and the number of neurons in each layer, as for the case of backpropagation using multi layers consisting of input, hidden, and output. The number of hidden layers 1 alone is sufficient to produce output that matches the target [23,24]. So that the network architecture designed for this research is 3 layers (input, hidden, and output) with 7 neurons for the input layer, 3 neurons for the hidden layer, and 1 neuron for the output layer [29].

Initialization of weights and bias is given before carrying out the training process of an existing network system in an artificial neural network. This initial initialization weight is given to each neuron that is interconnected. This weight factor defines the relationship between neurons with one another, where the greater the weight value of a relationship between neurons, the more important the relationship between the two neurons. Initialization of initial weights and bias is done randomly. Table III show the initial weights and bias of the input layer against the hidden layer and Table IV show initial weights and bias on the hidden layer to the output layer [25,26].

TABLE III. INITIAL WEIGHTS AND BIAS ON THE INPUT LAYER AGAINST THE HIDDEN LAYER

Variable				
		1	2	3
Bias	0	-0.53544204	1.256897569	-0.135689723
X_1	1	0.054219104	- 0.772365469	1.1365698754
X_2	2	0.617251692	0.678952432	2.5326546321
X_3	3	0.145236548	- 0.865321469	- 0.6987563212
X_4	4	-1.534256987	0.598756975	0.4659879464
X_5	5	0.568795643	0.123646897	
X_6	6	1.285695467	- 0.256458976	-1.856987965
X7	7	0.789653217	- 0.256987795	2.3698575426

Source: processed data

TABLE IV. INITIAL WEIGHTS AND BIAS ON THE HIDDEN LAYER TO THE OUTPUT LAYER

$W_{[j,]}$	$W_{[,k]}$
	1
0	1.3564795645
1	1.8563256987
2	0.2356489976
3	1.1698756975

Source: processed data

Training with the backpropagation algorithm is an algorithm with a supervised learning process. After obtaining the initial initialization weight and bias, a training process will be carried out on a network that has been designed for architecture and the parameters that have been determined using the training data that has been determined; the percentage is 80% of the total data. There are three phases of the training process for the backpropagation algorithm, namely feedforward, backpropagation, and weight modification. Feed forward (feedforward) at this stage, an error will be searched for the output in the forward direction (forward). Each input unit Xi (i: 1, ..., n) will receive the input signal xi then pass it to the hidden unit. Then all the weighted input signals will be calculated including the bias in each hidden unit Zj (h: 1, ..., p). The following is the input signal to the weighted hidden layer, including its bias.

After the hidden layer receives a weighted input signal (Table V) including the bias, the output signal will be calculated in the hidden layer from the input signal using the activation function. The following is the output signal on the Hidden Layer.

Then the output signal in the hidden layer in Table VI will play a role as an input signal in the output layer. The input signal will be forwarded to the output layer with the weights and bias hidden in the hidden layer against the output layer for each output unit.

After the output layer receives the input signal (Table VII) from the hidden layer, the input signal will be activated using the activation function. Table VIII shows output signals at the output layer.

Once activated in the output layer, the output will be distributed to all units in the output layer.

TABLE V. INPUT SIGNALS FROM INPUT LAYER TO HIDDEN LAYER

Z	Z_{netj}
1	0.116563201
2	0.623654261
3	3.133356987

TABLE VI. OUTPUT SIGNALS ON HIDDEN LAYER

Z	Z_{netj}
1	0.116563201
2	0.623654261
3	3.133356987

TABLE VII. INPUT SIGNAL FROM HIDDEN LAYER TO OUTPUT LAYER

Z	$Y_{net\ k}$
1	3.523657952

Source: processed data

TABLE VIII. OUTPUT SIGNALS AT THE OUTPUT LAYER

Y	Y_k
1	0.92361236

Source: processed data

To calculate the error between the input target and the output generated by the network with the result of the error factor () of -0.022507221. The error factor will be used to correct the weight (Wjk) and bias (W0k) in the lower layer (hidden layer) with the learning rate (α). The results of the weight improvements in the hidden layer to the output layer are as follows. Table IX show correction of weights and bias in the hidden layer ot the output layer.

Each hidden unit Zj accepts input delta weights and bias from the layer above it (output layer). The weight and bias delta input will be used to find the error factor in each hidden unit. Table X is the result of calculating the error factor in each hidden unit.

Input from the error factor in the hidden unit such as Table XI will be activated using the activation function with the following results.

After obtaining the error factor that has been activated with the activation function, the error factor will be used to correct or correct the weights and biases in the lower layer, namely the input layer to the hidden layer with calculations as in equation 16. Table XII shows correction of weights and bias in the input layer against the hidden layer, as follows.

TABLE IX. CORRECTION OF WEIGHTS AND BIAS IN THE HIDDEN LAYER OF THE OUTPUT LAYER

$\Delta W_{[j.]}$	$\Delta W_{[.k]}$
	1
0	-0.00021456
1	-0.00012341
2	-0.00019457
3	-0.00030123

Source: processed data

TABLE X. RESULT OF CALCULATING THE ERROR FACTOR IN EACH HIDDEN UNIT

Z_j	$\S_{net}$ j
1	-0.03265
2	-0.00365
3	-0.02456

Source: processed data

TABLE XI. INPUT FROM THE ERROR FACTOR IN THE HIDDEN UNIT

Z_j	$\S_j$
1	-0.01299
2	-0.00055
3	-0.00068

Source: processed data

TABLE XII. CORRECTION OF WEIGHTS AND BIAS IN THE INPUT LAYER AGAINST THE HIDDEN LAYER

$\Delta V_{[i.]}$	$\Delta V_{[.j]}$		
	1	2	3
0	-0.00010032	-6.06865E-06	-9.1251E-06
1	-5.006E-05	-3.56987E-06	-4.9658E-06
2	-1.7369E-05	-1.08658E-06	-1.5968E-06
3	-1.0852E-05	-6.06523E-07	-1.4236E-06
4	-1.0016E-05	-6.86523E-07	-9.1236E-07
5	-1.0865E-05	-7.36563E-07	-1.0563E-06
6	-1.093E-05	-6.69852E-07	-1.2354E-06

This training process will continue as long as the conditions are not met. Training will stop when an optimal error has been obtained, so that the final weight and bias for each layer is obtained. The final weights and biases with one step obtained from the artificial neural network with the backpropagation algorithm that has been designed using training data [27]. Table XIII show weights and final bias on the input layer against the hidden layer, as follows:

TABLE XIII. WEIGHTS AND FINAL BIAS ON THE INPUT LAYER AGAINST THE HIDDEN LAYER

Variable	$V_{[i.]}$	$V_{[.j]}$		
Bias		1	2	3
X_1	0	-1.43256	1.569857	-0.22568
X_2	1	-0.45628	-0.55698	1.196520
X_3	2	0.569863	0.896423	2.756982
X_4	3	0.063595	-0.25697	-0.89652
X_5	4	-1.73658	0.869545	2.698656
X_6	5	0.563985	-0.5879	0.756982
X_7	6	1.569758	-0.13995	-1.65327

Source: processed data

The artificial neural network predicts sales volume in the Tokopedia Indonesia marketplace with training data first to find out the accuracy of the network before testing the network. The prediction results using an artificial neural network model obtained using training data.

TABLE XIV. PREDICTION RESULTS OF THE NETWORK MODEL

No	Data To	Actual	Predictions	Time Range (Month)
1	56	78	78	32
2	166	62	61	9
3	218	45	61	13
4	90	73	62	33
5	90	73	63	8
.				
.				
.				
243	28	38	39	17
244	13	87	74	9
245	231	70	64	15

The results of the training data prediction using the network model obtained as shown in Table XIV above in predicting sales volume in the Tokopedia Indonesia marketplace with a span of time in months. Obtained is an accuracy rate of 95.75%. Testing of artificial neural networks with the backpropagation algorithm obtained after calculating the weight until it reaches optimal then testing the network obtained will be carried out. The network will be applied to the test data to determine the network's performance in predicting sales volume on the Tokopedia Indonesia's marketplace. The percentage of test data was 20% of the total data used, namely as many as 57 test data. The following are some of the results of the prediction of sales volume on the Tokopedia Indonesia marketplace using the network obtained [28].

TABLE XV. PREDICTION RESULTS OF TESTING DATA USING A NETWORK MODEL

No	Data To	Actual	Predictions	Time Range (Month)
1	1	77	71	29
2	6	101	56	24
3	75	79	78	14
4	76	56	61	11
5	159	68	79	16
.	.	.	.	
.	.	.	.	
.	.	.	.	
46	175	69	62	29
47	229	62	63	31
48	269	49	63	22

The prediction results from network testing using this test data show the network performance obtained. From the results of the prediction of sales volume in the Tokopedia marketplace with a span of time in months using test data, an accuracy rate of 95.75% is obtained. The network model obtained with the optimal level of accuracy is found when testing the network, so it can be said that the resulting network performance is very good. Regarding the sales volume prediction made, the seller at Tokopedia can provide stock according to the predictions obtained with the time span listed in Table XV, so that the seller can minimize the occurrence of losses and will increase their income [18, 29].

V. CONCLUSION

Based on phenomenon, research question, result and discussion the conclusion as follows:

1) The results of this study indicate that Backpropagation has a good level of accuracy in predicting sales.

2) The Artificial Neural Network method has an adaptive nature, namely the network tries to achieve data stability to achieve the expected output value.

3) The resulting artificial neural network architecture design results consist of three layers which include six neurons in the input layer, three neurons in the hidden layer, and one neuron in the output layer. The parameters used to form the network model include the learning rate with a value of 0.02 and the activation function used is binary sigmoid (logistic).

4) The resulting level of accuracy when testing the obtained network reaches an accuracy rate of 95.75%.

5) The sales forecasting process is to enter the estimated future sales data, to be processed using the backpropagation neural network to produce the desired data.

ACKNOWLEDGMENT

This work was supported by BINUS University, Jakarta, Indonesia.

REFERENCES

[1] Azhar Susanto & Meiryani. 2019. Antecedents of environmental management accounting and environmental performance : Evidence from Indonesian small and medium enterprises. International Journal of Energy Economics and Policy. 9(6), pp. 401-407.

[2] Bahadir, E. 2016. Prediction of Prospective Mathematics Teachers' Academic Success in Entering Graduate Education by Using Back-propagation Neural Network. Journal of Education and Training Studies. Vol. 4 (5): 113-122.

[3] Chakraborty, K., Mehrotra, K., Mohan, C. K., & Ranka, S. 1992. Forecasting the Behavior of Multivariate Time Series Using Neural Networks. Neural Networks.Vol. 5: 961-970.

[4] Fausett, L. 1994. Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. New Jersey: Prentice-Hall

[5] Fahruroji, A. 2014. Getting to know the E-commerce Business Model. https://afahrurroji.net/mengenal-model-bisnis-e-commerce/. Retrieved February 20, 2018.

[6] Fajri, N. 2011. Temperature Prediction Using Algorithms-Algortima in Artificial Neural Networks. Thesis. Bandung Institute of Technology: Bandung.

[7] Fausett, L. 1994. Fundamentals of Neural Networks: Achitectures, Algorithms, and Applications. New Jersey: Prentice Hall. Herdianto. 2013. Prediction of Induction Motor Damage Using Backpropagation Neural Network Method. Thesis. University of North Sumatra: Medan.

[8] Hidayatullah, A. I. 2017. Backpropagation Algorithm For Aircraft Delay Prediction Due to Weather. Thesis. Faculty of Mathematics and Natural Sciences, Indonesian Islamic University: Yogyakarta.

[9] Hardianto, H. N. I., Suyanto, & Purnama, B. 2011. Analysis and Implementation of Differential Evolution and Recurrent Neural Network for Time Series Data Prediction Case Study of Gold Selling Rate. Thesis. Telkom University.

[10] Hikmah, A. 2017. Time Series Forecasting Using Autoregressive (AR), Radial Basis Function Artificial Neural Network (RBF), and AR-RBF Hybrid in Indonesian Inflation. Unnes Journal of Mathematics. Vol. 7 (2): 1—14.

[11] Ita Qorry Aina. 2018. Implementation of Artificial Neural Network with Backpropagation Algorithm to Predict Sales Volume at Bukalapak. Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia. Yogyakarta.

[12] Kusumadewi, S. and Kiki. 2010. Analysis of Artificial Neural Networks with Backpropagation Method to Detect Psychological Disorders. FTI, Islamic University of Indonesia. Kusumadewi, S. 2003. Artificial Intelligence Techniques and Applications. Yogyakarta: Graha Science. McLeod, P. 2008. Management Information Systems. Jakarta: Salemba.

[13] Meiryani and Azhar Susanto. 2018. The Influence of Information Technology on The Quality of Accounting Information System. ACM International Conference Proceeding Series. Pp.109-115

[14] Nafi'iyah, N. 2016. Comparison of Linear Regression, Backpropagation, and Fuzzy Mamdani in Gold Price Predictions. Journal of the National Seminar on Innovation and Technology Application in Industry (Seniati) National Institute of Technology.

[15] Pratama, A.P. 2017. Development of Internet Users in Indonesia in 2016, the World's Largest. https://id.techinasia.com/pertumbuh-

penggunainternet-di-indonesia-tahun-2016. Retrieved February 20, 2018.

[16] Puspitaningrum, D. 2006. Introduction to Artificial Neural Networks. Yogyakarta: Andi Offset. 48 Ramadha, W. I. 2016. Stock Price Prediction Using Resilient Backpropagation Neural Network. Thesis. Not published. Faculty of Mathematics and Natural Sciences, Gadjah Mada University: Yogyakarta.

[17] Rebecca. 2016. Types of E-commerce & Examples. https://www.progresstech.co.id/blog/jenis-e-commerce/. Retrieved February 20, 2018. Resi, V.R. 2014. Comparative Analysis of Backpropagation Method and Radial Basis Function to Predict Rainfall with Artificial Neural Networks. Journal of the Faculty of Computer, Dian Nuswantoro University.

[18] Ryanda, et al. 2015. Designing a Mobile Application "Kiosku.Com" with Web Scrapping on the Olx.Co.Id, Berniaga.Com, and Bukalapak.Com Websites based on Android. e-Proceeding of Engineering: Vol. 2, No.2.

[19] Ryandi, et al. 2014. Application of Data Mining to Predict Members' Interest in Ptpn Vii Musilandas Cooperative Products. Journal of Informatics Engineering Students.

[20] Saputra, B. Y. 2015. Implementation of Artificial Neural Networks with Backpropagation Algorithm to Predict Earthquake Point Movement Patterns in Indonesia. Thesis. Faculty of Mathematics and Natural Sciences, Indonesian Islamic University: Yogyakarta.

[21] Salman, A. G., & Prasetio, Y. L. 2010. Implementation of Recurrent Neural Networks Using the Gradient Adaptive Learning Rate Learning Method for Rainfall Estimation Based on ENSO Variables. ComTech Journal. Vol. 1 (2): 418-429.

[22] Semen Indonesia. 2017. The Demand for Cement Will Continue to Increase. http://www.semenindonesia.com/perminta-semen-bakal-terus-naik/.

[23] Siang, J. J. 2004. Neural Networks and Its Programming Using Matlab. Yogyakarta: Publisher Andi.

[24] Suhada, B. 2009. Forecasting National Sugar Production Using Artificial Neural Network Approach. Derivatives Journal. Vol. 3 (1): 50-63.

[25] Susanti, L. A. D., Arna, F., & Sethiawardana. 2013. Forecasting Stock Prices Using Recurrent Neural Network with Backpropagation Through Time (BPTT) Algorithm. Final Project Papers. Surabaya: Sepuluh Nopember Institute of Technology.

[26] Sutono, S. B. 2008. Causal Forecasting Analysis Based on Principal Component Analysis of Artificial Neural Networks for Industrial Engineering Applications. Thesis. Pelalawan College of Technology: Riau. Turland, M. 2010. Php. | architect's Guide to Web scraping with PHP. Introduction-Web Scraping Defined, str, 2. Vermaat, S. C. 2007. Discovering Computers: Exploring Fundamental Computer Dubia. Edition 3. Jakarta: Salemba Infotek.

[27] William and Sawyer. 2007. Using Information Technology. Yogyakarta: Andi. Wong, J. 2010. Internet Marketing for Beginners. Jakarta: PT Elex Media Komputindo. 49 Zaira, Z. 2011. Implementation of Web Extraction for Hadiths Translated in Indonesian. Thesis. University of Indonesia: Depok.

[28] Valipour, M., Banihabib, M. E., & Behbahani, S. M. R. 2013. Comparison of the ARMA, ARIMA, and the Autoregressive Artificial Neural Network Models in Forecasting the Monthly Inflow of Dez Dam Reservoir. Journal of Hydrology (online). Vol. 476: 433—441.

[29] Winwin Yadiati & Meiryani. 2019. The role of information technology in E-Commerce. International Journal of Scientific and Technology Research 8(1), pp. 148-150.

Design and Evaluation of Bible Learning Application using Elements of User Experience

Frederik Allotodang[1], Herman Tolle[2]
Computer Science Department
Brawijaya University, Malang, Indonesia

Nataniel Dengen[3]
Computer Science Department
Mulawarman University, Samarinda, Indonesia

***Abstract*—Technological developments can encourage children to learn easily and help solve problems that often arise in the learning process. Sunday School students need learning media to make it easier to understand Christian Education. The method used in Sunday Schools still uses conventional methods which include face-to-face teaching and learning in class. This method often faces various challenges such as student's lack of focus. One of the solutions that are proposed in this paper is to design an Android-based learning application that will support the learning process. Application User Interface and User Experience (UI/UX) design will be built based on the Elements of User Experience methodology. The Elements of User Experience method will be used in the analysis and design process to maximize the usability and engagement level of the application. The learning materials will be designed based on Attention, Relevance, Confidence, and Satisfaction (ARCS) framework. ARCS will help the material design process to ensure the clarity and appropriateness of the material. The application will be implemented and tested on students to measure its effectiveness. The application trial has shown a promising improvement especially in student's engagement toward the materials.**

***Keywords*—Christian education; Sunday school; element of user experience; ARCS; android application**

I. Introduction

Christian Education nowadays is built on several objectives such as increasing student's religious knowledge and understanding of the Bible's contents and applying Christian values in their daily life [1]. Christian Education is implemented in many Christian churches through Sunday Schools. Sunday School is a Christian religious education activity intended for children aged 5-15 years and is usually held once a week every Sunday. Christian character is the concern of Christian parents in shaping the character of their children. The standard of Christian character to be achieved is a standard based on the Bible, not based on world philosophy [2]. Sunday school service is usually held in conjunction with the Public Service and is divided into several classes according to age ranges. The purpose of this class division is to adapt the material provided to suit the psychological age of each child. The teaching system in Sunday Schools also faces several problems. In modern and digital times, the teaching methods and media in many Sunday schools are considered to be lagging behind. Most students who are familiar with digital communication media such as cell phones and electronic tablets are more interested in accessing the internet and playing online games than in attending classes and activities in Sunday School. The teaching methods in most Sunday schools still use conventional methods. Teaching and learning activities include the teacher giving and presenting the material in front of the class. Teaching materials are given in written form on whiteboard in front of the class. As a result, many of the children did not pay attention to what the teacher gave because they were considered less attractive and the teachers also did not involve each child to contribute in the teaching and learning process. Sunday Schools are expected to be able to innovate so that the material taught becomes more interesting and relevant for children.

One of the innovations that want to be applied to overcome this problem is through designing a learning application for Sunday School using the Elements of User Experience methodology. The Elements of User Experience is a methodology used to develop a user interface and user experience (UI/UX) system which is divided into 5 stages, including Strategy, Scope, Structure, Skeleton, and Surface [3]. After the application is implemented, the next step is to measure the level of usability of the application and how effective it supports the learning process in Sunday Schools. This research is expected to provide solutions to overcome the problems faced in the learning process in most Sunday schools.

Several studies have been conducted in the application development process using the Elements of User Experience methodology. Most of the existing studies focus more on general-purpose applications and very few focus on educational applications especially for Christian Religious Education for Children. This research is expected to contribute in particular to combining application development methods with educational methods for children.

II. Previous Study

Several studies have been conducted based on the Elements of User Experience methodology. This method can be applied in web and mobile-based applications and several studies have tried to apply it. Ecelbarger, Hamlin, and McMcGrath have done this in the case of a web-based scheduling application [4]. Tong, Cui, and Chen have applied it to a more interactive mobile application like museum simulations [5]. Both studies have reported better results in terms of user experience. However, the application of this method is still very limited for educational-purpose applications. The biggest challenge in designing education-based applications is how to ensure that the applications made can increase the engagement of the learning process being carried out. Similar research was conducted by Nurul and Norasykin to develop a computer-programming learning application but it is more aimed at an adult audience with a simple and minimalistic user interface [6].

This research has its challenges because it has an audience of children aged 5-10 years who have a different application development approach to the adult audience. Several phases have to be modified to make a better approach to the problem. One of the main modifications is to combine the ARCS methodology at the phase of determining the application structure [7]. ARCS serves to ensure that the teaching materials provided are suitable for the intended learning abilities of the students.

III. Methodology

A. Elements of User Experience

The Elements of User Experience consists of five stages and starts from the lowest stage, Strategy stage. This stage is used to collect data related to system development that will be carried out to get product objectives and user needs. This stage determines the purpose of the application being made and the specific needs of Sunday School children in order to be able to design an appropriate system and suitable to user needs.

The second stage, Scope stage aims to determine the scope of the system you want to build and clearly define the system requirements. The strategy to be used is to design system requirements in accordance with the results of the questionnaire that was implemented in the previous stage.

The Structure stage contains a systematic design for how the system will run and how the system responds to users. The method used to describe interactions with users is to use a use-case diagram to explain in detail how users interact with the system being built, and for the information storage process, ER Diagram is used as an explanation of how the information will be stored.

The Skeleton stage is the stage for designing a system prototype to provide an overview of how the system works. The method will use mockups to explain how the existing display in the application will be developed.

The Surface stage is the final stage for designing the final result of the application being built. The entire process can be seen in Fig. 1.

B. Design Process

The stage of making the application structure provides functional description of how the system works. The functional description of the system is manifested in the form of a simple interface design of an application mockup. Mockups are structured models or replicas that describe the appearance of the application and the flow of application functions. An example of the application mockup can be seen in Fig. 2.

C. Implementation Process

At the implementation stage, mockups that were developed in the previous stage will begin to be built using the selected technology, React Native and MySQL database. React Native is a framework that allows mobile application development using the JavaScript programming language [8].

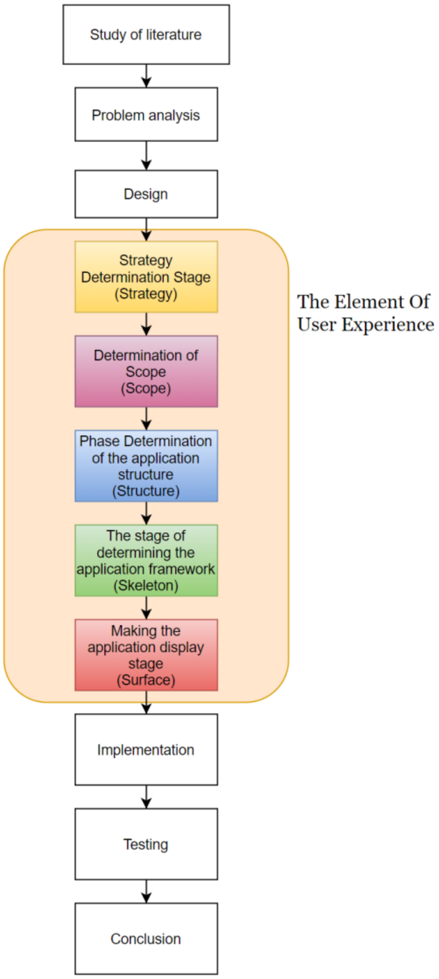

Study of literature
Problem analysis
Design
Strategy Determination Stage (Strategy)
Determination of Scope (Scope)
Phase Determination of the application structure (Structure)
The stage of determining the application framework (Skeleton)
Making the application display stage (Surface)
The Element Of User Experience
Implementation
Testing
Conclusion

Fig. 1. Research Methodology.

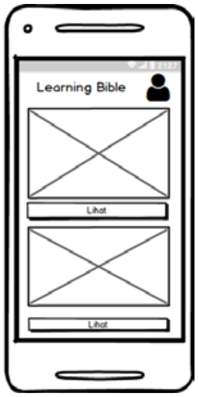
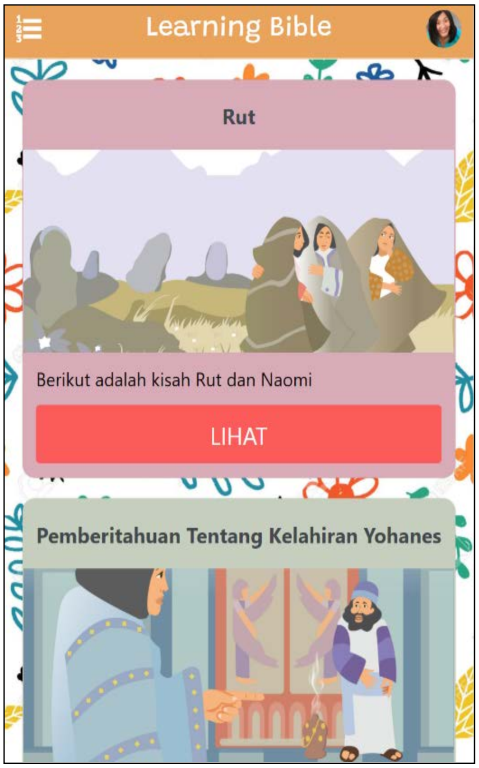


Fig. 2. Application Mockup.

User interface implementation is the final stage of The Elements of User Experience methodology. The mockup created in the previous process will be manifested as a user interface and becomes the blueprint of the actual application display development process. The existing application is designed to be as simple and attractive as possible, with relatively bright colors, considering the relatively young age range of users. An example of a user interface that has been developed can be seen in Fig. 3.



Berikut adalah kisah Rut dan Naomi

Pemberitahuan Tentang Kelahiran Yohanes

Fig. 3. Application user Interface.

D. Testing Process

This test was carried out at the Toraja Church in Samarinda, East Kalimantan. Tests will be carried out on students of the Toraja Church Sunday School who are in the age range of 5-10 years. Sunday School students will also be accompanied by teachers during the testing process. The method used in testing is the System Usability Scale (SUS) [9].

SUS is a method of testing the usability of an application containing 10 questions with an assessment range between 1 and 5 [10]. The SUS list of questions contains questions to determine whether the application is easy enough to use by its designated users, especially children. A list of SUS questions can be seen in Table I.

TABLE I. SUS QUESTIONS

No	Questions
1	I think I will use this application often
2	I think this application is too complicated but can be simplified
3	I think this application is easy to use
4	I think I need help from a technical person to be able to use this application
5	I found that there are various features that are well integrated in this application
7	I think there are many things that are inconsistent with this application
8	I think the majority of users will be able to learn this application quickly
9	I find this application very impractical to use
10	I really believe that I can use this application

After the user has filled in, the method for calculating the SUS score is as follows:

- For each odd question, the score given by the user must be reduced by 1 (formula: X-1).
- For each even question, subtract the value given by the user from 5 (formula: 5-X).
- The value obtained from each question is then added and multiplied by 2.5.
- The result is a SUS score and should be in the 0-100 range General guidelines on the interpretation of SUS scores can be seen in Table II.

TABLE II. SUS GUIDELINES

SUS Score	Grade	Predicate
> 80,3	A	Excellent
68 – 80,3	B	Good
68	C	Okay
51 – 68	D	Poor
< 51	E	Awful

IV. RESULTS AND DISCUSSION

After the application design process, the application was tested on 24 user respondents. The trial was carried out for 30 minutes for each respondent where 20-25 minutes was the application trial stage and 5-10 minutes was used to fill out the questionnaire. The questionnaire was designed based on the System Usability Scale (SUS) method which has 10 questions. Each respondent was also accompanied by a Sunday School teacher to assist in the testing process and filling out the questionnaire. The results of the application trial can be seen in Table III.

TABLE III. SUS GUIDELINES

QUESTIONS	Score Given by User Respondents				
	1	2	3	4	5
I think I will use this application often	1	1	2	16	4
I think this application is too complicated but can be simplified	2	16	3	2	1
I think this application is easy to use	0	4	1	17	2
I think I need help from a technical person to be able to use this application	3	15	5	1	0
I found that there are various features that are well integrated in this application	0	1	1	16	6
I think there are many things that are inconsistent with this application	2	16	6	0	0
I think the majority of users will be able to learn this application quickly	0	1	2	16	5
I find this application very impractical to use	4	17	2	1	0
I really believe that I can use this application	0	2	2	16	4
I have to learn many things first before I can use this application	3	17	2	2	0

The trial results were then calculated to obtain a score for each question. For odd questions, the formula used is X-1 where X is the score given by the user. The sum is then divided by 24 (according to the number of respondents). On the other hand, for even questions, the formula used is 5-X. The sum is then divided by 24. The SUS Calculation Results can be seen in Table IV.

TABLE IV. SUS CALCULATION RESULTS

Questions	**SUS Calculations**					**Score**
	1	**2**	**3**	**4**	**5**	
I think I will use this application often	0	1	4	48	16	2,88
I think this application is too complicated but can be simplified	8	48	6	2	0	2,67
I think this application is easy to use	0	4	2	51	8	2,71
I think I need help from a technical person to be able to use this application	12	45	10	1	0	2,83
I found that there are various features that are well integrated in this application	0	1	2	48	24	3,13
I think there are many things that are inconsistent with this application	8	48	12	0	0	2,83
I think the majority of users will be able to learn this application quickly	0	1	4	48	20	3,04
I find this application very impractical to use	16	51	4	1	0	3,00
I really believe that I can use this application	0	2	4	48	16	2,92
I have to learn many things first before I can use this application	12	51	4	2	0	2,88
Final SUS Score (Total score x 2.5)						72,19

The score obtained from each question is then added and multiplied by the constant 2.5 and the final SUS score is 72.19. The score based on Table II shows an outcome with a good predicate. A good SUS score indicates an increase in the ease of the learning process.

V. CONCLUSION

The Elements of User Experience allows the development of an Android-based application that suits the needs of users, especially students in the age range of 5 to 10 years old. The testing result using the System Usability Scale (SUS) method produces a final value of 72.19. Based on the SUS score guidelines, the score shows a good result that indicates an increase in the ease of the learning process.

These results also show that the Element of User Experience method can be applied to the development of mobile-based applications for educational purposes. This method will increase the usability of the application which is a major factor for its designated users.

Further research can be carried out using the same method but in different cases and age ranges. This study is focused on target users with an age range between 5-10 years old. Research at different age ranges will provide useful information on how effective this method is at younger or older ages.

REFERENCES

[1] Busthan Abdy (2019). Pendidikan Agama Kristen (PAK) dan Etika Kristen. Kupang: Yayasan Kasih Imanuel Mahawira.

[2] Hartono. Handreas. (2014). Membentuk Karakter Kristen Pada Anak Keluarga Kristen. Jurnal Kurios: Jurnal Teologi Dan Pendidikan Agama Kristen, Vol. 2, No. 1, s.

[3] Garrett, Jesse James. (2011) The Elements of User Experience: User-Centered Design for the Web and Beyond, Second Edition, California: New Riders.

[4] A.N Ecelbarger, P.D. Hamlin, S.C McGrath. Dkk. (2020) User Experience Design to Synchronize Government Acquisition Strategy and Schedule.

[5] Yanting Tong, Binyue Cui, and Yulin Chen. (2018) Research on UI Visual Design of Intangible Cultural Heritage Digital Museum Based on User Experience.

[6] Nurul F,Z & Norasykin M,Z (2018) Code Puzzle: ActionScript 2.0 Learning Application Based on Problem Based Learning Approach University Teknology Malaysia.

[7] Pratama, A. (2015). Pengembangan Multimedia Pembelajaran Berbasis Educational Games dnegan Menggunakan Strategi Pembelajaran ARCS.

[8] Xingwei Zhou, Wenshan Hu, and Guo-Ping Liu. (2020) React-Native Based Mobile App for Online Experimentation.

[9] Brooke, J. 1996 . SUS - A Quick and Dirty Usability Scale. United Kingdom: Redhatch.

[10] Rasmilla (2018). Evaluasi Website Dengan Menggunakan System Usability Scale (SUS) Pada Perguruan Tinggi Swasta di Palembang Volume 4, Hal. 89 – 98.

A Markerless-based Gait Analysis and Visualization Approach for ASD Children

Nur Khalidah Zakaria[1], Rozita Jailani[3]
School of Electrical Engineering, College of Engineering
Universiti Teknologi MARA (UiTM)
Shah Alam, Selangor, Malaysia

Nooritawati Md Tahir[2]
School of Electrical Engineering, College of Engineering
Institute for Big Data Analytics and Artificial Intelligence
Integrative Pharmacogenomics Institute (iPROMISE)
Universiti Teknologi MARA (UiTM)
Shah Alam, Selangor, Malaysia

***Abstract*—This study proposed a new method in gait acquisition and analysis for autistic children based on the markerless technique versus the gold standard marker-based technique. Here, the gait acquisition stage is conducted using a depth camera with a customizable skeleton tracking function that is the Microsoft Kinect sensor for recording the walking gait trials of the 23 children with autism spectrum disorder (ASD) and 30 typically healthy developing (TD) children. Next, the Kinect depth sensor outputs information is translated into kinematic gait features. Further, analysis and evaluation are done specifically the kinematic angles of the hip, knee, and ankle in analyzing and visualizing the pattern of the plots versus the kinematic plots acquired from the marker-based that is the Vicon motion system gait technique. In addition, these kinematic angles are also validated using the statistical method namely the Analysis of Variance (ANOVA). Results showed that the ρ-values are insignificant for all angles upon computing both the intra-group and inter-group normalization. Hence, these findings have proven that the proposed markerless-based gait technique is indeed apt to be used as a new alternative markerless method for gait analysis of ASD children.**

***Keywords*—Autism spectrum disorder (ASD); kinematic; marker-based; markerless-based; gait analysis**

I. Introduction

Autism Spectrum Disorder (ASD) is a development disorder that can be characterized by several difficulties such as in the learning process, communication, as well as social skills [1]. Note that the deficiency can be seen in early childhood, for instance as early as two years old [2] & [3]. Recently, there has been a growth in the literature regarding unbalanced and sensory disturbances, especially during walking for the ASD children [4], [5] & [6]. Due to the sensory conflicts, researchers face challenges in performing gait assessments to further provide suitable care and treatment [7]. On top of that, ASD children are found to face more challenges in receiving both the treatment and care as compared to other children with developmental delays, especially the requirement for these children to be examined in an unfamiliar testing environment [7], [8] & [9]. Here, gait analysis involving ASD children is indeed vital. This is because reliable gait information will help determine the level of deficit specificity as the evaluation tools [5].

II. Related Work

There are two categories of gait techniques namely the marker-based technique and markerless-based technique. The marker-based technique includes an optoelectronic system with synchronization between the camera(s) and the system. This system works by tracking the movement through an optical sensor (infrared camera) by identifying the position of the markers attached to the body or object followed by post-processing via specialized software [10]. This technique also offered an accurate three-dimensional (3D) model and is normally performed inside the laboratory [10] & [11]. Moreover, the position of the markers must properly adhere according to the system's template to produce accurate gait data and analysis [12]. For example, Hasan et al. [13] has placed 35 retro-reflective markers on the ASD and typically healthy developing (TD) subject’s body to analyze the gait pattern from two groups whilst, Eggleston et al. [14] used 19 retro-reflective markers to perform analysis of gait symmetry specifically at the lower body during the over-ground walking for the ASD children.

On the other hand, researchers attempt to overcome the limitation from the marker-based techniques with marker-free or markerless gait techniques and thus giving greater freedom of movement to the users [15]. This technique offers a low-cost system as compared to the marker-based technique [16]. The analysis for marker-less technique is based on the movement analysis specifically the subject’s silhouette in contrast to the background [17]. Thus, markerless requires post-processing analysis to enhance the appearance of the silhouette or to translate the movements into a three-dimensional (3D) coordinate point that can be mapped by the computer [17] & [18]. For example, Vilensky et al. [19] used two high-speed cameras, with one camera was placed perpendicularly with the walking track while the other camera was facing the track to measure gait disturbance between the normal and autistic children at the hip, knee, and ankle joints. Meanwhile, Zakaria et al. [20] performed ASD gait classification based on the calculated distance between joints captured by the depth camera.

In another study, Al-Jubouri et al. [21] classify full-body movement of the ASD and normal children based on the four-stage approaches which are augmentation, feature extraction, dimension reduction, and rough set. Based on previous researches, it was found that there are not many studies conducted related to the gait features using the markerless-based technique as reported in [22] & [23]. On top of that, there is a need to validate and evaluate the markerless-based features to provide an accurate analysis and evaluation method [24] that further can be used as the evaluation tools for ASD [5]. Earlier, the validity of the Kinect sensor was compared with the Vicon system as reported by Abiddin et al. [25]. The validity has been measured for the assessment of postural control based on several different types of movements. Here, selected body points from Kinect were imposed with the marker trajectories from Vicon to measure the outcome of anatomical landmark displacements or changes in angle relative to the ground [26]. Based on the Pearson correlation, the r-values obtained were greater than 0.90 for the majority of measurements as reported in [26]. In a different study, the intra-class (ICC) and Pearson analysis (r-value) of the vertical displacement of the knee marker for Parkinson Disease (PD) patients obtained using the Vicon system and Kinect sensor were compared and it was found that both techniques achieved ICC and r greater than 0.9 for walking on the same spot [27]. On the other hand, Cocchi et al. [28] have estimated the sagittal joints kinematics of children with cerebral palsy as well. In this study, the validity of the marker-less kinematic features was verified by comparing the markerless gait features with the marker-based gait features and results showed that the differences were heavily affected by the presence of the offset attained at the mean values of the joint kinematics. However, upon removing the angular offsets, the Pearson correlation specifically the r-values obtained were between 0.8 and 1, which indicated a good result [28]. To the extent of our knowledge and based on findings from previous studies, there is no formal study that has been done related to evaluation and validation for ASD using the markerless gait feature. Therefore, this study aims to evaluate and validate the kinematic gait features of both ASD and TD groups using a viable method like the Kinect sensor and the gold standard Vicon gait system. The use of the markerless-based gait technique can be of benefit to ASD children since most of these children often confront physical challenges such as lack of focus and anxiety during the laboratory gait trials acquisition and experiment.

III. Methodology

This section elaborated in detail the method used and the overall process methodology in this study. The experiment was conducted in the Human Motion and Gait Analysis (HMGA) Premier Laboratory of Universiti Teknologi MARA (UiTM), Shah Alam, Selangor. In this study, data acquisition consists of 23 ASD and 30 TD children. Prior to this study, parents were given an information sheet that needs to be completed in the Consent Form based on recommendation and approval by UiTM Shah Alam Ethics Committee.

The mean and standard deviation (SD) of the subjects is tabulated as in Table I. The ASD children have a mean age, height, and body mass of 8.391 (0.396) years, 1.267 (0.028) m, and 32.957 (3.485) kg accordingly. Meanwhile, the TD children have a mean age, height, and body mass of 9.021 (0.319) years, 1.316 (0.025) m, and 32.493 (1.991) kg, respectively.

For the laboratory layout, the HMGA Laboratory has an 8.3-meter wooden walkway with two force plates embedded at the center of the walkway as depicted in Fig. 1. Two digital video cameras which are installed on tripods were placed perpendicularly to one another for recording the walking trials from the side and front views as the subject walk along the walkway. In addition, all the infrared (IR) cameras of the Vicon System were wall-mounted approximately 2.5 meters above the floor level to ensure that the capability of the volume area to be captured at the center of the lab based on the length, width, and height of 4 meters, 3 meters, and 2 meters respectively. Meanwhile, the Kinect sensor is placed on a stand located 0.5 meters in height from the floor and is facing the walking direction, with a working range between 1.2 and 3.5 meters view from the camera at 57° horizontal wide and a vertical view of 43°. For this laboratory setup, the walking direction started at the point “A” and finished at the point marked as “B”.

A. Data Acquisition

For the data acquisition stage, a depth camera namely the Microsoft Kinect camera is used as a motion-sensing device for the markerless-based gait technique. The camera is equipped with a depth sensor, Red-Green-Blue (RGB) camera, microphones, and a tilt motor to provide full-body motion sensing for 20 primary body landmarks or coordinates. In addition, this device works at a frequency of 30 Hz. The IR depth sensor gives the Kinect its depth measuring capabilities. The emitter emits infrared light beams and the depth sensor reads the IR beams reflected to the sensor. The reflected beams are converted into depth information measuring the distance between the body and the sensor within its capture volume with a resolution of 640×480 to acquire the skeletal data. On the other hand, Vicon Motion System is used as a motion-sensing for the marker-based gait technique and also as the gold standard to establish the accuracy of the Kinect. Refer to Fig. 1, the Vicon system is equipped with eight optical cameras labeled as C1 – C8 (MX T-Series camera), wall-mounted, and operated at 100Hz. This Vicon system is also integrated with two force plates, a host computer is installed with the Vicon Nexus 1.8.5 software, and two digital video cameras namely DV1 and DV2.

TABLE I. SUBJECTS DISTRIBUTIONS

Items	Mean (SD)	
	ASD Group	TD Group
Gender	21 Male & 2 Female	14 Male & 16 Female
Age (year)	8.391 (0.396)	9.021 (0.319)
Height (m)	1.267 (0.028)	1.316 (0.025)
Weight (kg)	32.957 (3.485)	32.493 (1.991)
BMI (kg/m²)	19.509 (1.297)	18.206 (0.644)

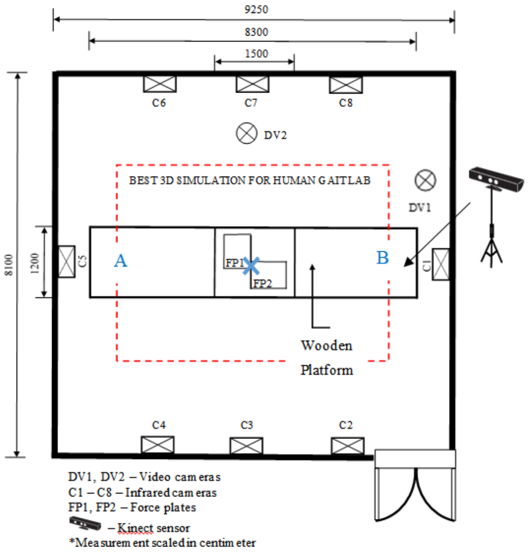


Fig. 1. Schematic Lab Layout.

B. Procedure

For the Vicon system, the subject's preparation was done before the experiment started. A total of 35 retro-reflective markers were attached to the subject‘s body based on the plug-in-gait full-body marker set that further produced the skeleton output or stick-figure model. Here, subjects need to perform a T-pose stand with the complete sets of markers so that the system can recognized and able to create the model based on the markers as in Fig. 2(a) [29]. On the other hand, a skeleton function was used to create the model from the Kinect device as shown in Fig. 2(b) [30].

Further, each subject performed walking trials while being concurrently monitored and captured using both the Kinect sensor and Vicon system. Refer to Fig. 2(c), for the walking task, subjects were required to walk freely barefooted with their comfortable normal speed to ensure their walking behavior is natural, without any assistant or walking aid. The walking task was carried out and repeated several times until all the ten successful trials per subject were completed. A successful trial is defined once the subject can successfully walk from point A to point B as in Fig. 1, without stopping or pausing during the walking trial or assisted by the caretakers. During the experimental, a walking trial is captured and saved simultaneously by both Kinect and Vicon systems.

C. Data Analysis

Data from the systems were screened and time-synchronized visually before data extraction. At this stage, a single gait cycle for the same walking trial session was acquired to represent each subject's walking pattern from both systems. For the data acquired via the Kinect sensor, the angles were derived using the cosine rule. As for the Vicon system, the output angles of the gait were extracted. Further, the kinematic angles from both systems were computed and analyzed.

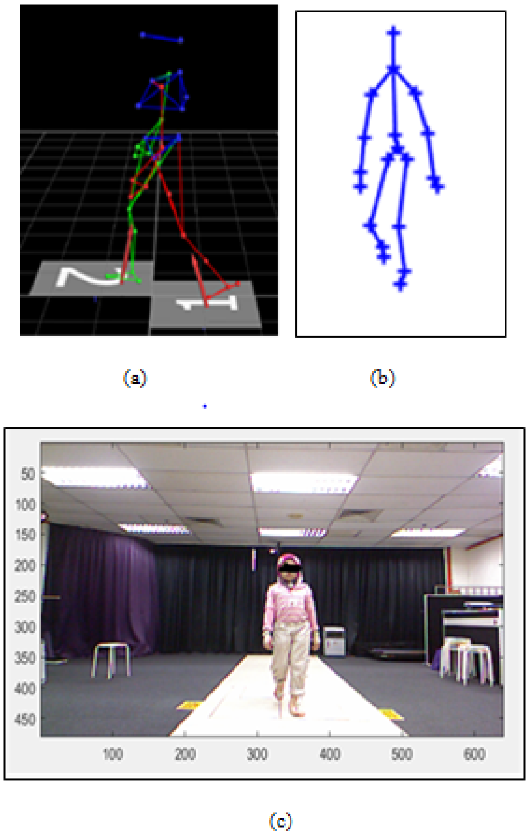


Fig. 2. Example of Stick-Figure Model from (a) Vicon System, (b) Kinect Depth Camera and (c) Subject Performed Walking Trial.

In order to analyze, evaluate and validate the walking gait among these subjects, the interpolation technique is used to standardize the number of frames that is fixed at 30 frames for all trials and further is expressed as the percentage of the gait cycle. Additionally, two types of normalization techniques are implemented before the evaluation of the gait data. Firstly, the intra-group normalization, x_{intra} was calculated to eliminate the influence of the subject‘s gait patterns which are different for each person and are expressed as in Equation (1) [31]. On the other hand, the inter-group normalization, x_{inter} was computed in minimizing the effect of gait patterns among the group as given in Equation (2) [31].

$$x_{intra} = \frac{x_i}{x_{\max(intra\ class)}} \qquad (1)$$

$$x_{inter} = \frac{x_i}{x_{\max(inter\ class)}} \qquad (2)$$

Further, the kinematic plots for the hip, knee, and ankle angles from both Kinect and Vicon systems are visualized based on the plot pattern. The *F*-statistic was used to assess the quality of variances by computing the ratio of variation between the sample means and variation within the sample means. Then, the Analysis of Variance (ANOVA) test set as the significant value, ρ is less than 0.05 ($\rho < 0.05$) with the null hypothesis as the mean values of the kinematic plots are similar for both systems was used to test any mean differences in all angles. Here, the mean differences indicated that the plot is not similar in terms of its form or magnitude.

IV. Experimental Results and Discussion

This section presents the results of the evaluation and validation of the gait data from both the gait techniques utilized in comparing the kinematic plots of the markerless-based gait technique that is from the Kinect sensor and the marker-based gait technique which is from the Vicon system. Fig. 3 showed the kinematic plots of the hip, knee, and ankle angles for the original, intra-group, and inter-group normalization gait data for both the ASD and TD children. The green color line represents the markerless method specifically the Kinect sensor while the magenta color line is for the Vicon system that is the marker-based technique. Further, the ANOVA results for all the kinematic plots of Fig. 3 are tabulated in Table II.

Firstly, the plots based on the original gait data similar shapes are visualized for the knee angles of the walking gait generated from both techniques in the ASD as well as the TD group. As for the plots for the hip angle, the plots are also similar however the amplitude values between the two gait techniques differ for both groups. The same goes for the ankle angle plots which are not similar using both techniques for the ASD and TD groups. Based on the original gait data, the differences in the knee angles and ankle angles of the kinematic plots are supported by the statistical results computed as tabulated in Table II. For the ASD group, the ANOVA value for *F* is 577.55 with ρ equals to 0.00 for the hip angles whilst the TD group with *F*=375.09. As for the ankle angles, the computed *F* values are 84.450 and 81.70 respectively for both the ASD and TD groups. These values resembled significant differences for the knee and ankle angles using the markerless and marker-based techniques for both the ASD and TD groups and matched results of the visualized plots. This showed that the hip and ankle angles owned large group means relative to the within-group variability as compared to the knee angle. Next, for the knee angles with *F*=1.443 and ρ is equal to 0.23 for the ASD group and with *F*=1.26 and ρ=0.23 which is larger than 0.05 resembled that there is no significant difference between both techniques for the ASD and TD groups. These values supported and proved that the visualization of both plots also showed similar patterns. Recall that this study is to prove that there are no significant differences amongst the kinematic features with the ρ-values must be greater than 0.05 since similar plots showed that the proposed markerless-based technique is similar to the gold standard which is the marker-based gait technique utilized.

Furthermore, the intra-group and inter-group normalization were computed to analyze and eliminate data redundancy thus maintains the same information that adheres to a common standard approach for datasets. All three angles of the hip, knee, and ankle of the kinematic plots are as plotted in Fig. 3 for both intra-group and inter-group normalization. As explained earlier, for the hip angle the magnitude gaps differ, however upon normalization, the differences between the magnitudes are significantly reduced especially towards the end of the gait cycle in both the ASD and TD groups. Next, the plots revealed that the knee flexion during the swing phase (60% to 100% of the gait cycle) has improved. However, there is a difference in flexion and extension that existed during initial contact (0% of the gait cycle) between the ASD and TD groups for the knee angle using the proposed markerless-based approach. As for the ankle angle plots, the normalization process failed to enhance or improve the plots. The only improvement based on the visualization of these plots is the similar shapes of the ankle dorsiflexion and ankle plantarflexion during the stance phase that is between 0% to 60% of the gait cycle and towards the end of the gait cycle. The ANOVA results for the intra-group and inter-group normalization showed that the mean of the hip, knee, and ankle angles for both gait techniques is similar for both groups, which means that there are no significant differences in the mean for the respective angles. The *F*-statistic revealed a small ratio for the hip, knee and ankle angles with *F*=0.00 respectively. Once again, the small *F*-value indicates that the group means are similar with minimal variability. Also, the ρ-value that evaluated the mean differences showed that there were no significant mean differences for the hip, knee, and ankle angles since the ρ-value was greater than 0.05.

TABLE II. ANOVA Results using the Marker-based Technique Versus the Proposed Markerless-based for both ASD and TD Groups using the Original Gait Features, Intra-group and Inter-Group Normalization

Item	Kinematic Gait Parameters	ASD		TD	
		F	**ρ**	***F***	**ρ**
Original Gait Data	Hip	577.55	0.00	375.09	0.00
	Knee	1.443	0.23	1.26	0.26
	Ankle	84.45	0.00	81.70	0.00
Intra-Group Normalization	Hip	0.00	1.00	0.00	1.00
	Knee	0.00	1.00	0.00	1.00
	Ankle	0.00	1.00	0.00	1.00
Inter-Group Normalization	Hip	0.00	1.00	0.00	1.00
	Knee	0.00	1.00	0.00	1.00
	Ankle	0.00	1.00	0.00	1.00

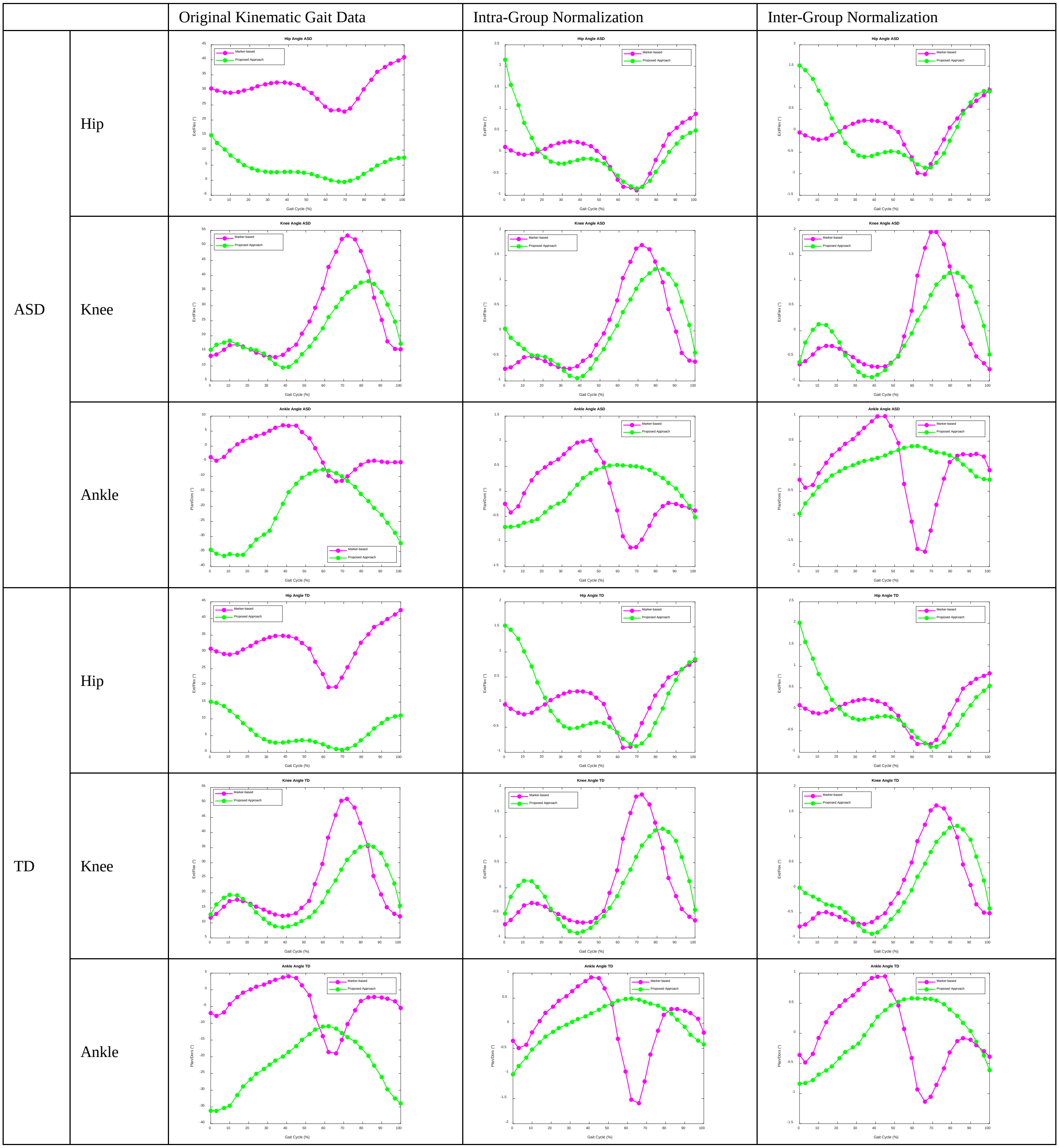


Fig. 3. Visualization of the Plot Pattern of the Kinematic Gait Angels using the Marker-based Technique (Magenta Color Line) Versus the Proposed Markerless-based Technique (Green Color Line) for both ASD and TD Groups using the Original Gait Features, Intra-Group, and Inter-Group Normalization.

VI. Conclusion

In conclusion, an evaluation and validation of gait data from the Vicon system versus the proposed markerless-based model are analyzed. Kinematic features namely the hip, knee, and ankle angles are evaluated based on visualization of the angles plots as further verified based on ANOVA. From the plots, the hip, and ankle angles showed differences between the two gait techniques for both ASD and TD groups. Further, the ANOVA test was performed to measure the significance of these angles between the marker-based and the proposed markerless approach.

The ANOVA results showed that there are significant differences for both the hip and ankle angles and these findings are in accordance based on the visualization of these plots. Further, the intra-group and inter-group normalizations are performed to minimize the data redundancy as well as maintaining similar information of the kinematic plots for both the marker-based and the proposed markerless technique. Upon normalization, the kinematic plots for the hip, knee, and ankle angles of both techniques showed similar shapes. With the ρ equals 1, hence these confirmed that there are no significant differences between these plots of gait features upon normalization. This showed that the proposed markerless-based gait technique is indeed suitable and potentially can be used as gait analysis for ASD children. The proposed markerless method is easy to set up, non-intrusive and portable as well. The next stage of work will include the use of the proposed markerless-based technique for Parkinson's disease (PD) or cerebral palsy (CP). This new proposed markerless gait approach could further assist in developing more suitable gait analysis intervention programs for autistic children and offer great research opportunities related to pathological gait.

Acknowledgment

This study is funded by PYPA Grant and Grant No: 600-IRMI/MyRA5/3/BESTARI (041/2017), Research Management Centre (RMC), Universiti Teknologi MARA, Shah Alam Selangor Malaysia.

References

[1] American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fifth ed. Arlington, VA: American Psychiatric Association, 2013.

[2] G. Esposito and P. Venuti, "Analysis of Toddlers' Gait after Six Months of Independent Walking to Identify Autism: A Preliminary Study 1," Perceptual and Motor Skills, vol. 106, pp. 259-269, 2008.

[3] M. Nobile, P. Perego, L. Piccinini, E. Mani, A. Rossi, M. Bellina, and M. Molteni, "Further evidence of complex motor dysfunction in drug naïve children with autism using automatic motion analysis of gait," Autism, vol. 15, pp. 263-283, 2011.

[4] C. Armitano, H. Bennett, J. Haegele, and S. Morrison, "Assessment of the gait-related acceleration patterns in adults with autism spectrum disorder," Gait & posture, vol. 75, pp. 155-162, 2020.

[5] O. Manicolo, M. Brotzmann, P. Hagmann-von Arx, A. Grob, and P. Weber, "Gait in children with infantile/atypical autism: Age-dependent decrease in gait variability and associations with motor skills," European Journal of Paediatric Neurology, vol. 23, pp. 117-125, 2019.

[6] H. L. Miller, P. M. Caçola, G. M. Sherrod, R. M. Patterson, and N. L. Bugnariu, "Children with Autism Spectrum Disorder, Developmental Coordination Disorder, and typical development differ in characteristics of dynamic postural control: A preliminary study," Gait & posture, vol. 67, pp. 9-11, 2019.

[7] K. Pope, J. Doll, A. Kyvelidou, H. Stessman, K. Nelson, and L. Jordan, "Clinician, caregiver and patient perspectives of the continuum of care for autism," Journal of Interprofessional Education & Practice, p. 100335, 2020.

[8] B.-O. Lim, D. O'Sullivan, B.-G. Choi, and M.-Y. Kim, "Comparative gait analysis between children with autism and age-matched controls: analysis with temporal-spatial and foot pressure variables," Journal of Physical Therapy Science, vol. 28, pp. 286-292, 2016.

[9] M. S. Nadeem, F. A. Al-Abbasi, I. Kazmi, B. N. Murtaza, M. A. Zamzami, M. A. Kamal, A. Arif, M. Afzal, and F. Anwar, "Multiple Risk Factors: A Challenge in the Management of Autism," Current Pharmaceutical Design, vol. 26, pp. 743-754, 2020.

[10] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. Farinella, "Computer vision for assistive technologies," Computer Vision and Image Understanding, vol. 154, pp. 1-15, 2017.

[11] S. Qiu, H. Wang, J. Li, H. Zhao, Z. Wang, J. Wang, Q. Wang, D. Plettemeier, M. Bärhold, and T. Bauer, "Towards wearable-inertial-sensor-based gait posture evaluation for subjects with unbalanced gaits," Sensors, vol. 20, p. 1193, 2020.

[12] J. D. Eggleston, J. R. Harry, and J. S. Dufek, "Lower extremity joint stiffness during walking distinguishes children with and without autism," Human Movement Science, vol. 62, pp. 25-33, 2018.

[13] C. Z. C. Hasan, R. Jailani, N. M. Tahir, I. M. Yassin, and Z. I. Rizman, "Automated classification of autism spectrum disorders gait patterns using discriminant analysis based on kinematic and kinetic gait features," Journal of Applied Environmental and Biological Sciences, vol. 7, pp. 150-156, 2017.

[14] J. D. Eggleston, J. R. Harry, R. A. Hickman, and J. S. Dufek, "Analysis of gait symmetry during over-ground walking in children with autism spectrum disorder," Gait & posture, vol. 55, pp. 162-166, 2017.

[15] K. Aminian, P. Robert, E. Jequier, and Y. Schutz, "Estimation of speed and incline of walking using neural network," Instrumentation and Measurement, IEEE Transactions on, vol. 44, pp. 743-746, 1995.

[16] D. A. Winter, Biomechanics and motor control of human movement: John Wiley & Sons, 2009.

[17] I. Rida, "Towards Human Body-Part Learning for Model-Free Gait Recognition," arXiv preprint arXiv:1904.01620, 2019.

[18] H. Ng, H.-L. Ton, W.-H. Tan, T. T.-V. Yap, P.-F. Chong, and J. Abdullah, "Human identification based on extracted gait features," International Journal of New Computer Architectures and their Applications (IJNCAA), vol. 1, pp. 358-370, 2011.

[19] J. A. Vilensky, A. R. Damasio, and R. G. Maurer, "Gait disturbances in patients with autistic behavior: a preliminary study," Archives of Neurology, vol. 38, pp. 646-649, 1981.

[20] N. K. Zakaria, N. M. Tahir, and R. Jailani, "ASD Children Gait Classification Based On Principal Component Analysis and Linear Discriminant Analysis," International Journal of Emerging Trends in Engineering Research, vol. 8, 2020.

[21] A. A. Al-Jubouri, I. H. Ali, and Y. Rajihy, "Gait and Full Body Movement Dataset of Autistic Children Classified by Rough Set Classifier," Journal of Physics: Conference Series, vol. 1818, p. 012201, 2021.

[22] M. Ebrahimi, M. Feghi, H. Moradi, M. Mirian, and H. Pouretemad, "Distinguishing tip-toe walking from normal walking using skeleton data gathered by 3D sensors," in Robotics and Mechatronics (ICROM), 2015 3rd RSI International Conference on, 2015, pp. 450-455.

[23] H. Moradi and I. Mohammad-Rezazadeh, "Recent Advances in Mechatronics Devices: Screening and Rehabilitation Devices for Autism Spectrum Disorder," in Advanced Mechatronics and MEMS Devices II, ed: Springer, 2017, pp. 283-296.

[24] T. B. Rodrigues, C. Ó. Catháin, D. Devine, K. Moran, N. E. O'Connor, and N. Murray, "An evaluation of a 3D multimodal marker-less motion analysis system," in Proceedings of the 10th ACM Multimedia Systems Conference, 2019, pp. 213-221.

[25] W. Z. W. Z. Abiddin, R. Jailani, A. R. Omar, and I. M. Yassin, "Development of MATLAB Kinect Skeletal Tracking System (MKSTS) for gait analysis," in Computer Applications & Industrial Electronics (ISCAIE), 2016 IEEE Symposium on, 2016, pp. 216-220.

[26] R. A. Clark, Y.-H. Pua, K. Fortin, C. Ritchie, K. E. Webster, L. Denehy, and A. L. Bryant, "Validity of the Microsoft Kinect for assessment of postural control," Gait & posture, vol. 36, pp. 372-377, 2012.

[27] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease," Gait & posture, vol. 39, pp. 1062-1068, 2014.

[28] I. Cocchi, G. Figari, N. Valeri, G. Paolini, U. Della Croce, A. Cereatti, E. Pantzar, A. Magnuson, and J. Riad, "A 2D markerless gait analysis protocol to estimate the sagittal joint kinematics of children with cerebral palsy," in 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT), 2019, pp. 192-196.

[29] Vicon Plug-in Gait Product Guide: Foundation Notes: Vicon Motion Systems Limited, 2010.

[30] A. Procházka, O. Vyšata, M. Vališ, and M. Yadollahi, "The MS kinect use for 3d modelling and gait analysis in the Matlab environment," Technical Computing Prague, vol. 270, 2013.

[31] N. M. Tahir and H. H. Manap, "Parkinson Disease Gait Classification based on Machine Learning Approach," Journal of Applied Sciences, vol. 12 (2), pp. 180-185, 2012.

Increasing the Steganographic Resistance of the LSB Data Hide Algorithm

A. Y. Buchaev[1], A. G. Mustafaev[2], V.S. Galyaev[3], A. M. Bagandov[4]
Department of Information Technologies and Management
Dagestan State University of National Economy
Makhachkala, Russian Federation

***Abstract*—The robustness of the security algorithm is one of the most important properties that determines how difficult it is to break it. Increasing the robustness of the algorithm directly affects the degree of secrecy when it is used for confidential transmission. The paper analyzes the steganographic algorithm Least Significant Bit, represents a method of counteracting the algorithm of the "visual attack" and statistical methods used against stego-containers generated using the LSB algorithm. To prove the increase in resistance, the study used the PSNR index, Chi-square test. The proposed technique involves the use of a uniform distribution and compression method. The paper presents the results of computer experiments demonstrating the effectiveness of the proposed technique.**

Keywords—Steganography; steganalysis; visual attack; least significant bit

I. Introduction

Steganography is the science of methods of transferring information or storing it, in which the fact of transfer or storage is hidden [1]. Currently, such directions as digital steganography (hiding information in digital objects) and network steganography (hiding information using the features of network protocols) are developing. Modern steganographic systems use steganography and cryptography algorithms together in order to not only encrypt and protect a message [2-4], but also to transmit it secretly.

Some steganographic algorithms have become widespread, such as algorithms for applying digital watermarks, which used to embed an image into an image and provide protection against illegal copying or dissemination of information [5]. The most famous and simple steganography algorithms have obvious drawbacks, for example, replacing the color of the hidden message with the background color, such a substitution is easy to notice and reveal the hidden data [6-7]. The book cipher (or its modifications, for example, the book cipher of Aeneas), in which each character of the secret message is replaced by a pointer (for example, the number of a row, column or table), has several significant disadvantages: transmission of small volumes of a secret message; storage and transmission of the key, which can be used to collect a "scattered" message from the so-called stego-container, weak degree of security [8]. Methods of hiding secret information in special fields of attributes of files of various formats or in service fields of network packets are also popular, but these methods have a significant limitation on the amount of information transmitted per unit of time [9-10]. Along with these methods, steganography includes various algorithms are based on distortion (introduction of changes into the structure of a digital object), statistical methods of concealment [11] and structural methods.

Specialized attacks are carried out on steganographic algorithms, the main purpose of which is to reveal the presence of an embedded secret message. The statistical method has become widespread, which makes it possible to determine the characteristic distortions both in the file structure and in the semantic information of a digital object [12]. Attacks are usually directed against specific vulnerabilities of a particular steganographic algorithm [13]. When hiding information in the service part of files or transmitted packets, an attacker can compare with reference values or empty containers to identify potential corruption [14-15]. In modern steganalysis, separate areas of investigation to identify hidden messages have been formed.

Reliable masking algorithms are being developing to counter attacks, for example, one of them hides secret information during a "handshake" when using the TCP data transfer protocol in a response packet [16], but this method is describing theoretically, is implementing only as part of the study and has small bandwidth.

Within the framework of this work, a technique makes it possible to reduce the efficiency of steganalysis methods of the "visual attack" type. The Least Significant Bit (LSB) method [17-18] was chosen as the most illustrative example of a steganographic algorithm. LSB is an efficient algorithm used to embed information into container files [19-20]. A hypothesis was put forward, according to which the algorithm can become more resistant to the mentioned types of attacks if the modified bytes are uniformly distributed in the container image. However, when using a uniform distribution of a large number of bytes, a large number of zones with a high density of modified bytes will inevitably appear [21-22]. Lossless compression methods are using to reduce the number of modified bytes.

II. Steganographic Algorithm least Significant Bit

LSB (Least Significant Bit) is a method of embedding a secret message into an image. The algorithm includes the following steps:

- Conversion of secret message into a binary code, followed by splitting into separate bits or into blocks of two bits.

- Replacing the last bit or two bits in the bytes of the container image with the corresponding number of bits of the transmitted secret message.

For example, the binary form of the secret message looks like this: 101101. We form groups of two bits: 10, 11, 01. We get three groups, the number of modified image bytes is equal to the number of groups, and therefore, the last two bits in the three bytes will be replaced in the container image. Suppose the bytes of the container image look like this: ... 10111010 11010001 10000011 ..., the last two bits in the given bytes will be replaced with bits from the previously received groups. The transformation will look like this: the set 10111010 11010001 10000011 goes to 10111010 11010011 10000001. This transformation leads to changes that a person does not perceive during normal visual observation.

Such conversions can be carried out with graphic file formats that do not compress data (BMP, PNG), otherwise the hidden information will be lost. The algorithm described above is also applicable to the WAV format [23], similar conversions can be performed with video clips and sound files.

For containers, either a unique digital object is often used, or images and sound files that are distributed over the network in a variety of ways. If you select objects that are in the public domain in the most common variant, then an attacker will be able to compare the source file with the container and reveal secret information.

III. Methods of Detecting the Fact of Transfer of Information

There are a number of attacks on image containers that reveal the presence of a hidden message. One of the common attack methods is "visual attack" [24-25]. The essence of the attack is the formation of new images from the low-order bits of the original image with amplification of values (maximization), an example of an image without hiding is shown in Fig. 1(a). At the same time, areas with a high data density usually appear on the images that are generated on the basis of stego-container files with an existing hidden message, as can be seen in Fig. 1(b). The embedded message is located in the areas of visual distortion.

When using a small container image (242 976 bytes), the result of a visual attack is more apparent. For example, when the container is filled by 30% (Fig. 2(b)), one can see a characteristic difference from the result of a visual attack on the original image (Fig. 2(a)).

To identify hidden information usually use method that named the Chi-square statistical test [26]. In steganalysis, the use of the criterion, which based on the fact that two neighboring colors (colors are adjacent if they are different only in the least significant bits) differ significantly in the number of points relative to the untransformed image [27]. Fig. 3(a) shows a chi-square rendering of an original container image. The clear difference between the rendering results of the original image and the image processed with classical LSB is shown in Fig. 3(b).

Because both steganalytical methods are based on identifying dependencies throughout the digital object, the research was aimed at finding modifications to the algorithm that would allow disguising the introduction of distortions when placing a steganographic message under random distortions or noise. To do this, it was necessary to achieve two parameters of the algorithm: to increase the coverage of the involved parts of the container, but at the same time to reduce the total number of modified bits. To increase the coverage of the involved areas of the container, a uniform distribution of modified bytes in the container image was used (Fig. 3(c)). This solution only partially improves the overall picture when performing visual attacks, as well as Chi-square analysis. To reduce the number of modified bytes, it was decided to use lossless compression methods.

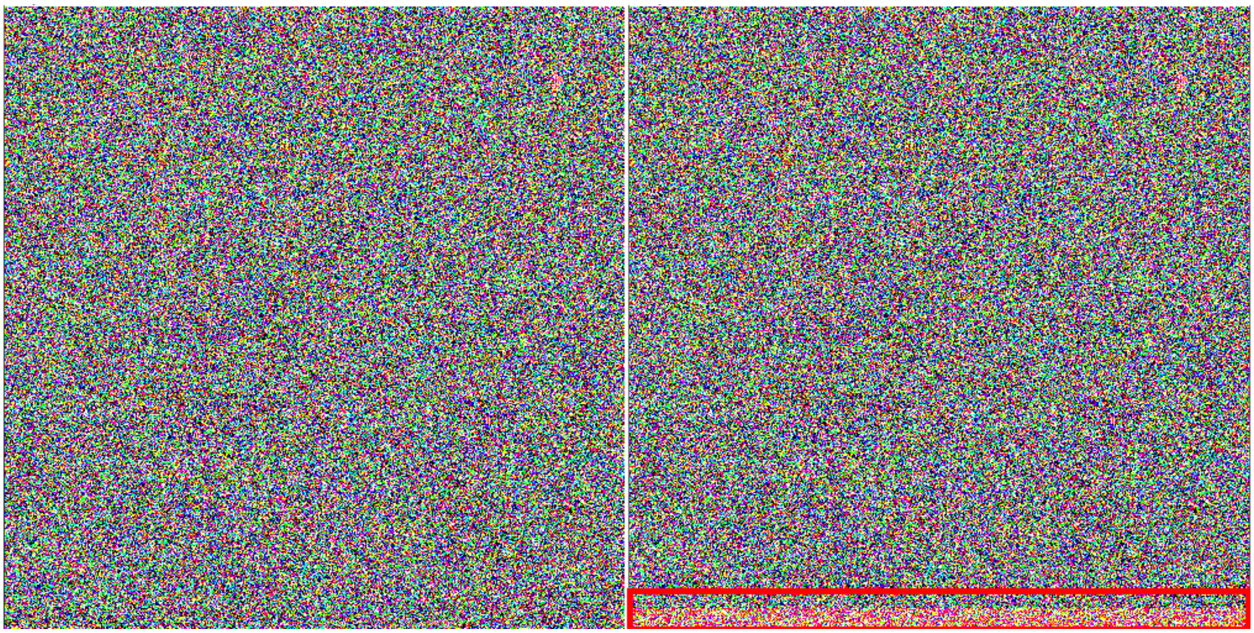

(a) (b)

Fig. 1. The Result of a Visual Attack on Images: (a) – Original Image; (b) – Image Filled with Standard LSB Method.

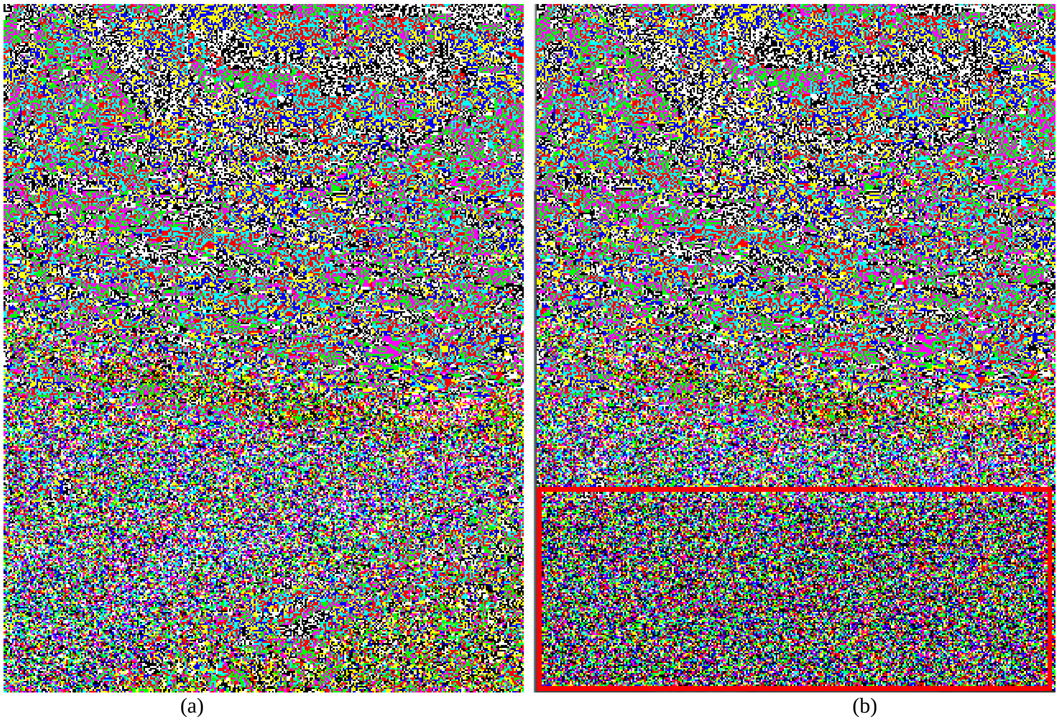
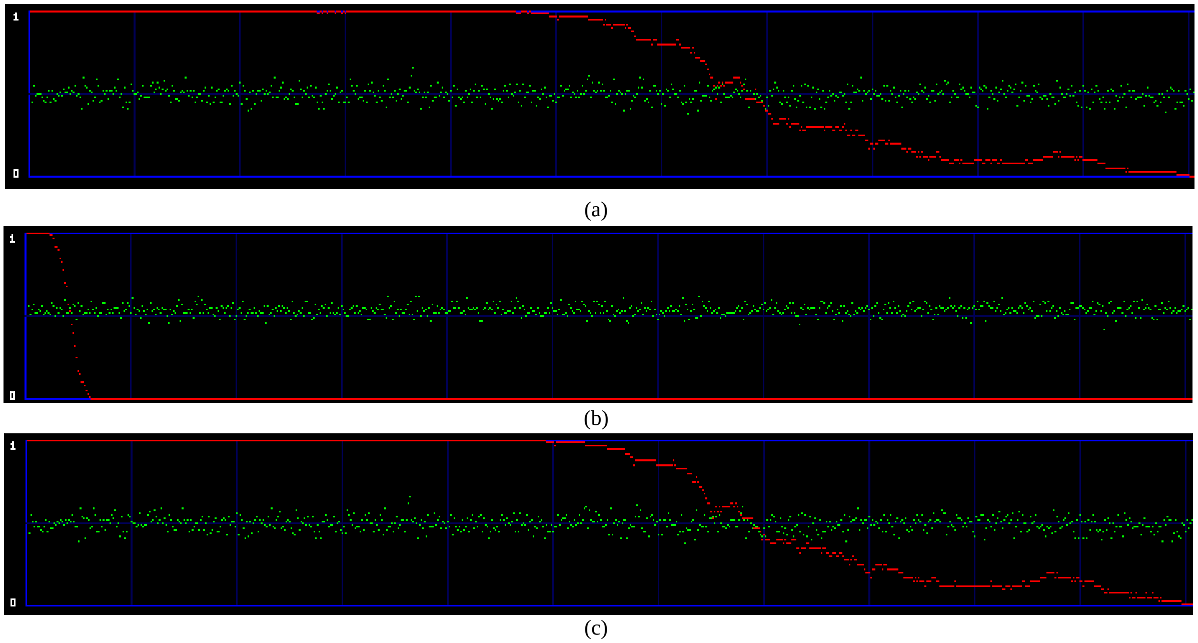

(a) (b)

Fig. 2. The Result of a Visual Attack on Small Image: (a) – Original Image; (b) – Image Filled with Standard LSB Method.



(a)
(b)
(c)

Fig. 3. Chi-Square Visualization: (a) – The Result of the Analysis of the Original Image; (b) – The Result of the Analysis of the Standard LSB; (c) - is the Result of LSB Analysis using a Substitution Table.

Studies have shown that for small messages, the size of the container has practically no effect on the result of the attacks described above. In other words, the attacks themselves are ineffective in finding low-capacity messages. The study of the results of the modified steganographic algorithm was tested on examples of hiding medium and large messages in the corresponding containers.

IV. Analysis of Compression Algorithms

To solve the problem of compression, several methods, which perform lossless data compression [28], have been considered. All the considered compression algorithms were implemented in the C ++ programming language and tested on several samples of medium and large texts.

Run-length encoding (RLE) [29]: An easy-to-implement algorithm with the best, average and worst compression ratios equal to 1/32, 1/2, 2/1. Test #1 results: text 608 characters long and 1115 bytes in size was converted to 1135 bytes. Test #2 results: a text of 3040 characters long and 5575 bytes in size was converted to 5669 bytes. Test #3 results: text 9120 characters long and 16725 bytes in size was converted to 17028 bytes. The increase in volume is due to the construction of key-value pairs. With a small number of repetitive sequences, the size of the original file will grow. This algorithm shows satisfactory results with texts containing repetitive sequences.

Compression of information based on binary coding trees (Huffman compression) [30]. Test #1 results: text 608 characters long and 1115 bytes in size was converted to 1321 bytes. Test #2 results: a text of 3040 characters long and 5575 bytes in size was converted to 3499 bytes. Test #3 results: text 9120 characters long and 16725 bytes in size were converted to 8944 bytes. With an increase in the volume of text, the compression ratio increases, but the work of the algorithm with medium-sized texts does not give positive results. This specificity is due to the implementation of the algorithm based on binary trees.

Sliding window compression (LZ77) [31]. Test #1 results: text 608 characters long and 1115 bytes in size were converted to 704 bytes. Test #2 results: text 3040 characters long and 5575 bytes in size were converted to 792 bytes. Test #3 results: text 9120 characters long and 16725 bytes in size were converted to 1009 bytes.

According to the results of the analysis (Table I) and testing of the methods, the LZ77 method is the most optimal for solving the problem of compressing medium and large texts.

TABLE I. COMPARISON OF THE RESULTS OF THE COMPRESSION ALGORITHMS

Compression algorithms	Text No. 1 608 characters, 1115 bytes	Text No. 2 3040 characters, 5575 bytes	Text No. 3 9120 characters, 16725 bytes
RLE	1135 bytes	5669 bytes	17028 bytes
Huffman compression	1321 bytes	3499 bytes	8944 bytes
LZ77	704 bytes	792 bytes	1009 bytes

V. KEY-GENERATED REPLACEMENT TABLE

The use of a uniform distribution of compressed data does not sufficiently improve the robustness of the steganographic algorithm, since any steganographic algorithm has the property of symmetry, in other words, it is possible to extract hidden information performing the reverse actions of hiding. To solve this problem, an analogue of the secret key was introduced. A byte replacement table in the container image is formed based on this key. The number of elements in the table is equal to the number of groups of two bits formed from the binary representation of the compressed bytes of the secret message. The replacement table forms a uniform distribution of the modified bytes in the container image (Fig. 4(b), 5(b)), due to which the embedded compressed message is similar to the noise that occurs in the image when it is projected in different channels. The important thing is that the receiver and the sender generate the table independently of each other, using only a shared secret key that both parties know. In this case, the replacement table for both sides is generated the same. This property is very important because transferring and storing such a large table is difficult, and the key is easy to use.

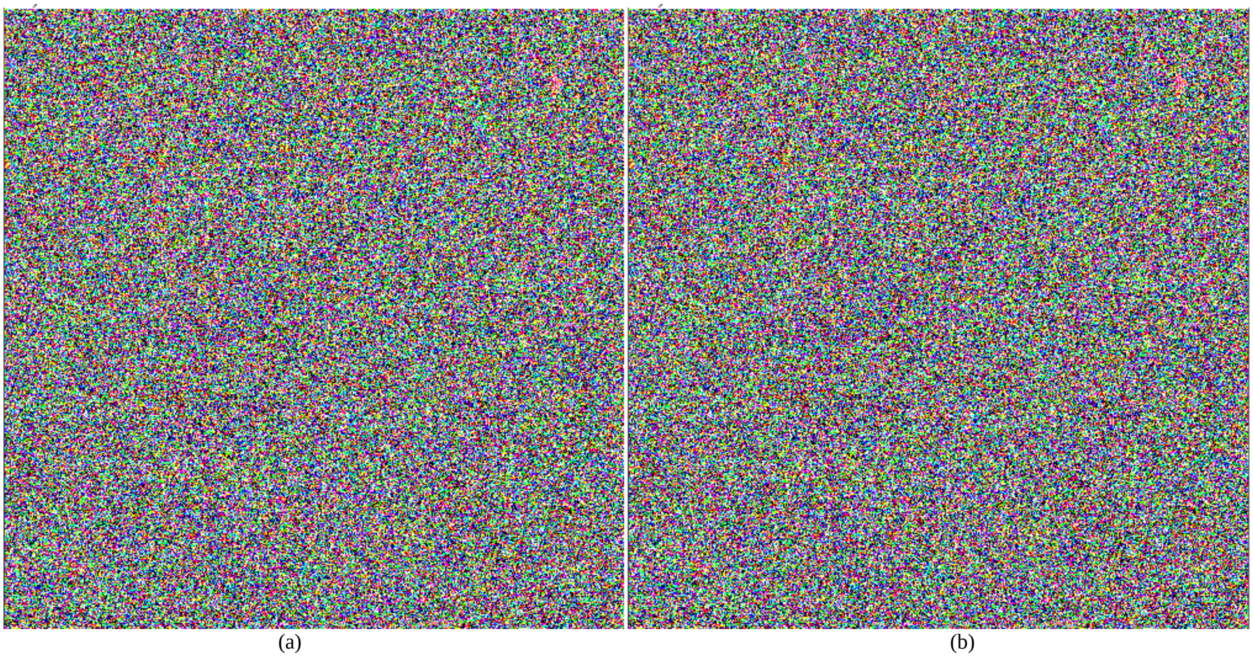


Fig. 4. Visual Attack on Images: (a) – Original Image; (b) – Image Filled with Improved LSB Method.

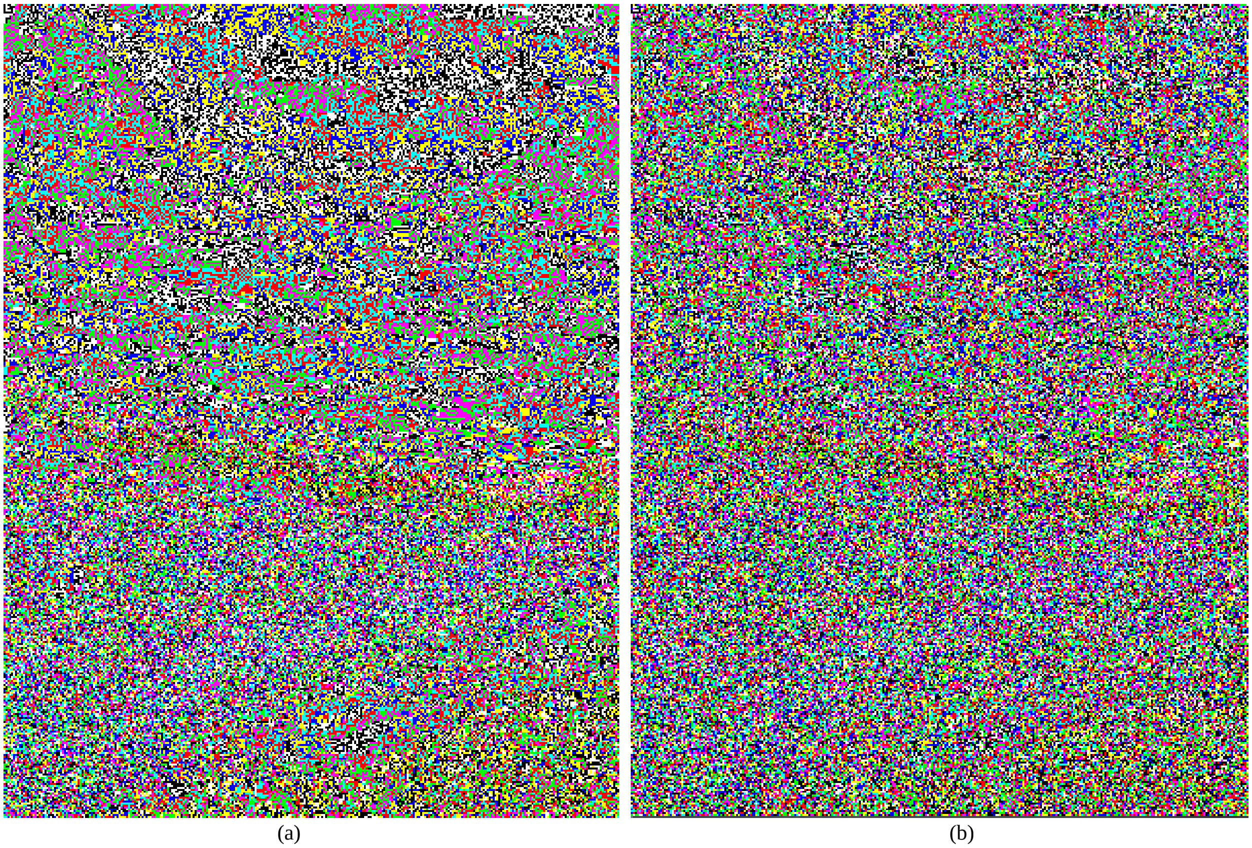


Fig. 5. Visual Attack on Small Images: (a) – Original Image; (b) – Image Filled with Improved LSB Method.

VI. Results of a Computer Experiment

A program that implements an improved version of the LSB hiding method was written in the Python programming language as part of the study. The results of testing this program are shown below. For comparison, the result of executing the original LSB algorithm will be shown. The comparison will be carried out on the same compressed data (the original text is 150 474 characters long, the text size is 150 948 bytes; the compressed text is 58 364 bytes).

VII. The Effectiveness of the Developed Method

For an objective assessment of the improvement of the algorithm, it was decided to use the PSNR (Peak Signal to Noise Ratio) index [32]. The index value indicates the similarity between the two images, and therefore, the higher the value, the more the similarity. When calculating the index, you need to calculate the mean square error (MSE) between the images.

$$MSE = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}(X_i - Y_i)^2$$

X and Y are equal to the values of the original image and the container image, N is the number of pixels in the image.

$$PSNR = 20\log_{10}\frac{MAX}{MSE}$$

MAX is the maximum value that the pixel color can take, equal to 255.

During the study, the index between the original image (Fig. 1(a)) and the image processed with the standard LSB (Fig. 1b) was calculated, which is 57.72449648521831. The index between the original image and the image processed by the improved LSB algorithm (Fig. 4(b)) is 61.39096278660813, which indicates a greater similarity with the original image. There is a slight improvement when calculating the index between the smaller images. So the index between the original image (Fig. 2(a)) and the image processed with the standard LSB (Fig. 2(b)) is - 49.47831069914614. The index between the original image and the image processed by the improved LSB algorithm (Fig. 5(b)) is 50.5708622069228. A brief comparison of the results of the computer experiment is presented in the Table II.

TABLE II. Comparison of PSNR Indices

	Standard LSB algorithm	**Improved LSB algorithm**
Large volume of hidden message, medium container volume	49.47831069914614	50. 5708622069228
Large volume of hidden message, large container volume	57.72449648521831	61.39096278660813

VIII. Conclusion

The study compared two aspects: PSNR index and security. Based on the results of calculations and comparison of visualizations of various types of attacks on images processed by the standard LSB method and images processed by the improved LSB method, we can conclude that the use of a substitution table with a uniform distribution gives a structural result similar to the original container image. Distortion areas are minimized, and fragments with a high data density have disappeared, from which it is possible to calculate the fact of information transfer, an analogue of a secret key has been added, without which an attacker will not be able to extract useful information from the container image. The PSNR index when using the improved algorithm increases by ~ 5.9% in the best case and by ~ 2.16% in the worst case. In addition, visualization of the result of the improved LSB algorithm by the Chi-square criterion also indicates an increase in steganographic resistance. Computer experiments have shown that the improved LSB algorithm is more resistant to visual attacks and the use of statistical analysis methods.

The improved algorithm can be applied to process a large number of images, for example, the covert transmission of information in social networks and services for storing and transmitting images.

In the future it will be analyzed the applicability of various distributions to increase the PSNR index and adapt the algorithm to other data formats, including video file formats.

References

[1] N.F. Johnson and S. Jajodia, "Exploring Steganography: Seeing the Unseen", Computer, vol. 31, no. 2, pp. 26-34, 1998.

[2] P.P. Aung and T.M. Naing, "A novel secure combination technique of steganography and cryptography", International Journal of Information Technology Modeling and Computing (IJITMC), vol. 2, no. 1, pp. 55-62, 2014.

[3] B. L. Sirisha, S. S. Kumar and B. C. Mohan, "Steganography based information security with high embedding capacity," 2015 National Conference on Recent Advances in Electronics & Computer Engineering (RAECE), Roorkee, India, 2015, pp. 17-21, doi: 10.1109/RAECE.2015.7510218.

[4] A. Shamir, "How to share a secret", Communication ACM, vol. 22, no. 11, pp. 614-613, 1997.

[5] J. Bloom, I. Cox, J. Fridrich, T. Kalker and M. Miller, Digital watermarking and steganography, San Francisco, CA, USA:Morgan Kaufmann Publishers, Inc., 2007.

[6] P. Johri, A. Mishra, S. Das and A. Kumar, "Survey on steganography methods (text, image, audio, video, protocol and network steganography)," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 2906-2909.

[7] N. Subramanian, O. Elharrouss, S. Al-Maadeed and A. Bouridane, "Image Steganography: A Review of the Recent Advances," in IEEE Access, vol. 9, pp. 23409-23423, 2021, doi: 10.1109/ACCESS.2021.3053998.

[8] Changda Wang; Shiguang Ju (2008). "Book Cipher with Infinite Key Space". 2008 International Symposium on Information Science and Engineering. p. 456. doi:10.1109/ISISE.2008.273. ISBN 978-0-7695-3494-7. S2CID 15768123.

[9] A. Kuznetsov, K. Shekhanin, A. Kolhatin, I. Mikheev and I. Belozertsev, "Hiding data in the structure of the FAT family file system," 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kiev, 2018, pp. 337-342, doi: 10.1109/DESSERT.2018.8409155.

[10] K. Shekhanin, A. Kolhatin, K. Kuznetsova and S. Kavun, "Steganographic hiding information in a file system structure," 2018 International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), Odessa, Ukraine, 2018, pp. 1-6, doi: 10.1109/UkrMiCo43733.2018.9047551.

[11] W. Bender, D. Gruhl, N. Morimoto and A. Lu, "Techniques for data hiding", IBM Syst. J., vol. 35, pp. 313-336, 1996.

[12] C. Zhi-li, H. Liu-sheng, Y. Zhen-shan, L. Ling-jun, and Y. Wei, "A Statistical Algorithm for Linguistic Steganography Detection Based on Distribution of Words," in Availability, Reliability and Security, 2008. ARES 08. Third International Conference on, pp. 558-563, 2008.

[13] Fridrich, M. Goljan and R. Du, "Detecting LSB steganography in color and gray-scale images", IEEE Multimedia, vol. 8, no. 4, pp. 22-28, 2001.

[14] T. Sharp, "An implementation of key-based digital signal steganography", Proc. of the 4th Information Hiding Workshop, pp. 13-26, 2001.

[15] F. Petitcolas, R. Anderson and M. G. Kuhn, "Information hiding – a survey", Proceedings of the IEEE, vol. 87, no. 7, pp. 1062-1078, 1999.

[16] Mazurczyk, Wojciech & Smolarczyk, Milosz & Szczypiorski, Krzysztof. (2011). Retransmission steganography and its detection. Soft Comput.. 15. 505-515. 10.1007/s00500-009-0530-1.

[17] Deepesh Rawat Vijaya Bhandari "Steganography Technique for Hiding Text Information in Color Image using Improved LSB Method" International Journal of Computer Applications Vol. 67 No. 1 April 2013 pp. 22-25.

[18] Shailender Gupta Ankur Goyal Bharat Bhushan " Information Hiding Using Least Significant Bit Steganography and Cryptography I.J. Modern Education and Computer Science 2012 Vol. 6 pp. 27-34.

[19] K. Thangadurai and G. Sudha Devi, "An analysis of LSB based image steganography techniques," 2014 International Conference on Computer Communication and Informatics, Coimbatore, India, 2014, pp. 1-4, doi: 10.1109/ICCCI.2014.6921751.

[20] V. Verma, Poonam and R. Chawla, "An enhanced Least Significant Bit steganography method using midpoint circle approach," 2014 International Conference on Communication and Signal Processing, Melmaruvathur, India, 2014, pp. 105-108, doi: 10.1109/ICCSP.2014.6949808.

[21] N. M. Abdali and Z. M. Hussain, "Reference-free Detection of LSB Steganography Using Histogram Analysis," 2020 30th International Telecommunication Networks and Applications Conference (ITNAC), Melbourne, VIC, Australia, 2020, pp. 1-7, doi: 10.1109/ITNAC50341.2020.9315037.

[22] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234-239, doi: 10.1109/WIFS.2012.6412655.

[23] Q. Liu, A. H. Sung and M. Qiao, "Detecting information-hiding in WAV audios," 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 2008, pp. 1-4, doi: 10.1109/ICPR.2008.4761650.

[24] R. Chandramouli and Nasir Memon, "Analysis of LSB Based Image Steganography Techniques", IEEE, no. 1019–1022, February 2001.

[25] A. Arora, M. P. Singh, P. Thakral and N. Jarwal, "Image steganography using enhanced LSB substitution technique," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2016, pp. 386-389, doi: 10.1109/PDGC.2016.7913225.

[26] M. Göl and A. Abur, "A modified Chi-Squares test for improved bad data detection," 2015 IEEE Eindhoven PowerTech, Eindhoven, Netherlands, 2015, pp. 1-5, doi: 10.1109/PTC.2015.7232283.

[27] Stanley, C.A., "Pairs of Values and the Chi-squared Attack", Department of Mathematics, Iowa State University (2005).

[28] R. J. van der Vleuten, "Low-complexity lossless and fine-granularity scalable near-lossless compression of color images," Proceedings DCC 2002. Data Compression Conference, Snowbird, UT, USA, 2002, pp. 477-, doi: 10.1109/DCC.2002.1000020.

[29] D. S. Bhadane and S. Y. Kanawade, "Comparative study of RLE & K-RLE compression and decompression in WSN," 2016 3rd International

Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2016, pp. 1-5, doi: 10.1109/ICACCS.2016.7586319.

[30] B. Ergude, L. Weisheng, F. Dongrui and M. Xiaoyu, "A Study and Implementation of the Huffman Algorithm Based on Condensed Huffman Table," 2008 International Conference on Computer Science and Software Engineering, Wuhan, China, 2008, pp. 42-45, doi: 10.1109/CSSE.2008.1432.

[31] C. Fraser, "An instruction for direct interpretation of LZ77-compressed programs", Technical report MSR- TR-2002-90, 2002.

[32] K. Joshi, R. Yadav and S. Allwadhi, "PSNR and MSE based investigation of LSB," 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), New Delhi, India, 2016, pp. 280-285, doi: 10.1109/ICCTICT.2016.7514593.

Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework

V. Laxmi Narasamma[1], Dr. M. Sreedevi[2]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, Andhra Pradesh, India-522502

***Abstract*—In recent years, the sentiment analysis using Twitter data is the most prevalent theme in Natural Language Processing (NLP). However, the existing sentiment analysis approaches are having lower performance and accuracy for classification due to the inadequate labeled data and failure to analyze the complex sentences. So, this research develops the novel hybrid machine learning model as Catboost Recurrent Neural Framework (CRNF) with an error pruning mechanism to analyze the Twitter data based on user opinion. Initially, the twitter-based dataset is collected that tweets based on the coronavirus COVID-19 vaccine, which are pre-processed and trained to the system. Furthermore, the proposed CRNF model classifies the sentiments as positive, negative, or neutral. Moreover, the process of sentiment analysis is done through Python and the parameters are calculated. Finally, the attained results in the performance parameters like precision, recall, accuracy and error rate are validated with existing methods.**

Keywords—Natural language processing; sentiment analysis; twitter data; Catboost; recurrent neural network

I. Introduction

In recent, several Artificial Intelligence (AI) [1] techniques are worn in NLP for many purposes like, sentiment analysis, question and answering system and so on [2]. The major reason of using NLP in big data is to reduce the time complexity. Moreover, the big data is applicable in all online application [3], so to handle the big data is the great deal. In this, the sentiment analysis is a key topic to evaluate sentiment values in customer suggestion in online application [4].Thus the sentiment analysis are processed in three levels that are document, sentence and feature level. Here the advanced level of sentiment analysis is feature level [5], which is proved in all research implementation, because it achieved high accuracy than document as well as sentence level [6]. In that, one of the broad social networking sites is twitter, a person uses the twitter for short message communication called tweets [7]. Twitter is defined as the online platform where publics can develop the messages, post, read, and update the text, which is called tweets. Moreover, the sentiment investigation based on the twitter statistics is mentioned as the scientific study of the tweets semantic parts. Subsequently, the sentiment analysis is the method of attaining data from numerous sources that are classified based on the sentiments. Generally, the tweets are reflecting the opinion from public based on the particular data about product, or any topic.

The public opinion is normally categorized into positive, negative and neutral tweets. However, the categorizations of tweets are very difficult for large quantity of data. In addition, if any one continuously following your tweets then your message is liked or attracted by the particular person [8], the twitter analysis is shown in Fig. 1. Even it has lot of facilities, the analyses of data in twitter is challenging task because of large volume of data [9]. This reason turned the interest of researchers towards this area [10]. Thus several researchers found much solution but it is not applicable for long time due to data complexity [11]. Also, in NLP text summarization is one of the schemes to identify the uniqueness of each document [12]. So, in NLP text summarization frame work is elaborated in better way: several machine learning techniques and vector based word embedding models were studied for the better classification [13, 30]. But for the complicated data these approaches are misbehave because the error removing model is not available in all machine learning model [14].

So, it failed to prune the error, this cause the difficulties to specify the sentiment value of data. The computer does not know the people language [15], so to make the human machine interaction machine learning is the advanced topic. In addition, the data can train the system in the form of 0 or 1 [16]. Because the machine only knows the binary value 0's and 1's, so the classification of sentiment value is in the form of decision making [17]. The sentiment analysis using the large quantity of data is done through the machine learning approaches [18]. Several machine learning approaches are found but still the issues are not end [19]. Thus, the present research work aimed to develop an efficient machine learning model to classify tweets data based on their sentiment values.

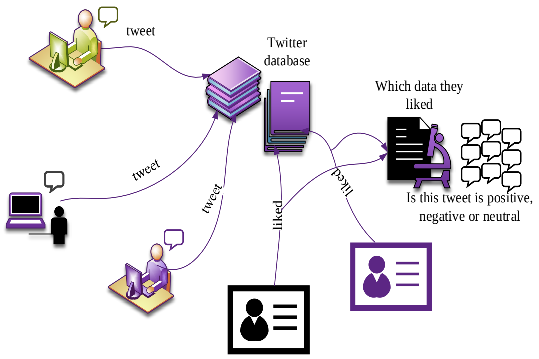


Fig. 1. Twitter Data Analysis.

This research work is organised as follows. The recent literature works based on sentiment analysis using twitter data is detailed in section 2. Also, the system model and problem statement is mentioned in section 3. Moreover, the developed methodology is elaborated in section 4 and the attained outcome of the proposed work is declared in section 5. Thus, section 6 detailed the end of the entire research.

II. Related Works

Several literature works correlated to the twitter data analysis is summarized below and detailed in Table I.

Ruz et al [21] introduced the bayes aspect manner to produce high real network. When comparing to the random forests and vector support in machine it gives competitive sentiment prediction result. However, approach cannot able to differentiate it behave in Spanish or English in RF and SVM. Finally conclude that, this network also allow to determine relation in the words, historically it gives interested quality data and catch socially the main headline of the dynamic act, accuracy in result and also reduce the exposure of misinformation.

Prediction of visiting next location using machine leaning by utilizing twitter information developed ensemble classified approach (ESA) has proposed by Kumar et al [22]. Moreover, this proposed work is utilized the twitter data for predicting the next visiting location of the user. Also, the developed ESA model attained the outcome of the prediction based on various classification models. For this prediction model, the voting technique is adopted to enhance the accurate sentiment calculation. This approach predicts accurate result with high desirable but it lack in security.

In recent, to assign a text for an emotion in classes automatically based on soft classified approach Hasan et al [23] developed a learning framework. That includes two tasks i.e., online and offline task. The result shows that the 90% of correct emotion of text can be created for real time. Finally, it gives best performance comparing to other approaches and also it doesn’t depend on other system. However, it attained high error rate.

TABLE I. Summary of Literature Survey

Author and year	Technique	Merits	Demerits
Ruz et al [21], 2020	Bayes aspect	Better information recall	It takes more time
Kumar et al [22], 2019	Ensemble classification approach	Accurate prediction	It attained Less privacy rate.
Hasan et al [23], 2019	Soft classification approach	Maximum Probability	High error rate
Barker, J. L. P., and Christopher JA Macleod [24], 2019	Prototype social geodata	Awareness during flood	Some time it lack in signal to predict the rain fall rate
Li al [25], 2019	patent analysis and twitter data mining	It required less time to process the mechanism	More complex

Barker, J. L. P., and Christopher JA Macleod [24] created prototype based social geo data from twitter to make the people aware from flood or huge disaster. Also, the decision mechanism is used to specify the rain fall and flood based data. Also, this model establishes the sentiment analysis using the twitter data based on pipeline extract tweets that involves 420000 tweets. Moreover, this supports a people lot to get aware about flooding.

Analysing the data is important in big data, Li et al [25] proposed patent analysis to determine the trends change in perovskite solar tech, and to identify response, expectation and sense is monitor using information from twitter mining. Finally, the comparison is made to identify the development of trends how the twitter users interested, this offer better in understanding and also it helps to find the development of trends in future but it may weak in capturing signals.

The key metrics of the proposed model is mentioned as follows,

- Initially, the twitter data based on the user opinion about COVID-19 is collected and trained to the system.
- Moreover, the novel CatBoost Recurrent Neural Framework (CRNF) is developed for analysing the sentiment value of twitter data.
- Subsequently, the developed CRNF model is utilized to remove the error while removing the leaf node layer that has increased the classification rate.
- Thus, the proposed approach effectively classifies the sentiments as positive, negative, or neutral.
- Additionally, the performance metrics such as recall, F-measure, accuracy, precision, and error rate are calculated and validated using existing approaches.

III. System Model and Problem Defination

Normally, the sentiment analysis or data analysis in natural language processing is done over big data dataset such as Facebook, twitter, etc. Moreover, sentiment analysis for large volume of data is some more difficult as because of its complexity and part of speech classification [20]. In addition, the sentence which contains positive words may also end with negative sentence. Thus the opinion or sentiment classification is one of the important tasks in NLP, which is mostly helpful for online service because the success of online business is based up on the customer review. Moreover, the process of sentiment analysis using twitter data is explained in Fig. 2.

Also to predict the uniqueness of each sentence, thus the classification of sentiment measure is more important. This motivate this research to find the scientific solution to enhance twitter data analytics using sentiment analysis in Natural Language Processing to reduce all kinds of issues

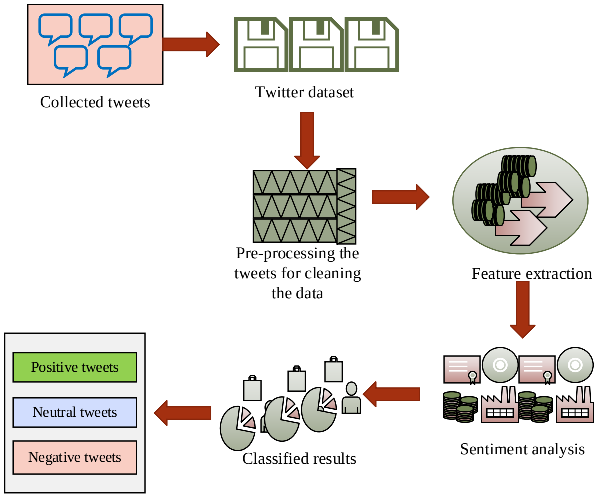


Fig. 2. System Model for Sentiment Analysis using Twitter Data.

IV. Proposed CRNF Methodology

In general, the sentiment analysis is one of the predictive modeling tasks that are trained with sentiments or textual data. However, the sentiment analysis using large data is the difficult task that provided lower efficiency for classifying the sentiments. In this research, the novel CatBoost Recurrent Neural Framework (CRNF) is developed for analysing the sentiments using twitter data. Here, the tweets are collected based on the COVID-19 vaccine from twitter that is utilized for training.

Moreover, the procedure of the developed CRNF approach is detailed in Fig. 3. Also, the developed CRNF approach is pre-processed the trained dataset and finally classifies the sentiment while removing the error. Thus, the proposed CRNF model is to remove the leaf node layer using pre-processing function and to enhance sentiment classification rate.

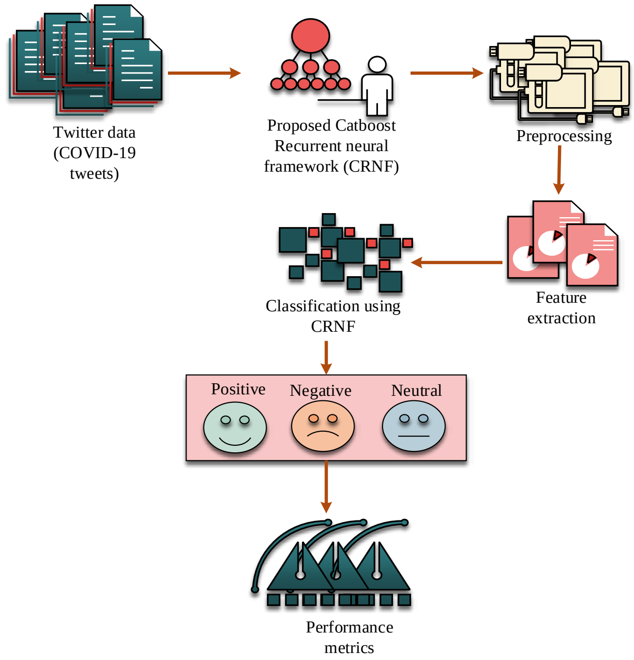

Twitter data (COVID-19 tweets)
Proposed Catboost Recurrent neural framework (CRNF)
Preprocessing
Feature extraction
Classification using CRNF
Positive
Negative
Neutral
Performance metrics

Fig. 3. Proposed CRNF Methodology.

A. Dataset Description

The proposed approach utilized the Twitter data for analysing the sentiments. Here, the twitter data based on COVID-19 vaccine tweets are collected from the kaggle.com that is processed in this research. The utilized dataset details about the 38460 numbers of tweets that involves the user name, location, description, friends, followers, favourites, text, tweet date, hashtags, and so on. In this, the collected tweets are based on the category of positive reviews, negative reviews and neutral reviews. Moreover, the collected dataset is given to the developed CRNF model for further processing.

B. CRNF Process for Sentiment Analysis

The proposed CRNF approach is processed on the twitter dataset for analyzing sentiments. Here, the developed catboost recurrent neural framework can reduce the error in the dataset, which is utilized for enhancing the classification accuracy [26, 27]. Here, the developed model performs the functions are pre-processing, feature extraction, and classification. The proposed CRNF model is the neural architecture that can reduce the training error, which is utilized to classify the sentiments in an effective manner. Furthermore the occurrence of CatBoost in the recurrent neural model can attain the enhanced classification accuracy as well as precision rate. Primarily, the dataset is initiated in the input layer of the network that is mentioned in eqn.(1),

$$d_T = \{(P_k)\}, k = 1,2,..., N \tag{1}$$

Where, P_k denotes the N^{th} quantity of tweets in the dataset d_T that involves positive, negative and neutral tweets. In this work, the dataset training process is done in the input layer. Here, P_k is the input and the output of the input layer is h^k that is given to the next layer. Moreover, the attained dataset having several errors or noise that should remove for attaining better results. So, the proposed model performing the pre-processing function.

- Pre-processing

This process is carried on the next layer of the network that is necessary for the dataset to remove the unnecessary data by cleaning the tweets, which involves the functions such as normalization, stop words removal and tokenization. Here, normalization process is utilized to remove the special characters, URLs, and emojis from the dataset. Also, the stop word deletion is processed to split the tweets that are compared using the stop words library, which involves the words not affects the original meaning the utilized sentences. Moreover, the raw is fragmented into the sentences or words with the use of tokenization process, which is employed to understanding the original meaning of text. Thus, the training errors in the tweets $(k = 1,2,3,...., N)$ are mentioned as H^{k-1} using additive manner that is represented by $H^k = H^{k-1} + \alpha h^k$ through the step size α and h^k utility removed the mistakes that is mentioned in eqn.(2),

$$h^k = d_T \left[\arg\min_{h \in H} \left(R^{k-1} + r_k \right) \right] \tag{2}$$

Where, the errors are mentioned as $\left(R^{k-1} + r_k\right)$ the repeated words r_k and the value of h^k is removed the errors and repeated words in a sentence, which are done in the hidden layer.

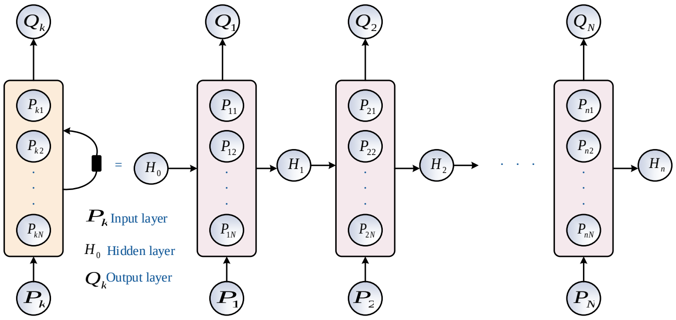


Fig. 4. Process of CRNF Network Model.

The dataset is given to the input layer of the proposed CRNF network that can be processed through the system. Moreover, the pre-processing and feature extraction function are done in the hidden layer of the network. Thus, the errors in the sentences are removed in the pre-processing process that is removed the leaf node layer. Additionally, the classification layer is utilized for classify the sentiments and finally, the output is obtained from the output layer of the network, which are represented in Fig. 4.

- Feature extraction

The pre-processed dataset is processed the feature extraction method that is done using the layer of the CRNF model. Here, the feature extraction is utilized for extracting the features from the dataset. Moreover, these features are utilized to identify the polarity of the sentences. In this approach, the feature extraction process is done using the factor φ_k and the activation function tanh that calculation is mentioned in eqn.(3),

$$f_k = \tanh \min \sum_{k=1}^{N} \left\{ \left(P_k \varphi_k \right) \right\} \tag{3}$$

Thus, the leaf node layer is removed while completing the pre-processing and feature extraction process. Moreover, the attained output is given to the classification layer that is performed the sentiment analysis process.

- Classification

The proposed method classifies the tweets like positive tweets P_t, negative tweets N_t, and neutral tweets N_l. The proposed model CRNF classifies the sentiments using the aspect terms in the sentences. Finally, the classification of sentiments is done using eqn.(4),

$$Q_k = d_T \rightarrow \sum_{k=1}^{N} P_k \left(A_W \left(P_t, N_t \right) \right) \tag{4}$$

Algorithm: CRNF for sentiment analysis

```
Input: COVID-19 Tweets from Twitter data
Output: classified output $(Pt, Nt, Nl)$
//where, $Pt$ -positive, $Nt$ -negative, and $Nl$ -neutral
Start
{
Initialization ()
        Initialize the dataset $d_T$ // COVID-19 Vaccine Tweets
        Import the dataset
Preprocessing()
        For all $d_T$ do
        Remove error, repeated words, noise, urls, numbers, special characters, stop words
        If misspelled (word) then
        Replace the word by correct word
        End if
        End for
Feature extraction()
        For all $d_T$ do
        Extract the features of the words
        End for
Classification()
        Mentioned the aspect terms $A_W(P_t, N_t)$
        If $A_W(P_t) > A_W(N_t)$ then
        $Sentiment \leftarrow P_t$ //(+1) positive tweet
        Else if $A_W(N_t) > A_W(P_t)$ then
        $Sentiment \leftarrow N_t$ //(-1) negative tweet
        Otherwise
        $Sentiment \leftarrow N_l$ // (0) neutral tweet
        End if
Classified sentiments
}
Stop
```

In this, the positive and negative aspect terms are saved in the layer that words are utilized to analyze the sentences. If the sentence having positive aspects then it is the positive tweet and if the sentence has negative aspects then it is negative tweets otherwise that sentence is considered neutral tweets. Finally, the results are attained from the output layer of the proposed network. Moreover, the positive tweets are represented as (+1), negative tweets are represented as (-1), and neutral tweets are represented as (0).

Thus, the sentiments have classified by the proposed catboost recurrent neural framework. The complete procedure of the proposed CRNF technique is detailed in the algorithm 1 and the flow chart is represented in Fig. 5.

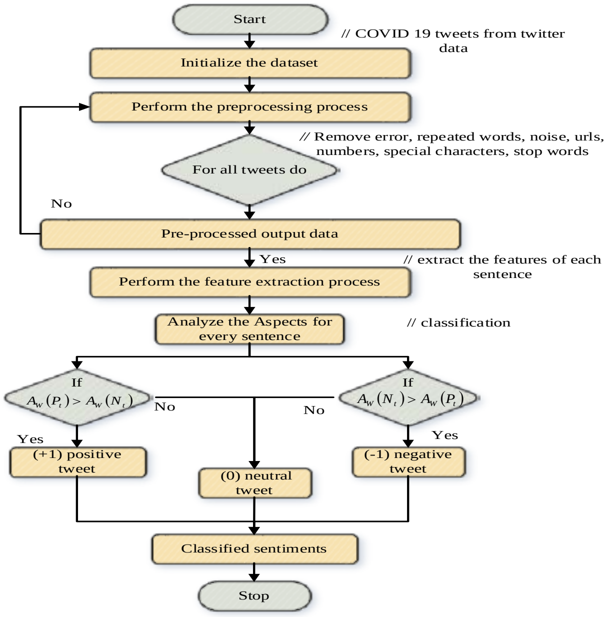

Start
// COVID 19 tweets from twitter data
Initialize the dataset
Perform the preprocessing process
// Remove error, repeated words, noise, urls, numbers, special characters, stop words
For all tweets do
No
Pre-processed output data
Yes
// extract the features of each sentence
Perform the feature extraction process
Analyze the Aspects for every sentence
// classification
If $A_w(P_t) > A_w(N_t)$
No
No
If $A_w(N_t) > A_w(P_t)$
Yes
Yes
(+1) positive tweet
(0) neutral tweet
(-1) negative tweet
Classified sentiments
Stop

Fig. 5. Flow Chart for CRNF Model.

V. Results and Discussion

In this work, the developed CRNF approach is simulated with the use of Python; moreover, the efficiency of the proposed strategy is evaluated with prevailing manners. Here, the comparison is carried out in the performance metrics like accuracy, recall, F-measure, precision, and error rate. In this, the developed is effectively classifies the sentiments using the proposed CRNF approach.

A. Case Study

In this paper, the sentiment analysis is done using the twitter data. Here, the tweets for COVID-19 are collected and processed using the proposed CRNF model. Several samples are for COVID-19 tweet and the classified results are mentioned in Table II. Here, the utilized dataset is initially pre-processed and feature extracted in the layers of the proposed CRNF network.

Subsequently, the positive and negative aspects are mentioned for classifying the sentiments. For example, vaccine, immunity, protect, etc., are considered as the positive aspects and sick, side effects, death, spread, etc., are considered as the negative aspects. Based on the considered aspects, each sentences are classified as positive tweet (+1), negative tweet (-1), and neutral tweet (0).

B. Performance Matrics

This research work performed the sentiment analysis using the developed CRNF approach, which is implemented using Python. Moreover, the performance metrics are calculated that are compared using existing methods for identifying the efficiency of the developed approach. Thus, the parameters like accuracy, precision, and error rate of the proposed model is validated with prevailing methods like Tree Augmented Naive Bayes (TAN) [21], Bag of words using machine learning (BOW-ML) [28], and Attention using Bidirectional CNN-RNN Deep Method (ABCDM) [29].

TABLE II. CLASSIFIED RESULTS FOR TWITTER DATASET ABOUT COVID-19

S.No	Text about COVID-19	Positive	Negative	Neutral
1	There are presently more than fifty COVID-19 vaccine contenders in the trials.	+1	-	-
2	The developed vaccine may cause the various side effects, which is related to the symptoms and signs of COVID-19.	-	-1	-
3	The Vaccine about COVID-19 is manufactured in Australia, which is supplied to the citizens at no cost. AFP quotes Prime Minister	+1	-	-
4	Got my CovidVaccine today. Ready to end this pandemic Protect your families.	+1	-	-
5	Got my covid vaccine! Tired, mild headache - work those antibodies, immune system	-	-1	-
6	Presently more than 50 numbers of COVID-19 vaccine candidates in trials.	+1	-	-
7	COVID-19 affected people develop mild to moderate disorder and recover without hospitalization.	-	-	0
8	Third stage of Russia's Covid-19 vaccine may initiate in seven to ten days	+1	-	-
9	COVID-19 is easily spread from one person to another like friends, family, and surrounding peoples.	-	-1	-
10	Masks are used to protect the people from COVID-19.	-	-	0

1) Accuracy: validation is utilized for determining the efficiency of the proposed framework. Also, it is identified the effectiveness of the developed model for classifying the sentiments, which is computed using eqn.(5),

$$Acc = \left(\frac{T_P + T_N}{T_P + F_P + F_N + T_N} \right) \tag{5}$$

Where, T_P represents the true positive that is the calculation for the total quantity of properly classified positive tweets, T_N is the true negative that represents the total quantity of properly classified negative tweets, F_P is the false positive that symbolizes the total quantity of improperly classified positive tweets, and F_N is the false negative that represents the total quantity of improperly classified negative tweets.

The accuracy calculation of the proposed CRNF model is compared with existing methods like TAN, BOW-ML, and ABCDM that are mentioned in Table III. The existing approaches TAN and BOW-ML approaches are attained lower accuracy as 80.8% and 85%.

Also, the ABCDM approach achieved nearly 93% accuracy. Thus, the proposed CRNF approach has attained high accuracy as 99.34% than other models while considering tweets data from twitter, which is represented in Fig. 6.

2) Precision: The calculation of precision is utilized for identifying the effectiveness of the proposed classifier. Here, the lower precision value denotes the high false positives and high precision rate denotes the less number of false positives. Moreover, the precision value of the proposed model is calculated using eqn. (6),

$$P = \left(\frac{T_P}{T_P + F_P} \right) \tag{6}$$

The precision value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table IV.

TABLE III. COMPARISION OF ACCURACY

No. of tweets taken	Accuracy (%)			
	TAN	**BOW-ML**	**ABCDM**	**CRNF [proposed]**
100	80.8	85.1	93.40	99.34
200	79.96	84.87	92.68	98.65
300	76.48	84.05	91.97	97.89
400	75	83.79	89.63	95.76
500	73.27	83.27	87.67	93.56

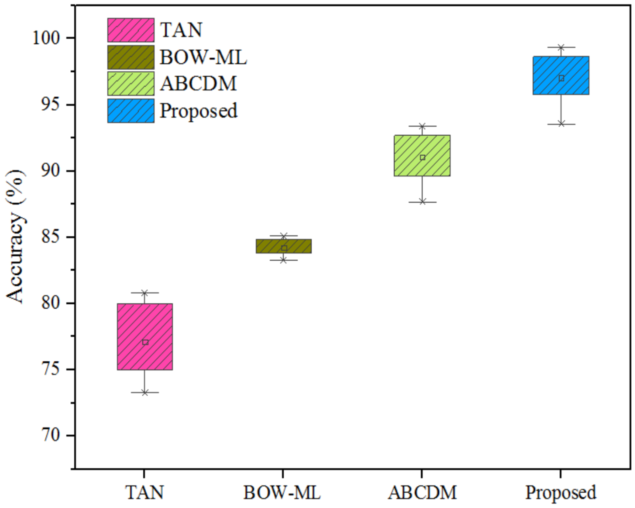


Fig. 6. Comparison of Accuracy.

TABLE IV. COMPARISON OF PRECISION

No. of tweets taken	Precision (%)			
	TAN	**BOW-ML**	**ABCDM**	**CRNF [proposed]**
100	90.6	83.6	95.70	98.38
200	88.04	82.87	93.78	97
300	85.58	80.96	93.27	96.78
400	83.47	79.67	92.09	96.07
500	82.97	77.86	90.83	95.36

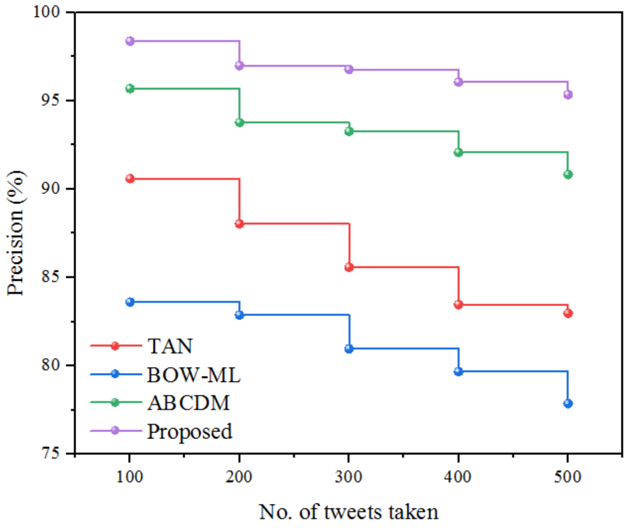


Fig. 7. Comparison of Precision.

Also, the existing BOW-ML model attained lower precision as 83.6%, TAN approach achieved 90.6% precision, and ABCDM model attained 95.7% precision value. Hence, the proposed CRNF model has achieved high precision rate as 98.38% than other methods that is represented in Fig. 7.

3) Recall: The calculation of recall is utilized for identifying the sensitivity or the completeness of the proposed classifier. In this, the lower recall value denotes the high false negatives and high recall rate denotes the less number of false negatives. Moreover, the recall value of the proposed model is calculated using eqn.(8),

$$R = \left(\frac{T_P}{T_P + F_N} \right) \tag{7}$$

The recall value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table V.

Also, the existing BOW-ML model attained lower recall as 88%, TAN approach achieved 85.4% recall, and ABCDM model attained 90.88% recall value. Therefore, the proposed CRNF model has achieved high recall rate as 97.45% than other methods that is represented in Fig. 8.

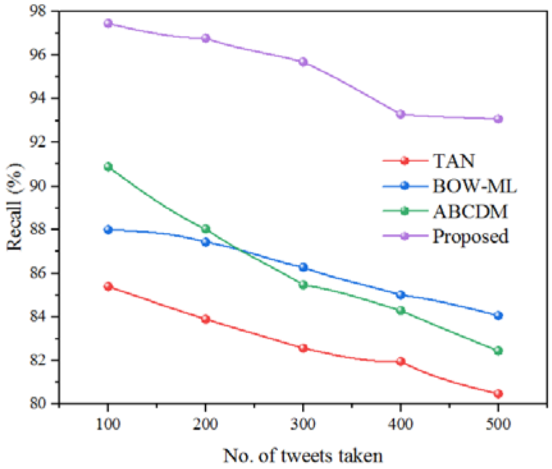


Fig. 8. Comparison of Recall.

TABLE V. COMPARISON OF RECALL

No. of tweets taken	Recall (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	85.4	88	90.88	97.45
200	83.9	87.43	88.04	96.78
300	82.57	86.28	85.47	95.69
400	81.96	85	84.30	93.29
500	80.48	84.07	82.45	93.08

4) F1-measure: The calculation of F1-score is defined the combination of the calculated precision and recall values, which is computed using eqn.(8),

$$F1-score=\left(2\frac{P*R}{P+R}\right) \tag{8}$$

The F1-measure value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table VI.

Also, the existing TAN approach achieved 87.9% F1-measure value, BOW-ML model attained lower F1-measure value as 85.8%, and ABCDM model attained 92.22% lower F1-measure value. Moreover, the proposed CRNF model has achieved high F1-measure value as 97.91% than other prevailing approaches that are characterized in Fig. 9.

5) Error Rate: This calculation is utilized to identify the classification error of the proposed model, which is computed using eqn.(9).

$$Error\_rate=\left(\frac{F_P+F_N}{T_P+T_N+F_P+F_N}\right) \tag{9}$$

The error rate value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table VII. These prevailing methods are attained higher error rate for classifying sentiments using twitter data.

TABLE VI. COMPARISON OF F1-MEASURE

No. of tweets taken	F1-measure (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	87.9	85.8	93.22	97.91
200	86.35	84.08	92.76	96.88
300	87.39	82.78	92.94	96.23
400	85.28	83.49	91.67	94.65
500	84.37	82.08	91.06	94.20

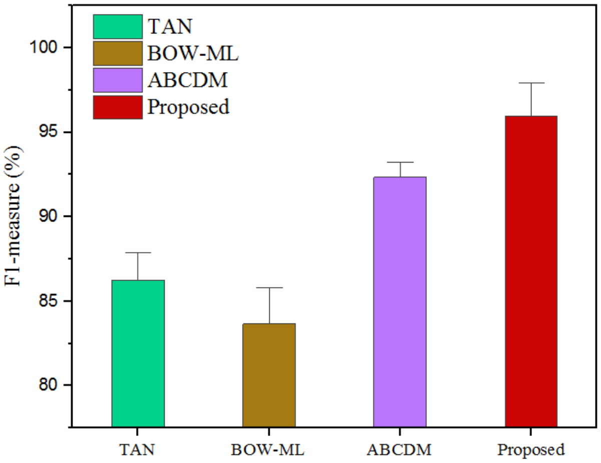


Fig. 9. Comparison of F1-Measure.

TABLE VII. COMPARISON OF ERROR RATE

No. of tweets taken	Error rate (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	19.2	14.9	6.6	0.66
200	20.04	15.13	7.32	1.35
300	23.52	15.95	8.03	2.11
400	25	16.21	10.37	4.24
500	26.73	16.73	12.33	6.44

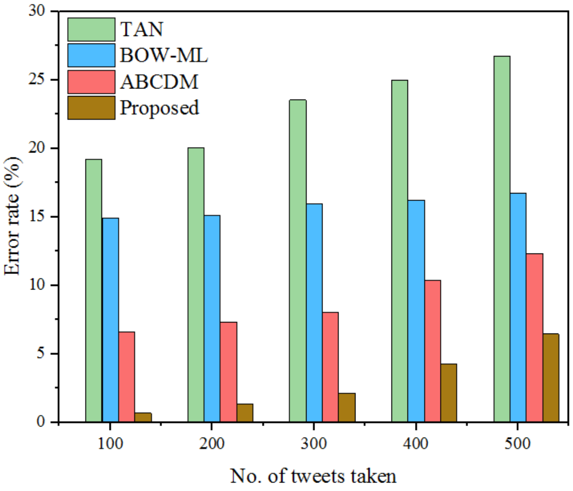


Fig. 10. Comparison of Error Rate.

Also, the existing TAN approach achieved 19.2% high error rate value, BOW-ML model attained 14.9% error rate value, and ABCDM model attained 6.6% error rate value. The comparison of the error rate value is characterized in Fig. 10. Moreover, the proposed CRNF model has achieved lower error rate value as 0.66% than other existing methods for classifying the sentiments using twitter data for COVID 19.

VI. Conclusion

In this research, the novel Catboost Recurrent Neural Framework (CRNF) is developed for performing sentiment analysis in the twitter dataset. Here, the tweets about COVID-19 are considered as the dataset that is utilized for classifying the sentiments as positive tweets, negative tweets, and neutral tweets. The noise, error, url, repeated words, stop words, numbers, special characters are removed by the pre-processing process. Also, the feature extraction method is utilized to extract the characteristics of each sentence. Subsequently, the classification of sentiments is done in the layer of the proposed CRNF model using the aspects words. Hence, the proposed model has achieved high accuracy as 99.34% with lower error rate as 0.66% than other existing approaches.

References

[1] Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). The turf is always greener: Predicting decommitments in college football recruiting using Twitter data. Decision Support Systems, 116, 1-12.

[2] D. Kandé, F. Camara, S. Ndiaye. "FWLSA-score: French and Wolof Lexicon-based for Sentiment Analysis", In 2019 5th International Conference on Information Management (ICIM), IEEE, 2019.

[3] Kursuncu, Ugur, et al. "Predictive analysis on Twitter: Techniques and applications." Emerging research challenges and opportunities in computational social network analysis and mining. Springer, Cham, 2019. 67-104.

[4] Liu, Xia. "Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods." Industrial marketing management 86 (2020): 30-39.

[5] Kumar, Sachin, and Mikhail Zymbler. "A machine learning approach to analyze customer satisfaction from airline tweets." Journal of Big Data 6.1 (2019): 62.

[6] Mandloi, Lokesh, and Ruchi Patel. "Twitter Sentiments Analysis Using Machine Learninig Methods." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.

[7] Ahmed, Hager, et al. "Heart disease identification from patients' social posts, machine learning solution on Spark." Future Generation Computer Systems 111 (2020): 714-722.

[8] Tahmasebi, Hossein, Reza Ravanmehr, and Rezvan Mohamadrezaei. "Social movie recommender system based on deep autoencoder network using Twitter data." Neural Computing and Applications (2020): 1-17.

[9] Hasan, Mahmud, Mehmet A. Orgun, and Rolf Schwitter. "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework." Information Processing & Management 56.3 (2019): 1146-1165.

[10] Gupta, Aakansha, and Rahul Katarya. "Social Media based Surveillance Systems for Healthcare using Machine Learning: A Systematic Review." Journal of Biomedical Informatics (2020): 103500.

[11] K. Sivakumar, N.S. Nithya, O. Revathy. "Phenotype Algorithm based Big Data Analytics for Cancer Diagnose", Journal of medical systems, 43(8), pp. 264, 2019.

[12] K. Weiying, D.N. Pham, Y. Eftekharypour. "Benchmarking NLP Toolkits for Enterprise Application", Pacific Rim International Conference on Artificial Intelligence, Springer, Cham, 2019.

[13] D. Benarji Tharini, V.V. Bulusu. "Development of a Micro Telugu Opinion WordNet and Aligning with TELOWN Ontology for Automatic Recognition of Opinion Words from Telugu Documents".

[14] M. Trupthi, S. Pabboju, N. Gugulotu. "Deep Sentiments Extraction for Consumer Products Using NLP-Based Technique", Soft Computing and Signal Processing, Springer, Singapore, pp. 191-201, 2019.

[15] B.A. Hammou, A.A. Lahcen, S. Mouline. "A Distributed Ensemble of Deep Convolutional Neural Networks with Random Forest for Big Data Sentiment Analysis", International Conference on Mobile, Secure, and Programmable Networking, Springer, Cham, 2019.

[16] K. Negi, A. Pavuri, L. Patel, C. Jain. "A novel method for drug-adverse event extraction using machine learning", Informatics in Medicine Unlocked, pp. 100190, 2019.

[17] H. Yang, L. Luo, L.P. Chueng, D. Ling, F. Chin. "Deep Learning and Its Applications to Natural Language Processing", Deep Learning: Fundamentals, Theory and Applications, Springer, Cham, pp. 89-109, 2019.

[18] J. von Bloh, T. Broekel, B. Özgun, R. Sternberg. "New (s) data for entrepreneurship research? An innovative approach to use big data on media coverage", Small Business Economics, pp. 1-22, 2019.

[19] Yang, Chao, et al. "Aspect-based sentiment analysis with alternating coattention networks." Information Processing & Management 56.3 (2019): 463-478.

[20] Chandra, Nidhi, Sunil Kumar Khatri, and Subhranil Som. "Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository." System Performance and Management Analytics. Springer, Singapore, 2019. 65-76.

[21] Ruz, Gonzalo A., Pablo A. Henríquez, and Aldo Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers." Future Generation Computer Systems 106 (2020): 92-104.

[22] Kumar, Sachin, and Marina I. Nezhurina. "An ensemble classification approach for prediction of user's next location based on Twitter data." Journal of Ambient Intelligence and Humanized Computing 10.11 (2019): 4503-4513.

[23] Hasan, Maryam, Elke Rundensteiner, and Emmanuel Agu. "Automatic emotion detection in text streams by analyzing twitter data." International Journal of Data Science and Analytics 7.1 (2019): 35-51.

[24] Barker, J. L. P., and Christopher JA Macleod. "Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities." Environmental modelling & software 115 (2019): 213-227.

[25] Li, Xin, et al. "Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology." Technological Forecasting and Social Change 146 (2019): 687-705.

[26] Huang, Guomin, et al. "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions." Journal of Hydrology 574 (2019): 1029-1041.

[27] Liu, Fagui, et al. "Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification."Neurocomputing 371 (2020): 39-50.

[28] Soumya, S., and K. V. Pramod. "Sentiment analysis of malayalam tweets using machine learning techniques." ICT Express (2020).

[29] Basiri, Mohammad Ehsan, et al. "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis." Future Generation Computer Systems 115 (2020): 279-294.

[30] V.Laxmi Narasamma, and M. Sreedevi. "Modeling of Tweet Summarization Systems using Data Mining Techniques: A Review Report." Indian Journal of Science and Technology 9 (2016): 44.

Image-based Onion Disease (Purple Blotch) Detection using Deep Convolutional Neural Network

Muhammad Ahmed Zaki[1], Sanam Narejo[2*]
Muhammad Ahsan[3], Sammer Zai[4], Muhammad Rizwan Anjum[5], Naseer u Din[6]
Department of Computer System Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan[1, 2, 3, 4, 6]
Department of Electronic Engineering, The Islamia University of Bahawalpur, Pakistan[5]

***Abstract*—Agriculture on earth is the biggest need for human sustenance. Over years, many farming methods and components have become computerized to guarantee quicker production with higher quality. Because of the enlarged demand in the farming industry, agricultural produce must be cultivated using an efficient process. Onion (Allium cepa L.) is an economically valuable crop and is the second-largest vegetable crop in the world. The spread of various diseases highly affected the production of the onion crop. One of the serious and most common diseases of onion worldwide is purple blotch. To compensate for a limited amount of training dataset of healthy and infected onion crops, the proposed method employs a pre-trained enhanced InceptionV3 model. The proposed model detects onion disease (purple blotch) from images by recognizing the abnormalities caused by the disease. The suggested approach achieves a classification accuracy of 85.47% in recognizing the disease. This research investigates a novel approach for the rapid and accurate diagnosis of plant/crop diseases, laying the theoretical foundation for the use of deep learning in agricultural information.**

Keywords*—Disease detection; disease classification; artificial intelligence; inceptionv3; deep convolutional neural network

I. Introduction

Farming production is the utmost important sector in many countries for contributing to domestic incomes [1]. It plays a significant role in stimulating the economy of agricultural countries like India, Bangladesh, Indonesia, Iran, Turkey, Afghanistan, and Pakistan which depend heavily on crop quality and growth [2]. Productivity in agriculture is the backbone of a country's economy. With a population of more than 220 million people, Pakistan is the sixth most populated country on earth. Approximately 64% of the population, around 140.8 million lives in rural areas [3]. Agriculture plays a crucial role to guarantee food safety for a relatively enormous and increasing population. In the agriculture market, vegetables are of prime importance. The share of vegetable production in Punjab is 95.4% for potatoes, 19.9% for onions, 21.1% for chilies, and 15.9% for tomatoes [4]. Punjab is by far the largest province for the farming industry in Pakistan. The total area under agriculture in Punjab is 40.55 million acres, which makes 71.53% of Pakistan's total agricultural area [5, 6]. Sindh, in particular for tomatoes, chilies [7], and bananas [8], is the second most important province for agricultural production. Onion prices have been falling in the global market in 2020 since February, as logistical problems from covid19 [9] hampered trade and demand in major producing countries, India and China remained lukewarm. Though demand in America has improved, the recent onion outbreak may have an impact on the United State (US) and Mexico export situation. Mexico's onion harvest is divided into two seasons, the spring season lasts from March to June and the fall season from August to January. The volume of onion exports to the US, its main importing destination has increased significantly over past years with a nearly 80% increase from April to July compared to the same period last year. Since the outbreak began demand for onion in the US has been stable with the bulk of it coming from retailers. Onion cultivation in the US appears to be quite consistent with no weather anomalies. According to the onion business the coronavirus had little influence on US onion shipments to Asian countries because the season extends from August to January the remaining seasons are supplied by New Zealand. With the increasing cargo charges moving the product across the country could be difficult but the foodservice business is warily hopeful about reopening which might handle the onion that began packing in August. During the early stages of covid19, trade flows were restricted due to logistical issues that made it difficult to export Chinese goods and lockdowns in Asian countries that slowed down exports. However, the trade situation has now returned to normal.

In Pakistan, a brief overview of the main crops is highlighted in Table I. Onion (Allium cepa L.) is the world's second-largest vegetable crop with an international supply of approximately 74.25 million tons [10] and is commercially the most important crop of the Alliaceae family [11]. Onion is the main relish commonly used throughout the year in all homes [12]. Recent research has shown that dietary onions can help prevent heatstroke and other disorders [13]. The onion bulb [14] is abundant in carbohydrates, calcium, and phosphorus.

Table II indicates the composition of the onion bulb. Food historians agree that onion originated in central Asia 5000 years ago, some suggest onion was initially grown in Iran and West Pakistan [15, 16]. Since onion has grown wild in different areas it has probably been eaten for multiple years all over the world [17]. Onion may be one of the initially cultivated crops since they were simple to harvest, less fragile, transportable, grown in a diversity of climates and soils than other crops [18] as shown in Fig. 1.

*Corresponding Author

TABLE I. CASH CROPS OF PAKISTAN

Major Crops	**Area (HA)**	**Production (Tons)**	**Global Ranking (Value)**
Cotton seed	2805700	4071400	3rd
Sugarcane	1128800	63749900	5th
Mangoes, Guavas	171289	1658562	6th
Wheat	8686602	24211400	8th
Onions	125900	1660800	8th
Rice	2789200	6798100	13th
Maize	1168490	4944210	22nd

TABLE II. COMPOSITION OF ONION BULB

Carbohydrates	11g	Calcium	0.2g
Moisture	87 g	Vitamin C	11 mg
Protein	1.3 g	Phosphorus	0.05 µg
Thiamine	80 µg	Riboflavin	0.01 mg
Fiber	0.65 g	Iron	0.8 mg
Minerals	0.45 g	Nicotinic acid	0.4 mg

Fig. 1. Fresh Onion.

Besides, onion was beneficial for human life to be sustained. Onion avoided thirst when food gets scarce could be dried and stored for later consumption. Onion helps lower blood pressure [19], lessen the occurrence of cancer [20], insect bite antiseptic, anti-diabetic, anti-aging, and stimulate hair regrowth [21]. The total world production of onions is 742.5 million tons with Pakistan ranked 7th [22] as presented in Table III.

TABLE III. WORLD TOTAL ONION PRODUCTION

S No.	**Countries**	**Million tons**
1	China	205.08
2	India	133.72
3	USA	33.22
4	Egypt	22.09
5	Iran	19.24
6	Turkey	19.01
7	Pakistan	17.02
8	Brazil	15.57
9	Russia	15.37
10	Republic of Korea	14.12

TABLE IV. STATUS OF ONION (PROVINCE-WISE) 2006-07

Province	**Area (HA)**	**% share**	**Production (Tones)**	**% share**	**Average Yield t/ha**
Punjab	35.6	27	315.7	17	8.9
KPK	12.0	9	206.1	11	17.2
Sindh	45.6	35	593.1	33	13.0
Balochistan	38.2	29	701.6	39	18.4
Pakistan	131.4		1816.5		13.8

Onions are generally grown in the field with a 10-12 cm seedling gap and 25-30 cm out in flatbed rows. To generate huge bulbs, space the crops 10-15 cm out in a row. Maintain onion weeds free by hoeing and shallow cultivation. Planting seeds directly in the field where the crop will mature, planting in a seedbed from which the plants will be moved subsequently to the field, and planting sets are the three most distinct techniques of planting onion. These sets can be bought or cultivated from seed by the grower. For early production, the transplanting approach is widely used. Onion is grown commercially on an area of 131.4 thousand hectares, producing 1.8 million tons [23]. The onion status (province-wise) for 2006-07 is shown in Table IV.

22 districts account for over 77% of Pakistan's total onion production [24]. In the world 66 syndromes including 38 fungal, 10 bacterial, 3 biological, 6 nematodes, 1 sponging plant, and 1 phytoplasmal ailment, 7 various infections and ailments are affecting onion. Numerous serious infections such as soil-borne ailments are becoming extensive and sufficiently extreme to restrict onion production worldwide. Fungi and bacteria can cause a variety of diseases in onion. Crop rotations, climate, storage, drying conditions, and disease control measures disease intensity. Botrytis blight, botrytis neck rot, black mold rot, pink root, botrytis bulb rot, bacterial soft rot, iris yellow spot virus, fusarium basal rot, purple blotch, translucent scale, and downy mildew are onion diseases worldwide [25]. Many onion diseases start on crops in the field and progress to the bulbs during transportation and storage. Understanding that postharvest diseases arise in the field is the first step toward effective control. Controlling onion diseases require suitable cultural practices such as removal of contaminated onion, crop rotations, culls, debris, and cultivar range [26]. Some common onion crop diseases found worldwide are presented in Table V and Fig. 2.

Following are the main contributions of this paper:

1) We demonstrated the feasibility of our approach by creating a dataset of 1000 images of onion crops.

2) Proposed an enhanced InceptionV3 classification model, which is used to identify purple blotch affected onion based on their images.

3) To improve classification accuracy, various pre-processing and training techniques are used.

4) The proposed model is created to aid in more accurate and efficient detection of onion disease (purple blotch).

TABLE V. COMMON DISEASES OF ONION CROP FOUND WORLDWIDE

Diseases	Caused By	Pathogen
Botrytis Blight	Fungus	Botrytis Squamosa
Botrytis Neck Rot	Fungus	Botrytis Allii
Purple Blotch	Spores Of The Fungus	Alternaria Porri
Fusarium Basal Rot	Fungus	Fusarium Oxysporum F. Sp. Cepae
Translucent Scale	Regular High Relative Humidity, and High Temperature (i.e. >90° F)	
Downy Mildew	Fungus	Peronospora Destructor

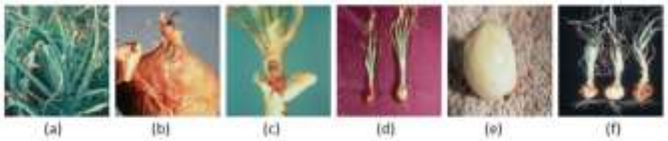


Fig. 2. Common Onion Crop Diseases (a) Botrytis Blight (b) Botrytis Neck Rot (c) Purple Blotch (d) Fusarium Basal Rot (e) Translucent Scale (f) Downy Mildew.

The remaining article is arranged as follows: Sections II describes onion disease (purple blotch). Section III presents and summarizes related work. The state of art models for disease detection is discussed in Section IV. Section V explains what knowledge is needed to analyze the proposed model. Section VI gives a dataset description. The implementation details, performance metrics, and experimental results are described in Section VII. Lastly, Section VIII presents the outlines of the article's findings.

II. PURPLE BLOTCH DISEASE

Younger leaves are more vulnerable than older leaves. Water-soaked lesions with a white center are the first signs of the disease. Lesions edges turn brown to purple the leaves above and below the lesions turn yellow. Dark brown to black concentric rings form over time inside the lesions. The fungus sporulates in these regions. Lesions can girdle the leaf as the disease progresses causing it to collapse and die as shown in Fig. 3. Seed stalks exhibit similar symptoms and infected stalks may collapse resulting in shriveled seed growth. The most common route for bulb infection is through the neck. If the fungus infects the bulb the infected area appears bright yellow at first, but then turns a distinctive red wine color.

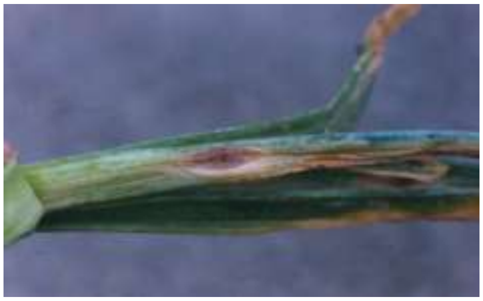


Fig. 3. Onion Disease (Purple Blotch) [27].

a) Circumstances for Disease Expansion: In leaf debris and cull piles the fungus survives the winter as mycelium. During humid nights leaf wetness cycles exceed more than 12 hours spores are developed. When the morning dew evaporates spores become airborne and spread to vulnerable onion tissues. It takes 1-4 days for signs to appear after infection. During extended periods of leaf wetness, disease production is at its peak.

b) Overcome: A broad-spectrum protection fungicide spray program can provide good protection before infection. Reduce disease growth by reducing leaf wetness using surface irrigation rather than sprinkler irrigation, good field drainage, and proper plant spacing. Rotating onions to unrelated crops for many years can also help to minimize disease.

III. RELATED WORK

Nihar et al. [28] suggested a neural network-based system for detecting plant disease that will aid in the development of the agricultural system which can properly determine whether a plant is infected or healthy and has a 97.7% accuracy. This technology enables the user to detect diseases faster allowing them to take acceptable precautionary steps and save crops.

Sangeetha et al. [29] compare different methods of image processing (K-mean clustering, artificial neural network (ANN), support vector machine (SVM), and fuzzy logic) then examine and summarize crop disease types. This study provides a thorough explanation of the machine learning (ML) models that were used to identify various agricultural diseases.

Liu et al. [30] presented a kiwi RGB-NIR-D dataset containing multi-modality associated images of kiwi fruits in farms. They suggested a unique method for using RGB-D sensors to combine associated near-infrared reflectance (NIR) and RGB images with deep learning (DL) techniques for fruit recognition. Using two modalities of aligned RGB and NIR images from the dataset faster region-based CNN is designed and deployed for kiwi fruit recognition.

Kim et al. [31] have proposed an image-built automated field monitoring method. Based on a weekly supervised learning approach the deep learning model was trained to identify crop disease. Using the field monitoring method the model was trained using captured onion crop field images, 6 groups including the sign of the disease were classified. As an ideal disease symptom localization threshold, 60% of the extreme value in the class activation map was calculated. The efficiency of recognizing disease symptoms was evaluated by intersection over union (IoU) using the mean average precision (mAP) metric.

Sharma et al. [32] enhanced decision-making with convolutional neural network (CNN) for various paddy crop diseases to prevent disease at initial stages and avoidance of mass loss in yield productivity. Paddy crop conditions are very fatal and can have serious effects on crops if initial treatment is not taken into account.

Karthik et al. [33] have used two distinct profound architectures to diagnose tomato leaf infection. To learn essential characteristics for classification the first architecture relates to residual learning. On top of the residual deep network (RDN) the second architecture applies an attention function. Experiments were carried out over 3 diseases namely leaf mold, late blight, and early blight from the plant village dataset. Using the attention mechanism, the suggested work exploited the features acquired by CNN at different handling hierarchies and accomplished a whole accuracy of 98% in the 5-fold cross-validation on the validation sets.

Pattnaik et al. [34] suggested a transfer learning system for the classification of pests in tomato plants based on an existing deep CNN structure. The database for the analysis was obtained from 859 images in 10 groups from online sources. A comprehensive contrast of the performance of 15 existing CNN models has been offered. The test results showed that 88.83% of the maximum classification accuracy was achieved with the DenseNet169 model.

Francis et al. [35] offered a detailed summary of the prevalent applications of DL and computer vision (CV) techniques in the field of farming demonstrating the need to recognize and classify diseases using a dataset of leaf images. It is proposed to clarify its working principle with a novel classification system. A new collection containing gradient images is created using the multi-space image reconstruction input. Through the original and reconstructed images, high-level semantic structures are mined by convolutional layers that are convolutional and depth-wise separable. Finally, for classification SoftMax was used. The hyper-parameters and the expense of computation are calculated mathematically providing researchers with an insight innovation. The output of the system is compared and estimated with the related works on the openly accessible dataset of apple leaf images.

In this article, we have presented a methodology to detect onion disease (purple blotch) by using CNN classification. The feasibility of our approach has been demonstrated by using a dataset of 1000 images for healthy and infected onion crops. Two different types of experiments were taken using an improved InceptionV3 for classification with different batch sizes. Experiment trials used 70% of the dataset for training, 25% of the dataset for testing, and 5% of the dataset for validation. The proposed method recognizes the disease with a classification accuracy of 77.05% and 85.47% for different batch sizes.

IV. The Proposed Method

The proposed method uses CNN to classify purple blotch disease in an onion crop. Finding an optimal architecture for CNN is one of the main challenges. Therefore we have used transfer learning where a model established for a mission is reprocessed as the beginning point for another task [36]. We have used a pre-trained CNN for recognition. This model is a functional model with multiple layers. The architecture of this model is shown in Fig. 4. This model is used to analyze the images and reliably detect them.

A. Input Layer

In image processing, it usually represents the image's pixel matrix. The parameter determines the dimension of the image (244×244×3) and due to data limitations, we augment our dataset by randomly generating images, adjusting the zoom and shear parameters of the original images. All images are duplicated five times. These copies are then fed into a pre-trained model, which is described further down.

B. Conv2D + ReLU Layer

Convolution is a linear procedure consisting of the multiplication between a two-dimensional (2D) weight array (filters) and an input data array of a weight set. We have 3 layers in the suggested design with zero padding and a filter of size 3×3. ReLU is abbreviated as a rectified linear unit that relates to the feature of non-saturating activation. It eliminates unwanted values from an activation map efficiently by setting them to nil. Without influencing the receptive fields of the convolution layer it rises the nonlinear assets of the resolution function and the whole network as specified in equation (1).

$$\text{ReLU} : f(x) = \max(0; x) \quad (1)$$

C. Pooling Layer

The pooling layer performs down-sampling feature maps by summarizing the presence of characteristics in the feature map patches. There are two kinds of pooling approaches max pooling and average pooling. We have used average pooling in the suggested design to determine the extreme value for each function map in each patch. The pooling average is set to 2×2 with 2 strides.

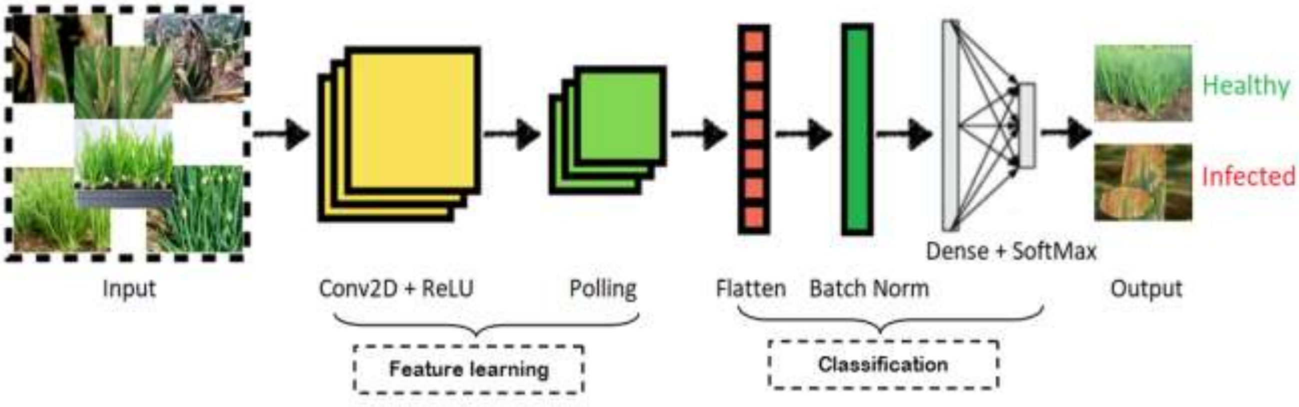


Fig. 4. CNN Architecture.

D. Flattening Layer

Flattening reduces file size by combining all apparent layers into the ground layer or transforms the data into a 1-dimensional array that is passed to the fully connected layer. To create a particular long feature vector we have dense the output of the convolutional layers.

E. Batch Normalization Layer

The batch normalization layer allows each layer to learn more individually. Learning becomes more rapid when batch normalization is utilized it can also be employed as a regularization to prevent model over-fitting. To standardize the outputs/ inputs, the layer is added to the model.

F. Dense + SoftMax Layer

The dense layer enhances an exciting non-linearity property, so any math function can be modeled. Though, they are always restricted in the logic that we still get the same output vector for the same input vector. On similar data, they can't identify recurrence in time or generate dissimilar responses. SoftMax layer is used to deliberate the feasible values of the preceding layer activation function. The values can be represented in two classes '0' and '1'. The logistic regression model is typically used for binary classification, although the SoftMax classification method is used for multi-classification. The classifier SoftMax is essentially an exponential normalized function as shown in equation (2).

$$\sigma(\hat{z}) = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \tag{2}$$

Where

$\sigma = \text{Softmax}$

$\hat{z} = \text{Input vector}$

e^{zi} = Standard input vector exponential function

K = Number of multi-class classifier classes

e^{zj} = Standard output vector exponential function

G. Output Layer

The output layer manages of generating the result. In a neural network, there must always be one output layer. The output layer receives the inputs from the layers above it executes the calculations using its neurons and then computes the output.

V. Model Consideration

One of the main goals of this research is to achieve state of art classification results using diverse transfer learning (DTL) models to both compensate for the small size of the trial data and to speed up the training procedure so that on modest hardware this can be achieved practically. These prototypes are all conveniently accessible as part of the Keras API and individually support TL in the form of support for pre-application to the model of the ImageNet weight.

A. Densenet201

DenseNet201 (dense convolutional network) is a 201 layers deep CNN that accepts an image input size of 224×224 [37]. DenseNet201 is a ResNet enhancement that requires dense connections between layers. In a feed-forward fashion, it ties each layer to every other layer. DensNet201 has $\frac{L(L+1)}{2}$ straight connections unlike typical CNN with L layers that have L connections. DenseNet has been able to boost efficiency compared to conventional networks by raising the necessity for computing, dropping the number of parameters, facilitating the reuse of features, and improving the propagation of features.

B. Inception-ResNet-V2

Inception-ResNet-V2 is a trained CNN on over a million ImageNet dataset images [38]. It is a cross procedure that incorporates the structure of residual connection and inception. The model takes a 299×299 dimension image and a list of approximate class probabilities in its output. The benefits of Inception-Resnet-V2 are the conversion of inception units to residual inception chunks and the adding of a new inception unit (Inception-A) to the stem module.

C. VGG16 and VGG19

VGG16 and VGG19 [39] are CNN architectures aimed to attain high accuracy in significant image identification applications with very narrow convolution filters (3×3). The depth [40] of the max-pooling, convolution, and fully connected layers vary between two implementations: 16 layers in VGG16 and 19 layers in VGG19.

D. MobileNet-V2

MobileNetV2 is an enhanced version of MobileNetV1 [41] it is consists of 54 layers and has an image input size of 224×224×3. Its key feature is that it uses two 1D convolutions with 2 kernels rather than having a single kernel 2D convolution. This implies that less storage, parameters, small and efficient model are needed for training. Two types of blocks can be distinguished: the first is a residual block with a stride of 1 and the other is blocked with a stride of 2 for downscaling. There are layers for each block: the first layer is ReLU6 1×1 convolution, the second layer is the deep convolution and another 1×1 convolution is the third layer but deprived of any non-linearity.

E. ResNet50

The ResNet was developed by executing a method of skip connections among layers known as residual learning to avoid the disappearing gradient problem characteristic in deep neural networks [42]. This design results in a network that trains extra effectively enable the design of deeper networks that have a positive effect on the accuracy of the model. ResNet50 is such a system implementing residual learning with 50 layers.

F. InceptionV3

The goal of InceptionV3 was to maximize the usage of computing assets within the network by raising the network's depth and width while retaining constant computing operations [43]. To define an enhanced network structure with avoided connections that are used as a building block the designers of this network coined the word 'Inception modules'. By assembling with max-pooling layers, [44] this initial module is spatially repeated to minimize dimensionality to a reasonable smooth for calculation.

G. Xception

The "extreme" type of Inception model was developed by Google Inc [45]. Xception is consisting of 71 deep layers. Only depthwise detachable convolution layers are used in the Xception architecture.

VI. Dataset Description

On our image dataset, we applied our offerings for instinctive binary classification. The following section gives the steps of our contribution in detail.

A. Dataset

This present work introduces an image dataset that contains onion crop images. Under the guidance of Dr. Imtiaz A. Nizamani, Department of Plant Protection, Sindh Agriculture University, Tandojam, Pakistan images of onion crops were gathered from Tandojam, Sindh onion fields for this study. In this case, there are a total of 1000 images (png and jpeg format) with 600 images of healthy crops and 400 images of infected crops with 'purple blotch disease' which affects onion. However, when evaluating the onion crop will only reveal whether it is healthy or unhealthy. When it is unhealthy the onion crop will display four stages (low, medium, high, and fully infected). The crop becomes highly unstable, leaf length diminishes, flower excision occurs, and so on. To protect the land the sick crop should be separated from the healthy crop. Fig. 5 illustrates an example of the dataset used in this research where (a) shows infected onion images (b) shows healthy onion images. For splitting of data, we used in this experiment 70% of the dataset for training, 25% of the dataset for testing, and 5% of the dataset for validation. We guarantee that the images selected for validation are not used throughout training to accomplish the binary classification task. Furthermore, we perceived that our database is unprovoked certainly 40% of the images represent the infected onion class. To overwhelmed this issue by using data augmentation we re-sampled our dataset. We produced two new images with various augmentation techniques from each single input image. The total quantity of images was therefore increased by two times in the healthy and infected onion classes.

B. Data Augmentation

After dataset splitting, and pre-processing for the training process data augmentation is used it helps to prevent the possibility of over-fitting. The image of the onion was resized to 224×224. To prevent distortion the black background of 224×224 pixels was applied to the images with various pixel ratios to accomplish a comprehensive transformation. Low-contrast images of the whole onion crop were omitted. To make the requisite variety to help CNN's generalization competencies, data augmentation is mandatory. In particular, the images are arbitrarily rotated by a maximum of 10^o and randomly moved vertically or horizontally in either direction by an all-out of 20 pixels. In this way, CNN acquires to be flexible in the variation of position and orientation alteration. Besides, geometric transformations such as rotations, re-scaling, shears, shifts, flips, zooms are the methods we have used, shown in Table VI.

TABLE VI. Data Augmentation

Argument	Explanation
Rescaling	Rescaling images from integers 0-255 to float 0-1
Rotation	Degree variety of the arbitrary rotation
Shift	Shifts vertically or horizontally by (10%) and is the fraction of given dimension
Shear	Permit being sheared the angle in an anti-clockwise
Zoom	Permits the image to be “zoom-in” or “zoom-out”
Flip	Permitted the input to be flipped vertically or horizontally throughout the training procedure
Fill	The neighboring pixel value is repeated and chosen for all the vacant values

C. Data Limitation

Some drawbacks are faced by the dataset which has to be listed. Infected onion crops are incorporated into a relatively small sample. Moreover, the onion field with severe symptoms may be used to generate this sample analysis which is required. In the existing dataset crops with minor symptoms are absent which is due to the strategy of people with onion fields who have slight disease indications.

D. Classification and Training Dataset

When data pre-processing, data augmentation, and data splitting techniques are done our training onion dataset size is increased and arranged for the feature extraction phase with the model proposed to extract the necessary and applicable features. To construct vectorized feature maps the features extracted from the suggested model are compressed together. To classify the image into the corresponding groups the generated feature vector is transferred to a multi-layer perceptron. Lastly, on test images using the training model the efficiency of the proposed method is assessed. Each experiment is replicated three times and we report the average performance.

VII. Experimental Setup

Our experiments have been carried out based on the following criteria for automatic binary classification established on our image dataset. All images in the dataset have been redimensioned to 224×224 pixels. We have set the batch size 8 and 16, the epochs are 500 to train our model. The samples are 155 and 90 respectively, for the training and validation. For optimization adam β1=0.89, β2=0.98, and learning rates are set at 10-5 and reduced to 10-6. Besides, we used decay weight to reduce our model over-fitting. The regularizers are supplied from Keras. The performance was specific to GPUs. Dell Inspiron Core(TM) i5-8250U CPU (8 CPU's) of 1.6 GHz, 16 GB Ram running on a Microsoft Windows 10 Professional (64-bit) is used in the implementation of the proposed model. The authors used Anaconda Navigator (Jupyter Notebook) version 3.7.4 to perform their research Keras, TensorFlow, OpenCV, NumPy, sklearrn are used for simulation as a backend for deep learning.

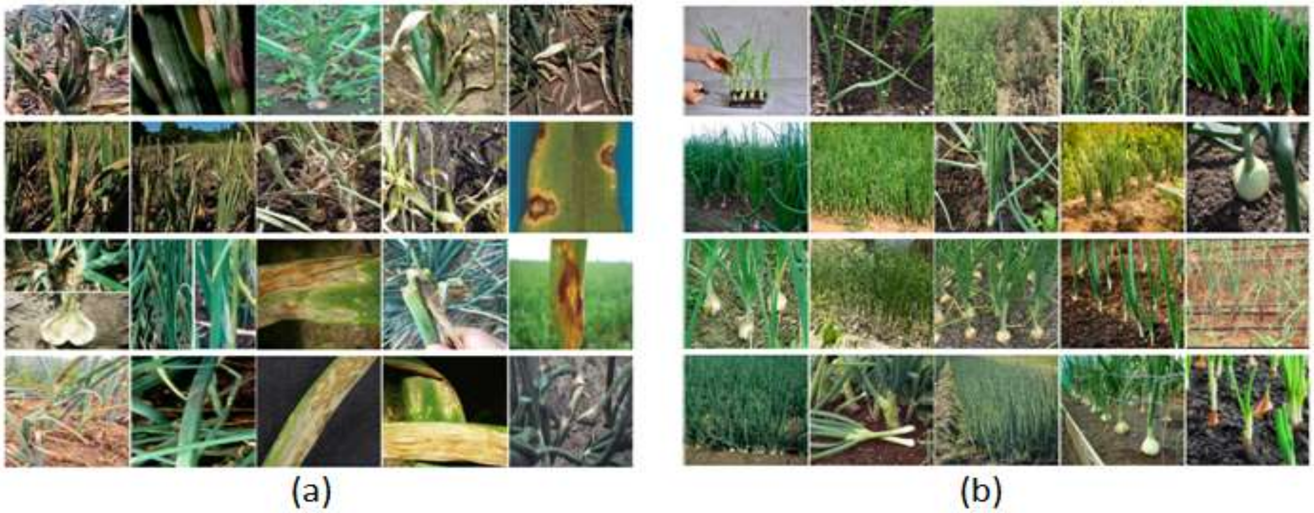


Fig. 5. Sample of Images used in Research (a) Infected Onion, (b) Healthy Onion.

A. Performance Metrics

The evaluation metrics that were used to measure the model's classification efficiency are explained in this section. For this, we use confusion matrix-based metrics [46]. Accuracy, recall, precision, F1-score, specificity, and sensitivity are examples of these metrics. We require the count of the following quantities to evaluate these measures: true positive, false positive, true negative, and false negative.

1) Accuracy: The ratio of correctly predicted predictions to the total number of predictions shown in equation (3).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \tag{3}$$

2) Precision: The ratio of true positive prediction compared to overall positive predictions shown in equation (4).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

3) Recall: The ratio of a true positive prediction made by the suggested model to the total number of predictions made shown in equation (5).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

4) F1-Score: The harmonic mean of recall and precision is shown in equation (6).

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{6}$$

5) Sensitivity: Measure the fraction of true positives correctly recognized shown in equation (7).

$$\text{Sensitivity} = \frac{TP}{\text{TP+FN}} \tag{7}$$

6) Specificity: (Also called TN rate) trials are performed on the fraction of true negatives properly-recognized shown in equation (8).

$$\text{Specificity} = \frac{TN}{\text{TN+FP}} \tag{8}$$

7) Confusion matrix: It is the evaluation of the model's results. It contrasts the true and predicted values in terms of true positive, false positive, true negative, and false negative shown in equation (9).

$$\begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix} \tag{9}$$

- True Positive (TP): True positive predictions are those that were initially positive and were also predicted to be positive by the AI model.
- False Positive (FP): False positive predictions are those that were initially negative but were predicted as positive by the AI model.
- True Negative (TN): True negative predictions are those that were initially negative and were predicted as negative by the AI model.
- False Negative (FN): False-negative predictions are those that were initially positive but were predicted as negative by the AI model.

B. Experimental Results

Our research aimed to classify purple blotch disease from onion crop images using an improved InceptionV3 classifier we present the result for the binary classification in this section. Besides, several experiments are carried out on the onion crop disease (purple blotch) dataset to verify the efficiency and robustness of the suggested model. The findings are shown separately in Fig. 6 and Fig. 7 using the accuracy, loss, and receiver operating characteristic (ROC) curve of training and validation. The experiment is detailed described in Table VII. It indicates the overall outcome of the first trial batch size 8 and 500 Epochs. Accuracy and loss are calculated each after 100 epochs. We present the two most used parameters of computer vision and deep learning before discussing these results: Train curve is considered from the training dataset that gives an idea of how well the model is learning, while the validation curve is considered from a hold-out validation dataset that gives an idea of how well the model is generalizing. Besides, the loss of validation and training is defined as a summing up of the errors made in the training sets of each case. The loss is not a fraction in comparison to accuracy. To summarize, the model that simplifies well, this is neither under-fitting nor over-fitting. The confusion matrix also displays a comprehensive illustration of what occurs to images later classification. We have observed that the accuracy increases until the value is 83.48% for the train and test accuracy from epoch 0 to epoch 7. The accuracy begins to be stable after epoch 10 were 87.01% and 84.42% for training and testing data respectively. In either the fast-growing interval from epoch 0 to epoch 32 where the loss is 2.98% or in the other interval where the decrease is slow and converges to 1.76% a good fit can be found for the loss curve of train results. As shown in the confusion matrix in Table IX, the model was able to classify 80 images as infected and 1 image as healthy for the infected onion class. The model with an improved InceptionV3 could predict 14 images as healthy and 27 images as infected for the healthy onion class. Similarly, Table VIII shows the overall result of the second trial with batch size 16 and 500 epochs. Likewise, accuracy and loss are calculated each after 100 epochs. We have observed that the accuracy improves until the value is 88.98% for the training and testing from 0 epoch to 45 epoch. The accuracy is stable at epoch 38 where it is equivalent to 87.79% for training data and 86.87% for testing data. The loss curve of train data can be seen as a good match in the steadily rising interval between 0 and 100 epoch where the loss is 1.388% at the other time the decrease is gradual and converges to 1.14% at the end converges to 0.756%. As illustrated, confusion matrix in Table IX. The model can classify 78 images as infected and 1 image as healthy for the infected onion class. It can be shown that the InceptionV3 model was capable of predicting 22 images as healthy and 16 images as infected for the healthy onion class.

In the last, a comparison is made between batch sizes 8 and 16 as shown in Table IX predicted results are shown in Fig. 8. Generally, the outcomes of this trial show that the model pays complete courtesy to the features of the spot disease and achieves excellent onion disease recognition efficiency and it is very clear that batch size 16 provides approximate results.

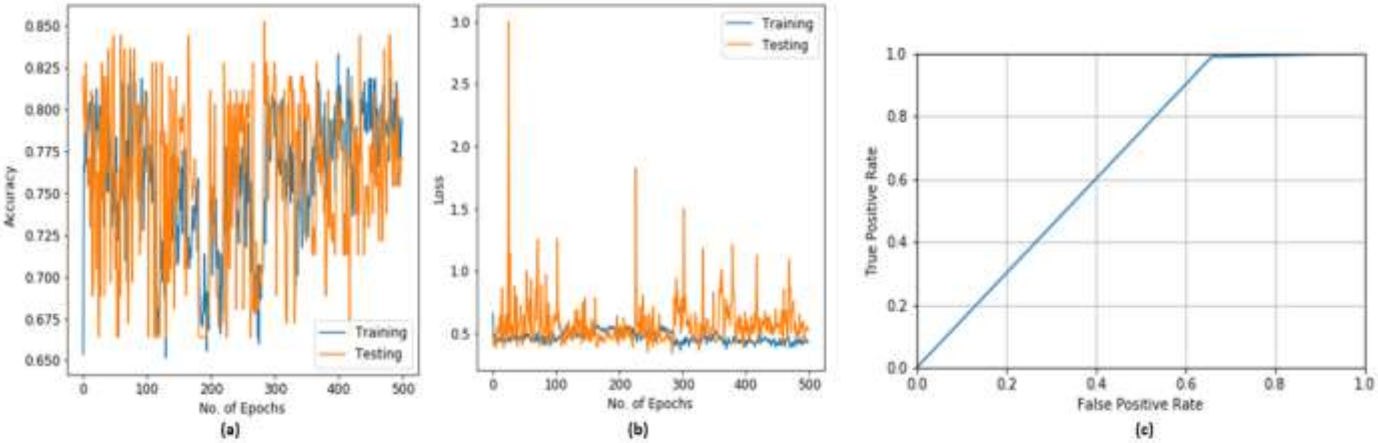


Fig. 6. CNN Model with Batch Size 8 (a) Accuracy, (b) Loss, (c) ROC Curve.

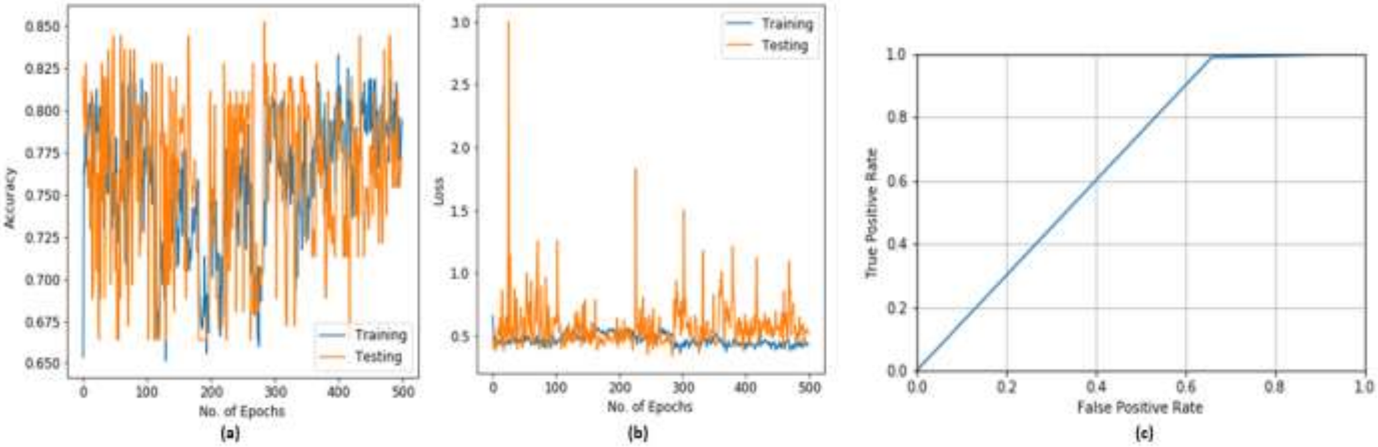


Fig. 7. CNN Model with Batch Size 16 (a) Accuracy, (b) Loss, (c) ROC Curve.

TABLE VII. RESULTS OF EXPERIMENT NO. 1

After	Accuracy/Loss
100 epochs	78.68% / 0.6275
200 epochs	81.15% / 0.4401
300 epochs	82.79% / 0.5441
400 epochs	78.68% / 0.5717
500 epochs	77.05% / 0.5197

TABLE VIII. RESULTS OF EXPERIMENT NO. 2

After	Accuracy/Loss
100 epochs	85.47% / 0.4114
200 epochs	87.18% / 0.4993
300 epochs	84.62% / 0.3868
400 epochs	84.62% / 0.6148
500 epochs	85.47% / 0.4337

TABLE IX. COMPARISON OF DIFFERENT BATCH SIZES

Batch Size	8	16
Confusion Matrix	[[14 27] [1 80]]	[[22 16] [1 78]]
Train Accuracy	79.38%	87.31%
Accuracy	77.05%	85.47%
Specificity	34.14%	57.89%
Sensitivity	98.76%	98.73%
Precision	74.76%	82.97%
Recall	98.76%	98.737%
F1-score	85.106	90.17

VIII. CONCLUSION

Alternaria porri is the fungus that causes purple blotch on onion crops. Leek, garlic, and chives are also affected by this fungus. On onion crops, spores germinate and create a thin water-soaked spot that turns brown. On several other crops, the oval-shaped lesion enlarges turns purplish, and forms the target spot appearance that Alternaria is known for (like early tomato blight). A yellow zone could encircle the margin. During wet weather, the surface of the lesion may be coated in brown to black masses of fungal spores. Infections in plants are a significant risk to worldwide food supplies. Extreme diseases in plants result in annual agricultural yield losses. The identification of diseases in plants at an initial stage is therefore very necessary for the avoidance of such dramatic losses in the future. This paper demonstrates the technical probability of deep learning to allow automatic infection analysis by image classification using the CNN approach. A new approach was discussed in this paper to use deep learning methods to spontaneously identify and detect crop disease from an image. The model established was able to differentiate between healthy and infected crops which can be diagnosed visually. The complete process was defined from the selection of images used for validation and training to the augmentation of images lastly, the deep CNN training procedure. We summarized the final results and concluded that through deep learning detection, segmentation, and classification our improved InceptionV3 achieves the highest precision as well as accuracy, F1-score, and recall. A deep CNN is accomplished to identify onion crops with a classification accuracy of 77.05% for batch size 8 and 85.47% for batch size 16 using our dataset of onion crops. Ongoing work aims to create a complete framework for crop monitoring. Besides, the performance can be improved by using a large dataset for biomedical image segmentation more advanced feature extraction techniques based on deep learning will be developed.

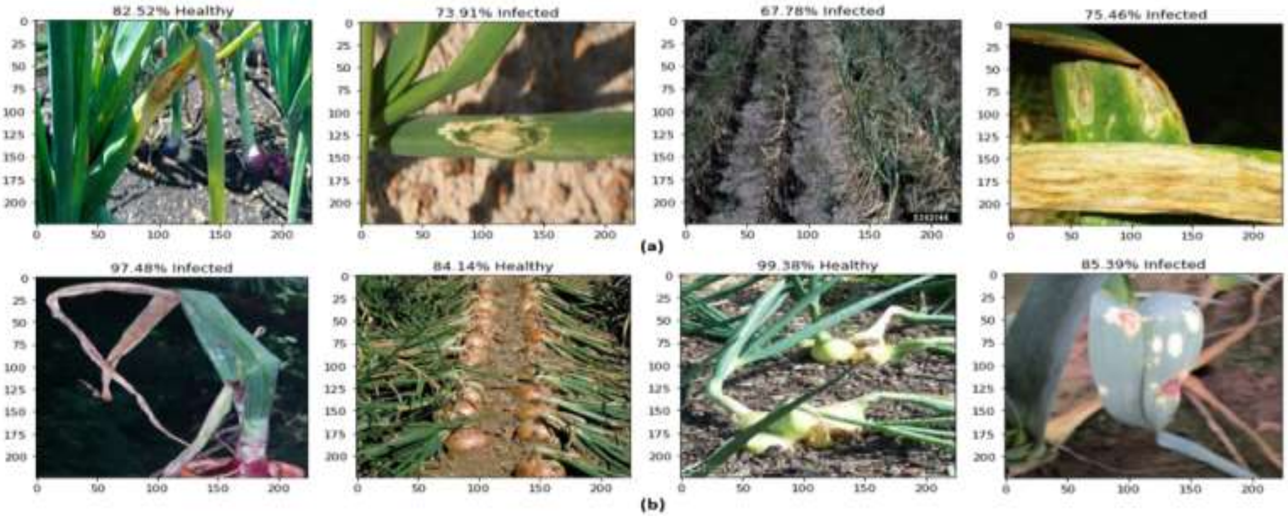


Fig. 8. Predicted Results from (a) Batch Size 8, (b) Batch Size 16.

IX. Future Directions

1) Diseases and pests manifest themselves in different ways at different stages of growth. As a result pest and disease images should be differentiated with greater caution diseases, and pests of the same class should be divided according to the growth cycle as a rule. The division of the dataset will be improved in the future.

2) The next step will be to collect a huge number of high-quality images of diverse types of diseases and pests refine and adapt the model and extend it to other crops/plants to improve the practicability.

3) The disease detection efficiency of the enhanced InceptionV3 model will be improved even further. The goal of early diagnosis and detection of onion diseases can be accomplished by capturing lesions in real-time detection.

Acknowledgment

The authors are extremely grateful to Dr. Imtiaz A. Nizamani, Department of Plant Protection, Sindh Agriculture University, Tandojam, Pakistan for helping in the collection of the onion dataset and their valuable clinical assessment throughout the study. We would like to thank Dr. WajeEha SidDiqui (WS), Peoples University of Medical and Health Sciences For Women, Nawabshah, Pakistan for her assistance and for providing us with the resources we needed to complete this study.

References

[1] M. Jhuria, A. Kumar, and R. Borse, "Image processing for smart farming: Detection of disease and fruit grading," in *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, 2013, pp. 521–526.

[2] S. Hena, L. Jingdong, A. Rehman, and O. Zhang, "A comparative analysis of agricultural development and modernization between China and Pakistan," *Int. J. Adv. Appl. Sci.*, vol. 6, no. 4, pp. 81–94, 2019.

[3] Y. Liu, A. Amin, S. F. Rasool, and Q. U. Zaman, "The Role of Agriculture and Foreign Remittances in Mitigating Rural Poverty: Empirical Evidence from Pakistan," *Risk Manag. Healthc. Policy*, vol. 13, p. 13, 2020.

[4] M. R. Faiz *et al.*, "Identification and prioritization of issues in growing and marketing vegetables in Punjab Province of Pakistan," *J. Innov. Sci.*, vol. 6, no. 1, pp. 54–59, 2020.

[5] A. Maqbool, S. A. Javeed, and A. Adeel, "Marketing of Carrot: A Case Study of Punjab Province, Pakistan."

[6] S. Abbas and Z. A. Mayo, "Impact of temperature and rainfall on rice production in Punjab, Pakistan," *Environ. Dev. Sustain.*, pp. 1–23, 2020.

[7] M. U. Shahbaz, T. Mukhtar, and N. Begum, "Biochemical and serological characterization of Ralstonia solanacearum associated with chilli seeds from Pakistan," *Int. J. Agric. Biol.*, vol. 17, no. 1, 2015.

[8] A. Hussain, M. A. Khathian, S. Ullah, and S. Ullah, "Increasing Crop Profitability through Adoption of Ridge Planting of Rice Crop in Punjab, and Wheat & Banana Crops in Sindh Province of Pakistan," *Proc. Pakistan Acad. Sci. B. Life Environ. Sci.*, vol. 56, no. 4, pp. 25–38, 2019.

[9] Hyesun Jang "Low Onion Prices of Major Asian Players, New Sources for European Markets" 2020. [Online]. Available: https://www.tridge.com/stories/low-onions-prices-of-major-asian-players-new-sources-for-european-markets

[10] C. B. T. Pal and G. C. Jadeja, "Microwave-assisted deep eutectic solvent extraction of phenolic antioxidants from onion (Allium cepa L.) peel: a Box–Behnken design approach for optimization," *J. Food Sci. Technol.*, vol. 56, no. 9, pp. 4211–4223, 2019.

[11] A. Nigussie, Y. Kuma, A. Adisu, T. Alemu, and K. Desalegn, "Onion production for income generation in small scale irrigation users agropastoral households of Ethiopia," *J. Hortic.*, pp. 1–5, 2015.

[12] L. Rajendran, "SCIENTOMETRIC ANALYSIS OF ONION (ALLIUM CEPA L) DURING 1980-2019: A STUDY BASED ON CAB DIRECT."

[13] G. Ekşi, A. M. G. Özkan, and M. Koyuncu, "Garlic and onions: An eastern tale," *J. Ethnopharmacol.*, p. 112675, 2020.

[14] N. B. Rathod and R. N. K. A. S. Karle, "Effect of organic, inorganic and integrated nutrient management on yield attributes, production of onion and soil properties, under vertisol of Maharashtra," *IJCS*, vol. 8, no. 1, pp. 937–940, 2020.

[15] C. Long, K. Hammer, and Z. Li, "The Central Asiatic region of cultivated plants," *Genet. Resour. Crop Evol.*, pp. 1–17, 2020.

[16] G. Messina, J. M. Peña, M. Vizzari, and G. Modica, "A Comparison of UAV and Satellites Multispectral Imagery in Monitoring Onion Crop. An Application in the 'Cipolla Rossa di Tropea'(Italy)," *Remote Sens.*, vol. 12, no. 20, p. 3424, 2020.

[17] H. Tamiru Geneti, "The response of onion (allium cepa l.) to applied water levels under pot planting at mehoni, raya valley of Ethiopia." Hawassa University, 2020.

[18] J. L. Brewster, *Onions and other vegetable alliums*, vol. 15. CABI, 2008.

[19] K. P. S. Kumar, D. Bhowmik, B. Chiranjib, and P. Tiwari, "Allium cepa: A traditional medicinal herb and its health benefits," *J. Chem. Pharm. Res.*, vol. 2, no. 1, pp. 283–291, 2010.

[20] S. Brankovic *et al.*, "Comparison of the hypotensive and bradycardic activity of ginkgo, garlic, and onion extracts," *Clin. Exp. Hypertens.*, vol. 33, no. 2, pp. 95–99, 2011.

[21] A. Kaur, T. G. Singh, S. Dhiman, S. Arora, and R. Babbar, "NOVEL HERBS USED IN COSMETICS FOR SKIN AND HAIR CARE: A REVIEW," *Plant Arch.*, vol. 20, no. 1, pp. 3784–3793, 2020.

[22] R. A. Baloch *et al.*, "Economic analysis of onion (Allium cepa L.) production and marketing in District Awaran, Balochistan," *Econ. Anal.*, vol. 5, no. 24, 2014.

[23] H. D. Lohano and F. M. Mari, "Spatial price linkages in regional onion markets of Pakistan," *J. Agric. Soc. Sci.*, vol. 1, pp. 318–321, 2005.

[24] M. A. Khokhar, K. M. Khokhar, and J. M. Khan, "WATER REQUIREMENT OF ONION CROP IN PAKISTAN."

[25] L. Black, K. Conn, B. Gabor, J. Kao, and J. Lutton, "Onion Disease Guide," *Seminis*, p. 71, 2012.

[26] S. Walker, N. Goldberg, and C. Cramer, "Onion Diseases in New Mexico," p. 12, 2014.

[27] L. Black, K. Conn, B. Gabor, J. Kao, and J. Lutton, "Onion Disease Guide," *Semin. grow Forw.*, p. 71, 2012.

[28] F. Nihar, N. N. Khanom, S. S. Hassan, and A. K. Das, "Plant Disease Detection through the Implementation of Diversified and Modified Neural Network Algorithms," *J. Eng. Adv.*, vol. 2, no. 1, pp. 48–57, 2021.

[29] S. K. B. Sangeetha, M. Sudha, R. Balamanigandan, and V. P. G. Pushparathi, "Comparison of Crop Disease Detection Methods - An intensive analysis," vol. 58, no. 2, pp. 10540–10546, 2021.

[30] Z. Liu *et al.*, "Improved Kiwifruit Detection Using Pre-Trained VGG16 with RGB and NIR Information Fusion," *IEEE Access*, vol. 8, no. January, pp. 2327–2336, 2020.

[31] W.-S. Kim, D.-H. Lee, and Y.-J. Kim, "Machine vision-based automatic disease symptom detection of onion downy mildew," *Comput. Electron. Agric.*, vol. 168, p. 105099, 2020.

[32] R. Sharma, S. Das, M. K. Gourisaria, S. S. Rautaray, and M. Pandey, "A Model for Prediction of Paddy Crop Disease Using CNN," in *Progress in Computing, Analytics and Networking*, Springer, 2020, pp. 533–543.

[33] R. Karthik, M. Hariharan, S. Anand, P. Mathikshara, A. Johnson, and R. Menaka, "Attention embedded residual CNN for disease detection in tomato leaves," *Appl. Soft Comput.*, vol. 86, p. 105933, 2020.

[34] G. Pattnaik, V. K. Shrivastava, and K. Parvathi, "Transfer Learning-Based Framework for Classification of Pest in Tomato Plants," *Appl. Artif. Intell.*, vol. 34, no. 13, pp. 981–993, 2020.

[35] M. Francis and C. Deisy, “Mathematical and Visual Understanding of a Deep Learning Model Towards m-Agriculture for Disease Diagnosis,” *Arch. Comput. Methods Eng.*, pp. 1–17, 2020.

[36] W. Koehrsen, “Transfer Learning with Convolutional Neural Networks in PyTorch” 2018. [Online]. Available: https://towardsdatascience.com/transfer-learning-with-convolutional-neural-networks-in-%09pytorch-dd09190245ce.

[37] G. Huang, Z. Liu and N. Van Der Maaten, Densely Connected Convolutional Networks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, arXiv:1602.07261, 2016.

[39] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv: 14091556, 2014.

[40] Q. Zhang, H. Wang, S. W. Yoon, D. Won, and K. Srihari, Lung Nodule Diagnosis on 3D Computed Tomography Images Using Deep Convolutional Neural Networks, Procedia Manufacturing, Vol. 39, pp. 363-370, 2019.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510-4520, 2018.

[42] K. He, X. Zhang, S.Ren and J. Sunet, Deep Residual Learning for Image Recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going Deeper with Convolutions, Technical report, 2014.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the Inception Architecture for Computer Vision, arXiv preprint arXiv:1512.00567, 2015.

[45] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017.

[46] M. A. Zaki, S. Narejo, S. Zai, U. Zaki, Z. Altaf, and N. U Din, “Detection of nCoV-19 from Hybrid Dataset of CXR Images using Deep Convolutional Neural Network,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 699–707, 2020. *doi: 10.14569/ijacsa.2020.0111281.*

Design of Multi-band Microstrip Patch Antennas for Mid-band 5G Wireless Communication

Karima Mazen[1]
Department of Electronics and Computer
Thebes Higher Institute for Engineering
Maaddi, Cairo, Egypt

Ahmed Emran[2], Ahmed S. Shalaby[3], Ahmed Yahya[4]
Department of Electrical Engineering
Al-Azhar University, Nasr City
Cairo-11371, Egypt

***Abstract*—Recently, the best antenna structures have considered microstrip patch antenna due to their simple construction, low cost, minimum weight, and the fact that they can be effortlessly integrated with circuits. To achieve multi-band operation an antenna is designed with an etching rectangle and circle slot on the surface of the patch to achieve multi-band frequency capabilities in mid-band 5G applications. Inset-fed structure type of fed of all antenna printed and fabricated on the brow of the Rogers RT5880 substrate. Then, prototype structures of the microstrip patch antenna were acquired during the design process until achieving the desired antennas. The antenna_1 achieved tri-band characteristics covering the WiMAX band including 2.51 – 2.55 GHz, WLAN, and S-band including 3.80 – 3.87 GHz and C-and X-band including 6.19 – 6.60 GHz. The antenna_2 gives dual-band characteristics covering C-band and X-band including (6.72 – 7.92 GHz) with a peak under -45 dB suitable for mid-band 5G applications. High impedance bandwidth increases between (70 MHz-1.25 GHz) for wireless applications. The proposed microstrip patch antennas were simulated using CST MWS-2015 and were experimentally tested to verify the fundamental characteristics of the proposed design, it offers multiple-band operation with high stable gain and good directional radiation characteristics results.**

***Keywords*—Band-width; microstrip; multi-band; notch slot; rectangle slot; 5 G**

I. Introduction

Generally, the microstrip patch antenna is an important component of communication systems that require characteristics such as compact size, lightweight, easy process of fabrication, and wide bandwidth. The microstrip patch antenna is primarily made of copper material or a perfect electric conductor (PEC). There are various types of geometries like a circular, rectangular, triangular, elliptical, square, ring, cone, etc. Nevertheless, the most commonly used shapes are rectangular and circular. The size of the patch antenna depends upon the substrate constant dielectric material (Ɛr). A Higher substrate dielectric constant leads to the lower size of the antenna [1][2]. The 1G /2G/3G/4G and Five generations (5G) introduce faster data rates, density connection higher, low latency [3]. Compact patch feeds square microstrip with right-angled isosceles Koch fractal antenna geometry on the edges is suitable for U-wide frequency band applications. The antenna is put on FR4-epoxy, (Ɛr = 4.4) substrate with dimensions 60 x 55 x 1.59 mm3. The antenna works at 4.3, 5.0, 6.1, 7.4, 8.9, and 9.2 GHz, this design is limited in bandwidth and gain, therefore required to strive to improve them [4]. Propose the patch antenna in the form of a Sierpinski fractal antenna, which can work in multi-band frequency, this design is limited in bandwidth and gain, therefore required to strive to improve them [5]. The designed microstrip antenna using CPW feeding technique, the microstrip antenna operates resonates at this is bands 0.45 GHz for GSM, 1.35, 1.92, 2.57 GHz for, WLAN, WiMAX, Walkie-Talkie. The antenna dimension is 46.32×25×1.6 mm3, substrate type is FR4 with Ɛr=4.3. The design is limited in bandwidth and gain, therefore required to strive to improve them [6]. The proposed microstrip is fabricated using an FR-4, the height of the substrate is 1.6 mm, and Ɛr =4.4. The ground plane and substrate dimension are the same, i.e. (70×60 mm2), the proposed microstrip antenna operates at desired frequencies 5.73, 1.8, 3.6, and 4.53 GHz, which can be used for many applications of the wireless [7]. Design patch antenna for satellite applications and operating frequency at 15 GHz with diverse slots cutting on microstrip antenna. The antenna achieving bandwidth frequency at 1.14 GHz, with S11 of -30.6 dB. The gain and directivity are 3.488, and 3.544 dB respectively. The RT duroid material size 9.5 x 8 x 1.6 mm3 to design microstrip antenna with Ɛr =2.2 have loss tangent 0.0009 [8]. The microstrip patch is fabricated and tested on substrate FR-4 with size 70×70×1.6mm3, Ɛr= 4.4. The choice bands are designed at 2.478, 2.313, and 2.396 GHz, and they have −10 dB impedance bandwidths of 2.42, 2.14, and 2.50%, respectively [9]. The microstrip patch antenna in the shape of the rectangle there is an array of L-slots and inverted them slot for multiband frequency is considered. The microstrip was designed for TM010 mode of frequency operation at 2.1 GHz. The configuration produced the best results at Penta frequency bands:1.48, 1.25, 1.8, 2.25, and 2.9 GHz. And they have -10 dB impedance bandwidth 3.4%, 3.2%, 3.33%, 4.3%, and 3.1%, respectively [10] All the design there is limited in bandwidth and gain, therefore required to strive to improve them. Design a simple microstrip antenna (UWB) for the wireless satellite. The design process is achieved in three steps. Firstly, a conventional antenna for 2.4 GHz FR4 substrate with Ɛr = 4.4 has been defined with dimensions of 50 × 55 × 1.6 mm3. Bandwidth is achieving between (2- 9.7 GHz) for different wireless applications [11]. Design microstrip a slotted pentagonal for multi-band applications. Designed on substrate material textile type with a dielectric constant is 4.4. Dimension size is 50 × 50 mm2 and thickness (h= 1 mm). The resonate frequency at 4 GHz with S11 is -31.84 dB. The antenna is achieved more directivity of 2.932

dBi and bandwidth from 2.85 to 9.30 GHz [12]. Simulated and fabricated a rectangular patch is introduce designed for 915 MHz band applications includes ZigBee and Bluetooth. A printed antenna on FR-4 epoxy substrate with a small size 60 × 30 × 1.6 mm3, with Ɛr =4.4. Achieving bandwidth from 902 MHz to 925 MHz for the ZigBee band applications. The results of the simulation are radiation pattern directivity is 2.83 dBm, the gain is 2.73 dBm, and return loss is -35 dB [13]. Design a novel broadband microstrip antenna printed on the FR-4 substrate with Ɛr =4.4. The simulated result of the patch gives four resonance frequencies at 1.59, 1.71, 1.81, and 1.986 GHz respectively, achieve a bandwidth of 29.6 % [14]. The microstrip U patch antennas on substrate Alumina 60 × 60 × 12.5 mm3 with dielectric constant Ɛr = 9.6 for wireless communication. The improvement in antenna bandwidth is recorded at 20 MHz, and efficiency 99.6%. The antenna operating frequency of the antenna is 3.8 GHz (3.06 –5 GHz), the antenna operates at single-band [15]. Design of a two-band patch using substrate PTFE, with a thickness of 0.8 mm, the dielectric constancy is 2.55. The patch antenna works at dual-band 9.96 GHz –12.84 GHz and the bandwidth of the antenna is 270 and 550 MHz respectively. The simulation result gives less value of S11 at f1 = 9.96 GHz and f2 = 12.84 is -33.6 and -39.1dB respectively. The bandwidth of the patch is 9.83 GHz and 12.57 GHz respectively [16]. Design of packaging of single antenna fed asymmetric slot-loaded triple-band patch antenna with LHCP and RHCP at two bands at (1.77645 GHz) for GPS L5 and (SDARS) at (2.320-2.345 GHz). The dimensions of the designed antenna 51 × 52 × 1.6 mm3 with substrate FR-4 material, and Ɛr = 4.4 [17]. Design a monopole antenna and implement them to operate at multi-band for 5G wireless communications and service. The antenna substrate dimension is 43 × 38 ×1.588 mm3 with dielectric constancy 4.4. The S11 parameter of the patch antenna is high than -20 dB, and VSWR is < 2 [18]. The design of the microstrip antenna is based on substrate RT/Duroid 5880 material with relative permittivity Ɛr =2.2. The patch designed and fabricated to operate in X-band is (8-12 GHz) and 60 GHz frequencies. The microstrip antenna was printed on two substrate heights (0.75- 1.57 mm) operate at 10 GHz and, (0.127 - 0.254 mm) for frequency band 60 GHz. Designing approach satisfied wideband and high gain antenna The antenna work at dual-band, but required to improve in bandwidth and achieve best reflection coefficient (S11) [19]. The microstrip patch was put on Duroid 5870 substrate h= 1.575 mm, a Ɛr of 2.33, and a dielectric loss of 0.0012, dual-band to cover K-band applications [20]. Design patch antenna printed on Rogers RT5870 with Ɛr = 2.2, and operate at 2.4 GHz. The thickness of 0.787 mm for applications such as IEEE 802.15.1 Bluetooth, ZigBee, WiFi, wireless USB. Bandwidth achieved 25.5 MHz with return loss -22.5 dB. The design is limited in operating at single band, bandwidth, and gain, therefore required to strive to improve them [21].

In this work, a new design approach is objective to realize tri-band and dual-band MPAs appropriate for 5G wireless applications. Two Compressed simple microstrip patch antennas (MPAs) have been designed to cover multiple frequency bands for wireless applications. The design idea of these microstrip patch antennas is almost based on rectangular patch antenna and are namely rectangular patch etching two rectangle slots antenna (antenna_1), and notch rectangle patch etching circle slot (antenna_2). All these antennas are fabricated on Roger RT5880 substrate and they are built the model and analyzed by using the tool of the CST MWS simulator [22]. The CST simulated result exposes that the proposed microstrip antennas are designed to guarantee the best performance results. In terms of resonant frequency bands and directional patterns as well as high gains, improvement impedance bandwidth, and total radiations efficiencies. The rest content of this work is orderly as follows. In section II, discussion the configurations of the proposed patch and the analysis process for designing these microstrip patches are offered in detail. Section III and IV are review discussion the performance of the experimental and simulated results along with the features of the designed patch antennas and parametric study. Conclusions and future work are drawn in Section V and VI.

II. Proposed Patch Antenna Design and Configuration

The simulation models of the two proposed design microstrip antenna structures are in Fig. 1(b) and Fig. 2(b), this is antennas, called the rectangular antenna etching two rectangle slots (antenna_1) etching on the reference (conventional) antenna (RA), and notch rectangle antenna with circle slot (antenna_2) by etching rectangle notch on corner of reference antenna and cutting circle slot on the middle of the patch antenna. The slot on the microstrip antenna is analyzed. The slot on the microstrip patch can be firm by using a duplicity relationship between the dipole and the slot [23] [24] [25]. The fabricated antennas are printed on the front side of the Roger RT5880 substrate. The height hs of 3.18 mm, relative permittivity Ɛr=2.2. The constant Ɛr of microstrip material should be between 2.2 and 12 for antenna designing [1]. The height of the substrate, h << λ0 (where, λ0 equal operating wavelength) [1]. The resonant frequency fr=2.4 GHz. The total substrate size (Ws x Ls) of all antennas is (94 × 78mm2). Each patch structure of these antennas size of patch (Wp) x length (Lp) of (47 × 38 mm2). The inset feeding type is used for the design because of its ease of fabrication in the PCB form, and easy matching with the existing system. The proposed microstrip antennas are configured to improve impedance bandwidth outcomes due to changes in substrate height and dielectric constancy. If the substrate height is increased the bandwidth of the antenna is also increased. This is because the bandwidth of the antenna is directly proportional to the substrate height. The feed width wf and length Lf, while the characteristic impedance is 50 Ω. A full ground plane is on the backside of the substrate material. Copper is used as the conducting material for patches and ground. The dimensions of the proposed microstrip antenna are calculated by using the well-known microstrip patch antenna formulas using this equation from (1) -(5) [1].

The width of the microstrip patch antenna (w_P) is given by the following equation [1].

$$w_P = \frac{c}{2 f_r \sqrt{\frac{\varepsilon_r + 1}{2}}} \tag{1}$$

Sub. ε_r=2.2, c=3x10^8m/sec, f_r= 2.45GHz

The effective dielectric constant ε_{reff}:

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + 12\frac{h}{w_p}\right)^{\frac{-1}{2}} \tag{2}$$

Effective length due to fringing effects:

$$\Delta L = 0.412h \frac{(\varepsilon_r + 0.3)\left(\frac{w_p}{h} + 0.264\right)}{(\varepsilon_{reff} - 0.2680.3)\left(\frac{w_p}{h} + 0.8\right)} h \tag{3}$$

The effective length of the patch (L_{eff}):

$$L_{eff} = \frac{c}{2 f_0 \sqrt{\varepsilon_{reff}}} \tag{4}$$

The actual length of the patch:

$$L_p = L_{eff} - 2\Delta L \tag{5}$$

To determine the ground plane dimensions length (L_g) and width of the ground plane (w_g)

$$L_g = 6h + l_p \tag{6}$$

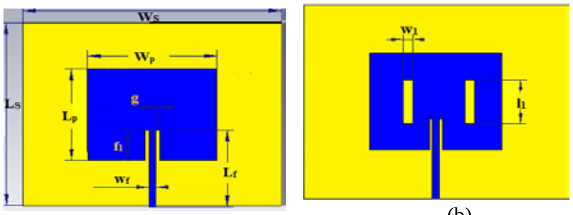
$$w_g = 6h + w_p \tag{7}$$

Where W_p the width of the microstrip patch, f_r = 2.45GHz is the resonant frequency, ε_r =2.2 is the dielectric constant of the substrate material, h_s =3.18 mm is the height of substrate material. The patch antenna L_p= 38 is the length of the patch and w_p= 47 mm is the width of the patch. After many series of optimization by using CST simulator, then, the final parameters and the optimized value of the parameters are illustrated in Table I.

Reign successively conventional patch antenna (MPA) was designed founded and based on the equations (1) -(5) for resonating frequency at the proposed frequency. Initially, MPA is designed with the same geometrical parameters. The obtained CST simulated reflection coefficient(S_{11}) and radiation pattern realized gain of reference antenna (RA) shown in Fig. 1(a) or 2(a) and 3, an antenna is resonating at f_r = 2.45 GHz with -10 dB S_{11} impedance bandwidth 70 MHz ranging from (2.43 – 2.50 GHz) for (WiFi), return loss is -23 dB, and radiation pattern of the antenna. It is demonstrated from observations in Fig. 4 that the RA inset-feeding technique with a full ground plane has a better gain of 7.31 dB and, a boresight directional radiation pattern suitable for wireless communication application.

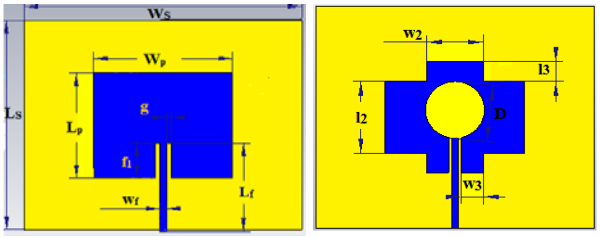
TABLE I. DIMENSIONS OF PROPOSED ANTENNAS

Dimension	W_s	L_s	L_f	f_1	h	G	w_1
mm	94	78	32	12.7	3.18	1.0	3.4
Dimension	t	l_1	l_2	D	l_3	w_2	w_3
mm	0.07	17.5	25.20	18.4	9.60	19	7.35



(a) (b)

Fig. 1. The Proposed Microstrip Patch Structures, (a) AR (b) Antenna_1.



(a) (b)

Fig. 2. The Proposed Microstrip Antenna Structures, (a) RA (b) Antenna_2.

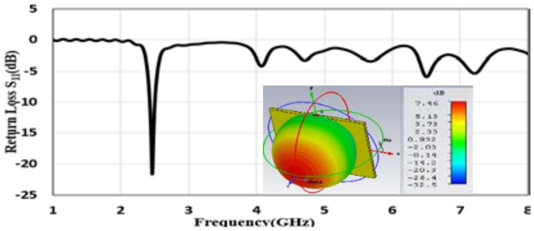
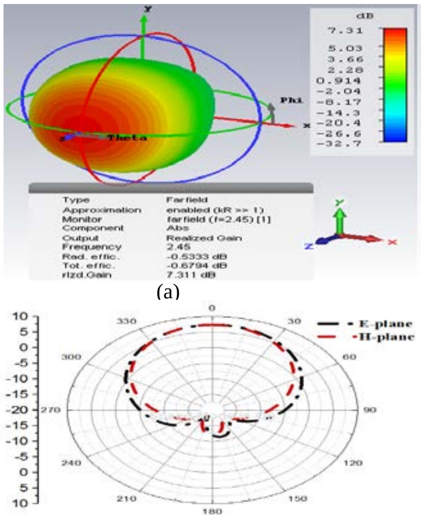


Fig. 3. S_{11} Plot and Radiation Pattern for the Reference Antenna.



Type Farfield
Approximation enabled (kR >> 1)
Monitor farfield (f=2.45) [1]
Component Abs
Output Realized Gain
Frequency 2.45
Rad. effic. -0.5333 dB
Tot. effic. -0.6794 dB
rlzd.Gain 7.311 dB

(a)

b

Fig. 4. Far-field Realized Gain Radiation Pattern of RA at 2.45 GHz. (a) 3D. (b) 2D in yz- and xz-plane.

TABLE II. MICROSTRIP PATCH ANTENNA (MPA) PERFORMANCE PARAMETERS

Parameter	Inset-feed with full ground plane
f_r (GHz)	2.45
f_L (GHz)	2.43
f_H (GHz)	2.5
BW(MHz)	70
Gain(dB)	7.31

Table II summarizes the main performance parameters of this reference antenna, lower and higher-frequencies, resonance frequency, bandwidth, and gain.

III. DISCUSSION OF SIMULATION RESULTS

After that, the procedure for microstrip patch antenna structures in the previous section is suggested for the proposed antennas for operating in the wireless communication WLAN, WiMAX, C- band, and -X band frequency ranges for mid-band 5G wireless applications [27]. The proposed design concept will be Verified by the main performance parameter results concluded to gain, reflection coefficient (S_{11}), and current distribution on the patch as well as the radiation pattern directivity and realized gain.

A. Reflection Coefficient

The proposed microstrip patch antenna structures are realized in simulated of the CST Microwave Studio ver. 2015. Its time-domain solver is used to obtain these results. The reflection coefficient (S_{11}) for the proposed microstrip antennas, antenna_1 and antenna_2 is discussed and investigated in detail. The results of the return loss of the whole prototype proposed antenna structures have been recapped in Fig. 5 and 6, shows the S_{11} plot for intermediate antennas that belong to each type of the two proposed antennas. Observed that the antenna_1 covers tri-band characteristics, whereas the antenna_2 gives dual-band. The frequency bands of the proposed antennas can be summarized as follows:

1) Antenna_1: Cover WiMAX band 2.53 GHz (2.51 – 2.55 GHz, 1.5%) and WLAN/C-band and S band 3.86 GHz (3.80 – 3.87 GHz, 2%). C-band 6.45 GHz (6.19 – 6.60 GHz, 6%) for 5G application [26].

2) Antenna_2: Cover C band and X band for mid-band 5G service [27], 6.92 and 7.707 GHz (6.72 – 7.92 GHz, 17.5%).

Thus, these microstrip antennas cover the helpful frequency bands that useful for wireless communication.

VSWR results proposed microstrip antenna have been shown in Fig. 7. Their quantitative analysis in terms of f_r, f_l, *and* f_h, between RA and their proposed antennas, is listed in Table III. The realized gain respect to frequency for the proposed microstrip patch antennas at desired frequency bands is observed in the experimental result in section 3. Then the gain of antenna_1 at the lower frequency band is the less one whereas it is the greater one in the high band.

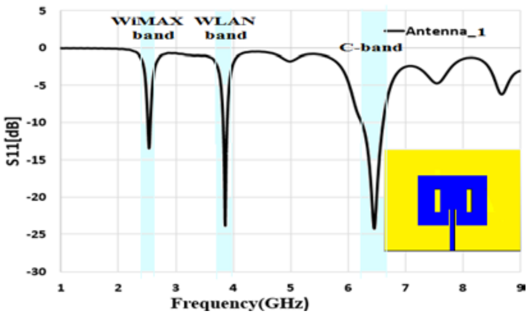


Fig. 5. Simulated Return Losses S_{11} Curves for Antenna_1 at 2.53, 3.86 and 6.45 GHz.

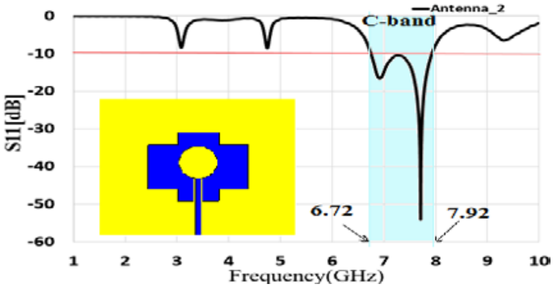


Fig. 6. Simulate Return Loss S_{11} Curve for Antenna_2 at 6.93 GHz and 7.707 GHz.

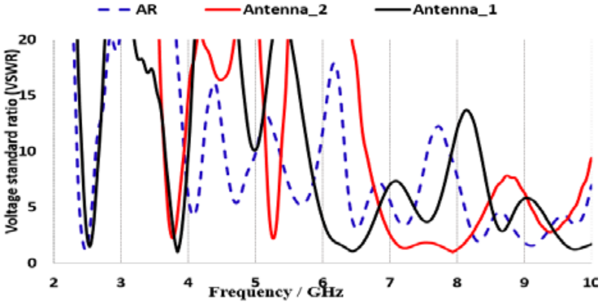


Fig. 7. Simulation VSWR for the Proposed Antennas

TABLE III. THE PROPOSED ANTENNAS TO RA COMPARISON FROM THE BANDWIDTH, (FL- FH), FR PERSPECTIVE IN (GHZ)

Antenna	Band 1	Band _2	Band 3	B.W (MHz)
RA	(2.43– 2.50), 2.45 GHz	------	-------	70
antenna_1 (Tri-band)	(2.51–2.55), 2.53 GHz	(3.80 3 .87), 3.86 GHz	(6.19-6.60), 6.45 GHz	50,70,410
antenna_2 (Dual-band)	------	(6.75-7.92), 6.92 GHz	(6.71-7.94), 7.707 GHz	1250

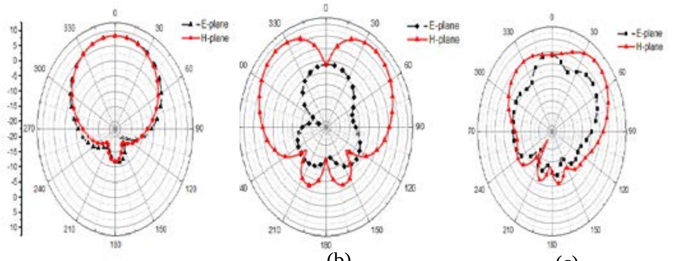
B. 2-D and 3-D Radiation Pattern

Fig. 8 shows the 2-D radiation pattern in terms of E- or yz-plane and H-or xz -plane at band 1, band 2, and band 3 observed at Table III. It is observed from this figure that all antennas have a directive radiation pattern in H-plane and E-plane at the lower band (2.53 GHz), and near to omnidirectional at all other middle and higher frequencies (3.86 GHz & 6.45 GHz) bands. The radiation pattern in E-plan shown in Fig. 8(a) at 2.53 GHz shows close to bidirectional nature with angular width (3) of 75.0 deg., and main lob directional 1.0 deg. Whereas, the radiation pattern at 3.87, 6.45 GHz shown in Fig. 8(b) and Fig. 8(c) shows close to omnidirectional nature with greater angular widths 59.2 and 31.6 degrees respectively. Besides, the main lob directional 4.0 and -5.0 degrees respectively. Therefore, the radiation pattern is changed from bidirectional to omnidirectional when the angular width is reduced.

Fig. 9 shows the 3-D radiation pattern for tri-band frequency (a) for 2.53 GHz, (b) 3.86 GHz, (c) 6.45 GHz.

Fig. 10 shows the 3-D radiation pattern for dual-band frequency (a) for 6.92 GHz, (b) 7.707 GHz.

The simulated results verified that the proposed antennas are achieved multi-band antennas. Table IV shows the simulation results of proposed antennas that have a good radiating element antenna, return loss less than -20 dB, good realized gain, VSWR is less than 2, and good impedance bandwidth.

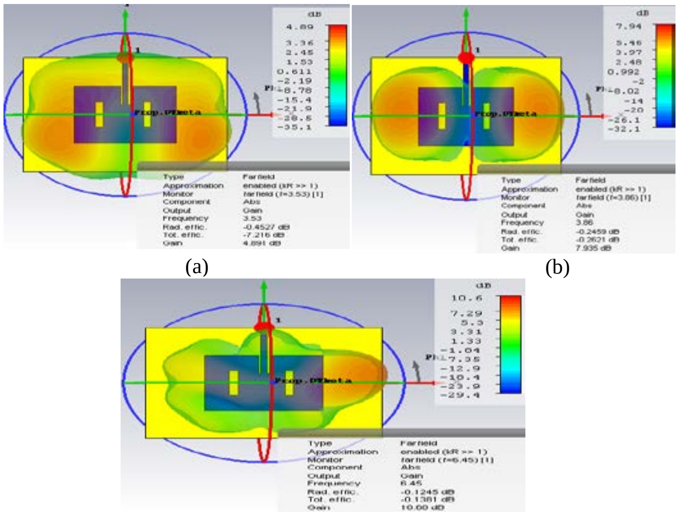


(a)

(b)

(c)

Fig. 8. Simulated E and H-plane for Antenna_1 at (a) 2.53 GHz, (b) 3.87 and (c) 6.45 GHz.



Type Farfield
Approximation enabled (kR >> 1)
Monitor farfield (f=3.53) [1]
Component Abs
Output Gain
Frequency 3.53
Rad. effic. -0.4527 dB
Tot. effic. -7.216 dB
Gain 4.891 dB

(a)

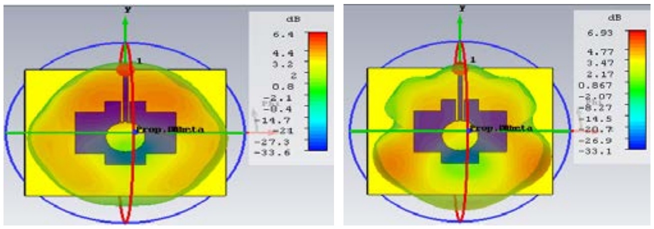
Type Farfield
Approximation enabled (kR >> 1)
Monitor farfield (f=3.86) [1]
Component Abs
Output Gain
Frequency 3.86
Rad. effic. -0.2459 dB
Tot. effic. -0.2621 dB
Gain 7.935 dB

(b)

Type Farfield
Approximation enabled (kR >> 1)
Monitor farfield (f=6.45) [1]
Component Abs
Output Gain
Frequency 6.45
Rad. effic. -0.1245 dB
Tot. effic. -0.1381 dB
Gain 10.60 dB

(c)

Fig. 9. 3D Radiation Pattern Gains of Proposed Antenna_1at (a) 2.53 GHz, (b) 3.86 GHz, (c) 6.45 GHz.



(a)

(b)

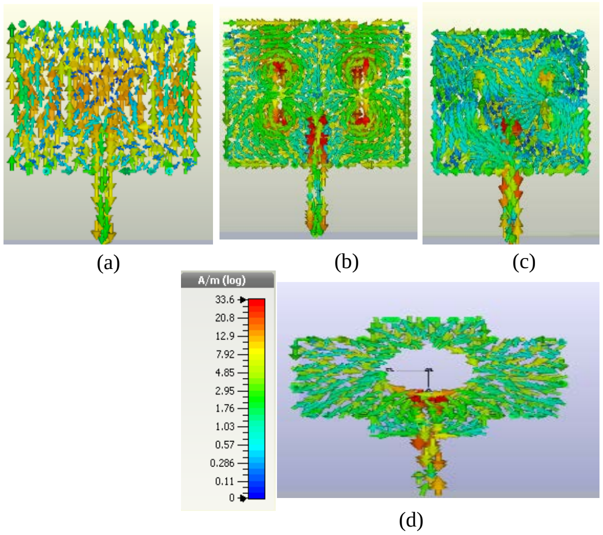
Fig. 10. 3D Radiation Patterns Gain of the Proposed Antenna_2 at (a) 6.92 GHz, (b) 7.707 GHz.

TABLE IV. SIMULATED MULTIBAND ANTENNAS -10 DB BANDWIDTH, GAIN, VSWR, DIRECTIVITY AND IMPEDANCE BANDWIDTH

Results	**antenna_1**			**antenna_2**	
Frequency f_r(GHz)	2.53	3.86	6.45	6.92	7.707
S_{11} dB	-13.38	-23.86	-24.26	-16.92	-55
VSWR	1.4	1.1	1.1	1.3	1.4
Dir (dBi)	8.34	8.22	10.73	6.38	7.28
Gain dB	8.18	7.97	10.6	5.56	6.22
B/W MHz	50	70	410	1250	

C. *Surface Current*

To analyze the effectiveness, of the proposed microstrip patch antenna, the higher modes make nulls and some side effects lobes this is due to the effect of the vertical and horizontal distributions current on the surface slotted microstrip antenna, as the operating band frequency increases. The surface current of antenna_1 shows in Fig. 11, and analyzed at frequency band in Fig. 11(a) shows the surface current distributed have mainly flowed throw feed line part and in a horizontal line on the patch at a lower frequency. However, the path and the direction of the density are entirely changed with the insertion of two rectangle slots. As shown in Fig. 11(b) the current distribution along with two rectangle slot more concentrated interior slot and exterior of two rectangle slot, mainly flows along feed line part. Thus, two rectangle slots modify the surface current distribution to generate resonance at 3.86 GHz. Fig. 11(c) shows the distribution mainly on the feed line part and inside two rectangle slots of the patch in relatively higher frequency (6.45 GHz). From Fig. 11(d), it can be observed the surface current distribution the path, and the direction of the current are entirely changed with the insert of the circle slots. The current distributed concentrated throw feed line under an arc of the circle and its below. The current distributed interior edge of the circle and outside of circle slot, and less current density throw the notch.



(a) (b) (c)

(d)

Fig. 11. Simulated the Current Distribution on Antenna_1 at, (a) 2.53 GHz, (b) 3.86 GHz, (c) 6.45 GHz (d) The Surface Current on Antena_2 at 7.707 GHz.

D. Parametric Study

To reach up with a final design of acceptable performance worked at in-demand frequencies; dense simulations to test the insight of each antenna dimension and slots on its performance have been done using parametric sweep –time-domain solver CST simulator. The antenna_1 is optimized to operate at triple-frequency bands.

Fig. 12 shows the effects of variation on S_{11} characteristics, it indicates that the second resonant at 3.86 GHz is the most affected resonant, and its value increased -29.30 dB as w_1 increased. While the first and third resonance for the patch itself; is not nearly affected. The rectangle length-shaped slot of antenna_1 is l1 changed in two rectangle slots to find its effect on antenna_1.

Fig. 13 shows the effects of variation in l_1 on S_{11} characteristics. The center frequency of the second resonance is increased return loss is -35.29 dB at 3.86 GHz as l_1 increased while decreasing when the value l_1 decreases and the center frequency of the third frequency varies to change according to change the value of l_1. But first band resonance for the patch itself; is not nearly affected.

It can be observed that the impedance fractional bandwidth is wider for the third frequency band at a different value of l_1. Bandwidth 440 MHz (6.17 - 6.61GHz, 6.87%) is resonating at 6.45 GHz achieving a return loss of −27.19 dB is obtained when (l_1 = 18 mm). The effects result of variation in l_1 for impedance bandwidth and S_{11} are tabulated in Table V.

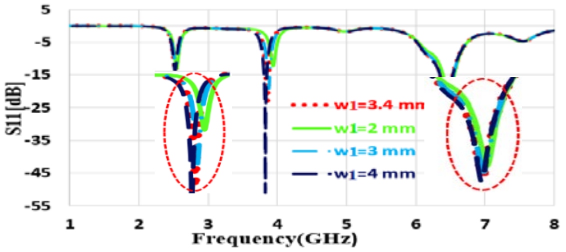


Fig. 12. Simulated S_{11} for Antenna_1 with Variation of width of slot (w_1).

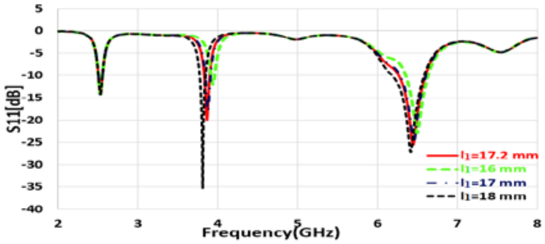


Fig. 13. Simulated S_{11} for Antenna_1 with Variation of l1=16 -18 MM.

TABLE V. COMPARATIVE RESULTS OF ANTENNA_1 VARIED LENGTH L1= 16 - 18 MM

l_1(mm)	Return loss (dB)	Triple resonant freq. (GHz)	Bandwidth MHz
l_1=18	-14.13 -35.29 -27.19	2.53 3.81 6.41	(2.50-2.55 GHz) is 50 MHz (3.77-3.83 GHz) is 60 MHz (6.17-6.61GHz) is 440 MHz
l_1=17.2	-14.13 -17.97 -24.93	2.53 3.87 6.44	(2.50-2.55GHz) is 50 MHz (3.82-3.91GHz) is 90 MHz (6.25-6.61GHz) is 360 MHz
l_1=16	-14.13 -11.63 -22.87	2.53 3.94 6.49	(2.50-2.55GHz) is 50 MHz (3.90-3.91GHz) is 100 MHz (6.35-6.63GHz) is 280 MHz

From Fig. 14, it can be observed the circle slot parameter sweep has also its influences on the antenna_2 performances. The radius slot of the circle (r = D/2) is selected to be presented the effects of variation in r on S_{11} shown in Fig. 14. The first frequency band at 6.92 and 7.707 GHz resonances are which respond to any change in the dimensions of the slots by shifting up/down or degrading/ improving the return losses. The analysis reveals that as the radius of a circle (r =9.2 and 9.25 mm), the wideband characteristic of the antenna represents dual-band at 6.933 and 7.707 GHz the fractional bandwidth increases. While radius value (r = 8.5 and 10 mm) the bandwidth decreases. The maximum bandwidth (1.25 GHz) of 16% (from 6.72 – 7.94 GHz) resonating frequency at 7.707 GHz, given reflection confections, return loss of -55 dB is obtained when r = 9.2 mm and -16.92 dB at 6.93 GHz. After the optimization, the parameter of the proposed antennas has the best bandwidth and return loss for a radius value of r = 9.2 mm. The effects of variation in r on impedance bandwidth and S_{11} are tabulated in Table VI.

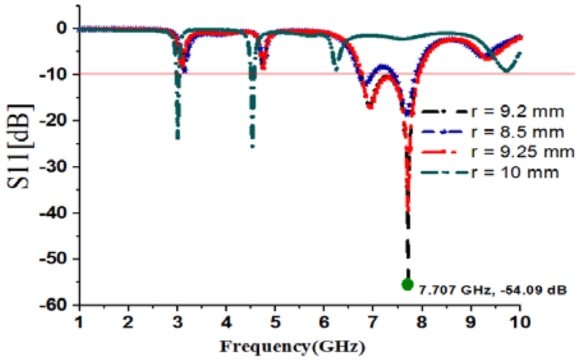


Fig. 14. Simulated S_{11} for Antenna_ 2 for Variation of r.

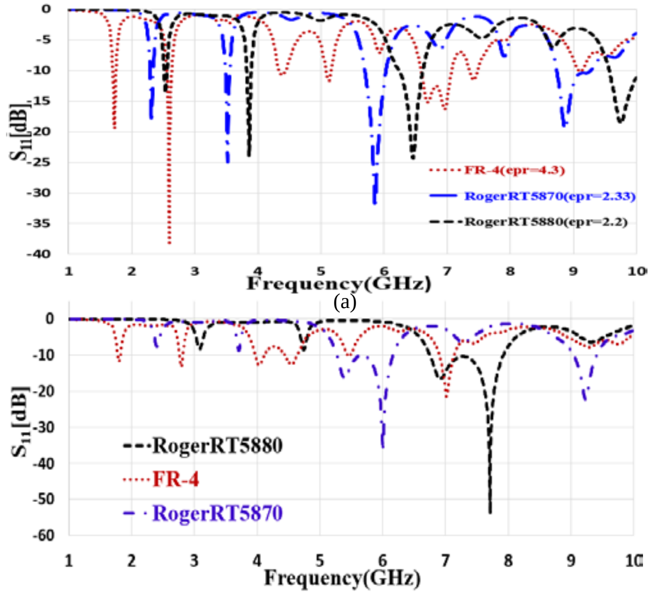
TABLE VI. COMPARATIVE RESULTS OF THE RADIUS OF A CIRCLE(R) FOR ANTENNA_2

Radius of circle	Return loss(dB)	Resonant freq.(GHz)	Bandwidth(MHz), %
r= 8.5 mm	-18.60 -12	7.68 6.83	450 MHz, 6% 260 MHz, 5%
r =9.2 mm	-55 -16.92	7.707 6.93	1250 MHz, 16%
r=9.25 mm	-39.60 -16.92	7.69 6.933	1170 MHz, 16%
r= 10 mm	-23.72 -25.50	2.99 4.53	50 MHz, 2% 50 MHz, 2%

Continue to study the effected of reflection coefficient (S_{11}) concerning frequency for proposed microstrip patch antennas with the different substrate material will be shown in Fig. 15. Observed from this figure, the return loss vs frequency response is verified by applied different substrate materials Roger RT5880, FR-4, and Roger RT5870, without changing the dimensions of slots. Whereas, after optimization and from a results acts comparison between them, it is that Roger RT5880 material realized best results of reflection coefficient (dB) and better wider impedance bandwidth respect to these of the other substrate materials used.

E. Comparison with other Studies

From Table VII, it can be observed analysis comparison between the proposed multiband rectangular microstrip patch and other rectangular antenna structures over previous multiband patch antenna design, such as [4], [5], [6], [7], [8], [9], [10], [15], [19] and [21]. The design recorded in [4-5-6-7-9-10] the patch antenna work in multi-band frequency, but band-width and gain less than compared with the proposed design. Reference [8-21] in this works antenna operates in single-band therefore, bandwidth, gain and reflection coefficient (S_{11}) less than comparing with multiband frequency proposed antenna. The design recorded in [19] the patch operates in dual-band compared with proposed multi-band antenna. The objective of this work to design a microstrip patch antenna that operates at multi-band realized high gain and enhanced bandwidth for wireless applications.



(a)

(b)

Fig. 15. Simulated S_{11} for Various Substrate Materials for given Microstrip Antennas (a) Antenna_1, (b) Antenna_2.

TABLE VII. A COMPARISON BETWEEN THE PROPOSED ANTENNAS WITH OTHER REFERENCE ANTENNAS

Ref.	No. of bands	Sizes(mm)	Resonant frequency(GHz)	Bandwidth (MHZ)	Gain (dB)
[4]	Multi-band	60 × 55 ×1.59	4.3, 5.0, 6.1, 7.4, 8.9, 9.2	68.6, 126.7, 132, 124.3, 191.2, 530.6	1.08, 3.23, 3.36, 2.77, 3.07, 4.87
[5]	Multi-band	70×70×1.58	1.75, 3.65, 4.12, 5.55, 6.5, 7.77	170, 60, 110, 120, 140	7.2, 11.2, 11.3, 7,
[6]	Multi-band	46.32×25×1.6	0.45, 1.35, 1.92, 2.57	185, 151, 77, 218	4.484, 2.59, 3.27, 4.39
[7]	Multi-band	70×60×1.6	1.81, 3.6, 4.53, 5.73	70, 290, 680	5.71, 5.54, 5.01, 5.32
[8]	Single-band	9.5 × 8 × 1.6	15	1140	3.44
[9]	Multi-band	70×70×1.6	2.313, 2.396, 2.478	50, 60, 60	1,1.2,0.7
[10]	Multi-band	33.7×33.7×1.6	1.25, 1.48, 1.8, 2.25, 2.9	3.2%, 3.4%, 3.33%, 4.5%, 3.1%	1.1, 1.12, 1.15, 1.39, 1.4
[15]	Single-band	60×50× 5	3.8	UWB	---------------
[19]	Dual-band	29.52×34.35×1.57	10 , 60	384	13.5, 13
[21]	Single-band	47×39×0.787	2.4	25.5	6.65
Proposed antennas	Multi-band	47×38×3.18	2.45, 2.53, 3.86, 6.45 6.93, 7.707	70, 60, 70, 410 1420	7.31, 8.18, 7.97, 10.6, 5.56, 6.22

IV. Experimental Results

The fabricated prototype of two proposed antennas, the first antenna has a tri-band and the second has dual-band frequencies. The proposed patch antennas are fabricated and measured to verify the performance of the proposed multi-band microstrip antenna printed on the dielectric constant of 2.2 of Rogers RT5880 substrate with a loss tangent of 0.025 and a 3.18 mm thickness. a copper layer thick of 0.07mm on each side for the patch and ground plane for the proposed microstrip antennas. The front side view and measurement of the proposed antennas are shown in Fig. 16.

The prototype model structure is fabricated at the national telecommunication institute (NTI) Cairo Egypt, while the experimental verification of the S_{11} results and the far-field measurements are carried out using the anechoic chamber at Microwave lab, Ain-Shams University, Cairo Egypt.

The measured return loss(S_{11}), VSWR, realized gain, directivity, and far-field results of antenna_1 and antenna_2 are realized in good agreement with simulation results illustrated in Fig. 17. Fig. 17(a) shows the measured return loss(S_{11}) exhibits good tri-band frequency response, and experimental results verified multi-band frequency. The measured S_{11} and impedance bandwidth is shown in Table VIII.

The measured impedance bandwidths for antenna_1 are shown in Fig. 17 (a) as (2.36 -2.477 GHz, 5%) at the 2.402 GHz, (6.32-6.63 GHz, 4%) at the 6.55 GHz band, and (7.077-7.387 GHz, 4.5%) at the 7.25 GHz band.

Fig. 17(b) shows S_{11} of antenna_2 which exhibits a good dual-band frequency response. The measured S_{11} impedance bandwidth is 1.42 GHz (6.518-7.949 GHz, 20%) at 7.707 GHz, which is better than the simulated bandwidth (1.25 GHz) valid for C-band and X-band for 4G, and suitable for mid-band 5G wireless applications. The measured results and simulated S_{11} microstrip are comparisons presented in Table VIII.

The experimental results return loss (S_{11}) is in close with its simulation result. However, there, exist some slight discrepancies caused during the implementation of the microstrip patch antenna precision and interface deviation due to loss material, Rogers RT5880 plate for the manufacture of the antenna prototype which is not typical in each country. The fabricated method and measurement techniques generate the differences between simulated and measurement results. Also, S_{11} levels are accredited to tolerance during the fabricated and measurement steps.

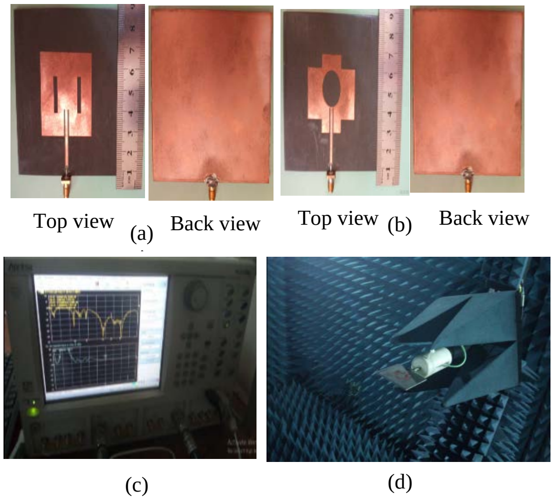


Fig. 16. Fabrication and Measurement for Proposed Microstrip Patch Antennas, (a) Fabricated Antenna_1 (b) Fabricated Antenna_2 (c) The Proposed Microstrip Patch Antennas Measurement on Network Analyzer (d) Experimental Test which was set up of Proposed Antennas in an Anechoic Chamber.

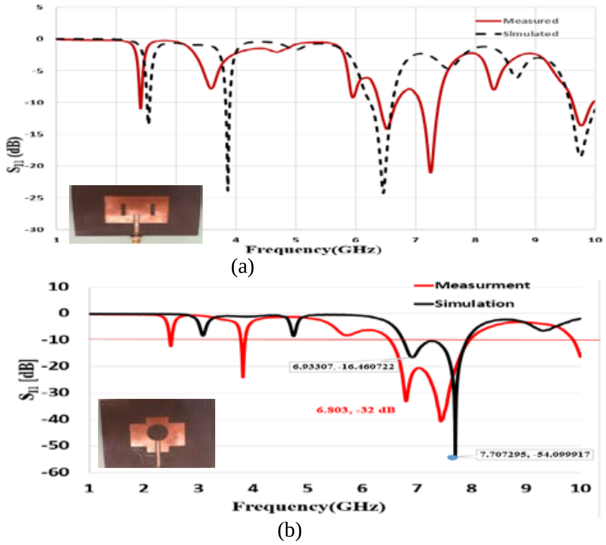


Fig. 17. Measured and Simulated of Reflection Coefficient S_{11} for (a) antenna_1, (b) antenna_2.

TABLE VIII. The Experimental Measurement and Simulated Comparison Results, TRI-Band of Antenna_1, and a Dual-Band of Antenna_2

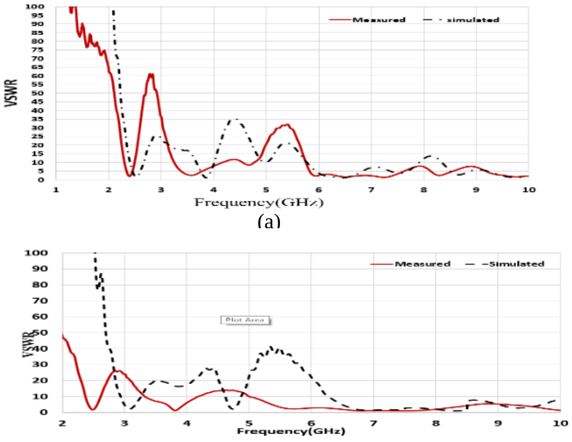
antenna_1	**Resonant freq. (GHz)**	**Directivity (dBi)**	**Realized Gain (dB)**	**Bandwidth (MHz)**	**antenna_2**	**Resonant freq. (GHz)**	**Directivity (dBi)**	**Realized Gain (dB)**	**Bandwidth (MHz)**
Measured	7.25 6.45 2.40	12 8.5 8.9	9 6.5 7.2	310 240 50	Measured	7.44 6.80	8.5 6.9	6.5 5	1420
Simulated	6.45 3.86 2.53	10.62 8.22 8.34	10.5 7.97 8.18	410 70 50	Simulated	7.707 6.93	7 6.9	6.22 5.56	1250

Illustrates the performance of measured VSWR from Fig. 18, for the proposed patch antennas, which verified value less than 1.5 at all resonant frequencies compared with simulated results. At all frequencies, minimum reflected power is inferior to -20 dB. The conclusion derived from the analysis of both Fig. 17 and 18, measured results of return loss and VSWR respectively, confirms that the designed multi-band microstrip patch antenna ensures obtaining good performance.

From Fig. 19, it can be observed realized gain plots vs frequency measured and simulated for the antenna_1 and antenna_2. The simulated gain is 10.6 dB at high frequency (6.45 GHz) and 8 dB at low frequency (2.53 GHz). The measured gain of 7.2 dB at 2.402 GHz frequency, while 6.5 dB at 6.55 GHz frequency is shown in Fig. 19(a).

Fig. 19(b) shows the experimental and simulated result realized gain of antenna_2. The measured gain is 5 dB at 6.8 GHz, 5.6 dB at 7.44 GHz, while the simulated realized gain is 5.56, and 6.22 dB at 6.93, 7.707 GHz, respectively.

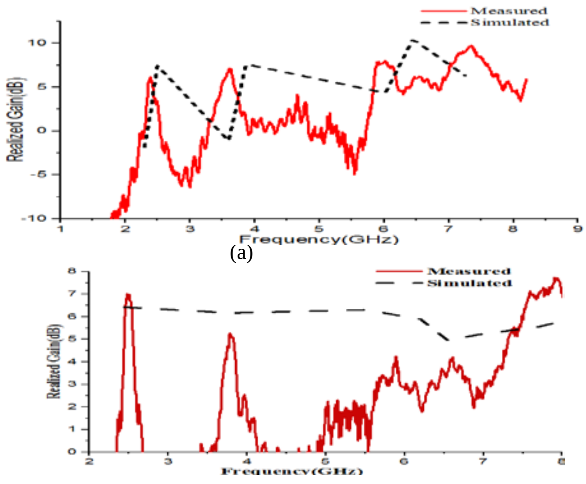
Fig. 20 shows the measured and simulated directivity against frequency for the antenna_1 and antenna_2.



(a)

(b)

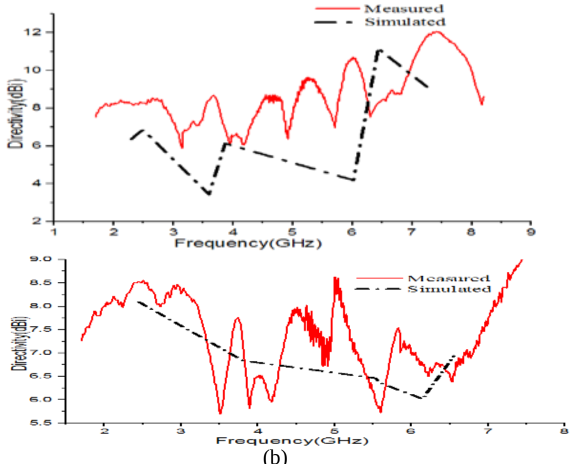
Fig. 18. VSWR Measured and Simulated for (a) Antenna_1 (b) Antenna_2.



(a)

(b)

Fig. 19. Measured and Simulated Realized Gain of (a) for Antenna_1, and (b) for antenna_2.



(b)

Fig. 20. Measured and Simulated Directivity for (a) Antenna_1, and (b) Antenna_2.

The realized gain, directivity, and impedance bandwidth of proposed antennas are tabulated in Table VIII.

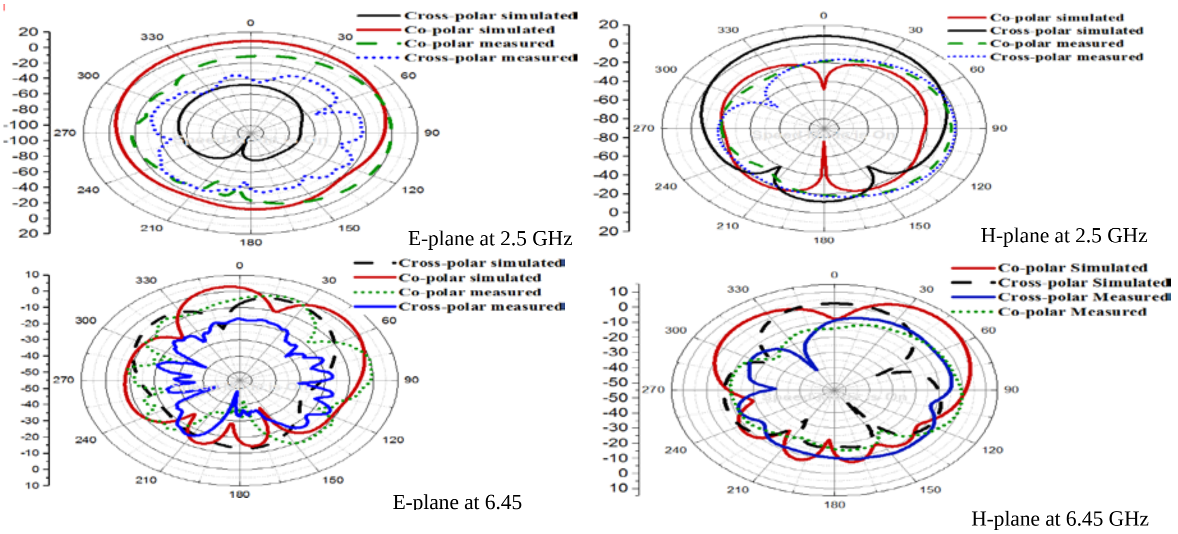
Fig. 21(a), (b) display the experimental measurement and simulation radiation patterns for the antenna_1 along with two elevation cuts (xz and yz planes) exhibit dual-polarization, E and H-plane co and cross-polarization at frequency bands 2.53 and 6.45 GHz.

Fig. 22 displays the experimental measurement and simulated radiation patterns for an antenna_2 along with two elevation cuts (xz and yz planes) exhibits dual-polarization, E and H-plane co, and cross-polarization at 7.707 GHz. The deviation of the measured results from the simulated results may be attributed to fabrication imprecision and measurement errors. Besides, the dielectric slab will be cause-effect, with little differences in the side lobes of the radiation patterns between measurements and simulation results.

From Fig. 21 and 22 discussions of co-polarization and cross-polarization between measurement and simulated for proposed antennas. Co-polarization is defined as the polarization the antenna was intended to radiate, while Cross its perpendicular pair [1].

The cross-polar and co-polar in E and H-plane for antenna_1 have been shown in Fig. 21(a) and Fig. 21(b) at 2.53, 6.45 GHz respectively, which is close to omnidirectional nature. Fig. 21(a) has shown radiation pattern in E-plane, simulated co-polar radiation angular width 27.3 deg. has been measured with an angular width (3) of 25.3 degrees at 6.5 GHz. But simulated cross-polar 39.7 deg. and measured 35.5 deg. at 6.5 GHz. While the measured co-polar in E-plane angular width of 68.8 degrees at 2.5 GHz and simulated angular width of 75.0 GHz at 2.53 GHz. But radiation pattern in E-plane the measured cross-polar angular width about 58.8 degrees at 2.5 GHz and simulated 76.2 degrees at 2.53 GHz.

Fig. 21(b) shows simulated co-polar in H-plane of 68.5 degrees and measured 75.2 degrees at 2.53 GHz whereas, the simulated cross-polar in H-plane angular width 68.8 degrees at 2.5 GHz, and measured angular width of 63.4 degrees at 2.5 GHz.



E-plane at 2.5 GHz

H-plane at 2.5 GHz

E-plane at 6.45

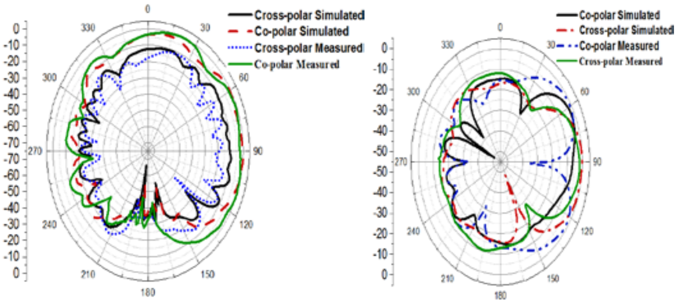
H-plane at 6.45 GHz

Fig. 21. Measured and Simulated Co-polar and Cross-polar of Antenna_1 at 2.5 and 6.45 GHz.

The simulated cross-polar in H-plane angular width of 58.0 degrees at 6.45 GHz and measured 59.8 degrees at 6.5 GHz. Whereas, simulated co-polar in H-plane of 41.8 degrees at 6.45 GHz and measured of 58.3 degrees at 6.5 GHz. It is possible to omission the far-field radiation in E-plane, compared to H-plane at frequency band 2.53 GHz, whereas, it is noted that the cross-polarization of the E-plane is lower than the cross-polar in the H-plane. Similarly, the cross-polar in far-field at 6.45 GHz band in the E-plane is lower than the cross-polar in H-plane**.** Therefore, the angular-width is reduced when the far-field radiation pattern is changed from Omni-to-bidirectional.

Fig. 22 has shown co-polar and cross-polar in the H-and E plane. Fig. 22(a) illustrated measured a co-polar 74.5 degree and simulated 70.1 degrees at 7.707 GHz. Also, a measured cross-polar angular width of 49.7 degrees and simulated angular width of 51.0 degrees at 7.707 GHz has been shown.

Fig. 22(b) shows measured co-polar in H-plane angular width of 54.8 degrees and simulated 58.3 degrees. Also, the cross-polar angular width of 35.2 and 38.3 degrees was measured and simulated respectively at 7.707 GHz.



(a)

(b)

Fig. 22. Measured and Simulated Co-polar and Cross-polar for Microstrip Antenna_2 at 7.707 GHz, (a) E-plane, (b) H-plane.

V. Conclusion

In this work, we design and fabricate two proposed microstrip antenna covers multi-band microstrip patch antennas. These proposed antennas cover the useful frequency band of modern wireless communication systems. Antenna_1 covers tri-band frequency, for WiMAX band 2.53 GHz (2.51 – 2.55 GHz), WLAN/C-band band 3.86 GHz (3.80 – 3.87 GHz), and C-band 6.45 GHz (6.19 – 6.60 GHz) which has potential for C- band in 5G services. Antenna_2 covers dual-band for C- band, and X-band 6.92/ 7.707 GHz (6.72 – 7.92 GHz, 1420 MHz) which is serving for C- band and suitable for mid-band 5G application. The proposed design of microstrip patch antennas is characterized as simple structures to be manufactured (94 × 76 × 3.18 mm^3). Besides, Experimental results verified good conformity with simulation results such as return loss, gain, bandwidth, and radiation pattern of these antennas.

VI. Future Work

Simulation the antenna on another simulator and compare the simulation results obtained with two simulators: Improvement the bandwidth and radiation pattern; Fabricated microstrip antennas with other material is low cost.

References

[1] Balanis, “Antenna Theory: Analysis and Design”, John Wiley & Sons, Inc., Hoboken, New Jersey, February 2016.

[2] D. Pardhan, “Circular patch with circular slit patch the antenna used for ultra-wideband application”, International Journal of Electrical, Electronics and Data Communication, Vol. 5, Issue2, Feb.2017.

[3] A. Gupta, R. Kumar, “A Survey of 5G network: Architecture and emerging technologies”,2015, IEEE.

[4] M. Gupta and V. Mathur, “Koch boundary on the square patch microstrip antenna for ultra-wideband applications”, Alexandria Eng. Elsevier, 2017.

[5] Roopa, Jayadeva, & Kumarswamy, “Enhancement of performance parameters of sierpeinsiki antenna using the computational technique”, [WiSPNET Conference, 2016].

[6] M. A. Amin, Shimanto and Md. R. Raihan, “Design and Performance Analysis of a multiband microstrip patch antenna for GSM, WiMAX, WLAN, Walkie-Talkie and ATC Application”, 5th International Conference on Informatics, Electronics and Vision (ICIEV), IEEE, 2016.

[7] L. Prasad, B. Ramesh, K.S.R. Kumar, and K.P. Vinay, “Design and implementation of multiband microstrip patch antenna for wireless applications” Advanced Electromagnetics, vol. 7, 2018.

[8] M. H. Reddy, R. M. Joany, G. Manikandan, and A. S. A. Nisha, “Design of microstrip patch antenna with multiple slots for satellite communication”, International Conference on Communication and Signal Processing, IEEE, 2017.

[9] T. Dabas, B. K. Kanaujia, D. Gangwar, A. K. Gautam, K. Rambabu, “Design of multiband multi polarized single feed patch antenna” IET Microw. Antennas Propag., Vol. 12 Iss. 15, pp. 2372-2378, 2018.

[10] A. Ghosal, S. Kumar, and A. Das, “Multi-frequency rectangular microstrip antenna with an array of L-slots”, International Journal of Electronics and Communications (AEÜ), Elsevier, 2019.

[11] U. Keskin, B. Döken, and M. Kartal, “Bandwidth improvement in microstrip patch antenna”, IEEE, 2017.

[12] R. Kushwaha, V. K. Singh, N. K. Singh, A. Saxena, and D. Sharma, “A Compact pentagonal textile microstrip antenna for wideband application”, Springer Nature Singapore, Pte Ltd., 2018.

[13] S. Srivastava, D. Somwanshi, “Design and analysis of rectangular microstrip patch antenna for ZigBee application”, IEEE International Symposium on Nanoelectronic and Information Systems, 978-1-4673-9692-9/15, IEEE, 2015.

[14] Km. Kamakshi, A. Singh, M. Aneesh, J. A. Ansari, “Novel design of microstrip antenna with improved bandwidth”, Hindawi, ID 659592, 7 Pages,2014.

[15] S. K. Hasan, A. C. Shagar, “Design and analysis of U-shaped microstrip patch antenna”, 3rdInternational Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB17), IEEE, 2017.

[16] Lv. Hong, H. Zhixiang, S. Dengzhi, “Design of a dual-band microstrip antenna”, IEEE, 2016.

[17] N. Agrawal, A. K. Gautam, K. R. Faisa, “Design and packaging of multi-polarized triple-band antenna for automotive applications”, International Journal of Electronics and Communications (AEÜ), Elsevier, 2019.

[18] S. Patiland, V. Rohokale, “Multiband smart fractal antenna design for converged 5G wireless networks” International Conference on Pervasive Computing (ICPC), IEEE, 2015.

[19] M. S. Rabbani and H. G. Shiraz, “Improvement of a microstrip patch Antenna gain and bandwidth at 60 GHz and X bands for wireless applications”, Microwaves, Antenna & Propagation, IET, vol. 10, 2016.

[20] M. M. Islam, M.T. Islam, and M. R. I. Faruque, “Dual band operation of a microstrip patch antenna on a duroid 5870 substrate for Ku- and K-bands”, Hindawi, ID378420,10 pages, 2013.

[21] N. A. Muhammad. and Y. Idris, “Design and analysis of microstrip patch antenna for global WLAN applications using Matlab and CST micro-studio software. Vol. 6, Issue no. 4 ISSN 2321 3361, IJESC, 2016.

[22] CST Microwave Studio, ver. 2015, Computer Simulation Technology, Framingham, MA, 2015.

[23] Shivnarayan, S. Sharma, B. R Vishvkarma, “Analysis of slot loaded microstrip patch antenna”, Indian Journal of Radio and Space Physics, vol. 34, pp. 424-430, December 2005.

[24] N. M. Tarpara, R. R. Rathwa, D. N. A. Kotak, “Design of slotted microstrip patch antenna for 5G Application”, International Research Journal of Engineering and Technology, vol. 05, 2018.

[25] N. Gupta, “Effects of slots on microstrip patch antenna”, International Research Journal of Engineering and Technology, vol. 04 Issue: 02, 2017.

[26] J. Stewart, C. Nickerson, T. Lewis, “5G Mid-Band spectrum global update”, REF:2020391-62, MARCH 2020.

[27] “Global update on spectrum for 4G & 5G”, December 2020.

Stacked Autoencoder based Feature Compression for Optimal Classification of Parkinson Disease from Vocal Feature Vectors using Immune Algorithms

K. Kamalakannan[1]
Ph.D., Research Scholar (Part Time)
PG and Research Department of Computer Science
Adhiparasakthi College of Arts and Science
Kalavai, TN, India

Dr.G.Anandharaj[2]
Assistant Professor & Head
PG and Research Department of Computer Science
Adhiparasakthi College of Arts and Science
Kalavai, TN, India

***Abstract*—Parkinson’s disease (PD) is a neurological progressive disorder and is most common among people who are above 60 years old. It affects the brain nerve cells due to the deficiency of dopamine secretion. Dopamine acts as a neurotransmitter and helps in the movement of the body parts. Once brain cells/neurons start dying due to aging, then it will lead to a decrease in dopamine levels. The symptoms of Parkinson’s are difficultly in doing regular/habitual movements, uncontrollable shaking of hands and limbs may encounter memory loss, stiff muscles, sudden temporary loss of control, etc. The severity of the disease will be worse if not diagnosed and treated at the early stages. This paper concentrates on developing Parkinson’s disease diagnosing system using machine learning techniques and algorithms. Machine Learning is an integral part of artificial intelligence it takes huge data as input and train by making use of existing algorithms to understand the pattern of the data. Based on the recognized pattern, the machine will act accordingly without any human intervention. In this work, two major approaches have been employed to diagnose PD. Initially, 26 vocal data of PD affected and healthy individual datasets are obtained from the UCI Machine Learning data repository, are taken as initial raw data/features. In pre-processing, the mRMR feature selection algorithm is employed to minimize the feature count and increase the accuracy rate. The selected features will further be extracted using the Stacked Autoencoder technique to improve and increase the accuracy rate and quality of classification with reduced run time. K-fold cross-validation is used to evaluate the predictive capability of the model and the effectiveness of the extracted features. Artificial Immune Recognition System – Parallel (AIRS-P), an immune inspired algorithm is employed to classify the data from the extracted features. The proposed system attained 97% accuracy, outperforms the benchmarked algorithms and proved its significance on PD classification.**

Keywords*—Immune algorithms; Parkinson’s disease; stacked autoencoder; airs-parallel; machine learning

I. Introduction

Parkinson's disease is a complex neuro-related disorder, having more prevalence among elderly people around the world. It is essential to diagnose it early to treat it accordingly. Although it has several treatments, medications, and surgery, it is always better to recognize the symptoms at the early stages. So, it helps in better recovery of the PD affected patients. Medication plays an important role in controlling the symptoms of PD. Medications include Dopamine Promoter, Anti-depressant; Anti-tremor can help in overcoming the effects of PD. The most prescribed medicine is L-dopa (Levodopa) combined with Carbidopa. The medicine will be converted into dopamine by the brain cells and thus it balances the level of dopamine needed for the motor actions of the nervous system. However, recognizing the symptoms of PD and early diagnosis helps to control the severity before it gets worse.

The genetic mutations and environmental factors may be the genesis for Parkinson’s disease. The usage of herbicides, fungicides, and pesticides is the emergence of acquiring Parkinson’s disease. The related studies unveiled the 70% of the people affected by the PD due to the excessive usage of the pesticides. There are several transformations obtain in the brain of Parkinson’s Disease afflicted people includes, including the clumps of some certain element in the interior of brain cells called Lewy bodies. Lewy bodies retain the key value to identify the genesis of the PD. Despite the possibility of many elements present in the Lewy bodies from the elements the vital protein termed asalpha-synuclein (a-synuclein). Alpha-synuclein is encountered in Lewy bodies in a clustered form, the cells in the clumped formcould not be decomposed.Various genes are decisively linked up with the Parkinson’s disease such as LRRK2, DJ-1, PRKN (Parkin), PINK1, GBA (Glucocerebrosidase – beta) and SNCA. Parkinson's disease is closely similar to other diseases such as Progressive supranuclear palsy (PSP), Corticobasal degeneration (CBD), and Multiple System Astrophys (MSA). The above three can be described as the name of Parkinson’s plus Disease.

Attributable to Parkinson’s plus disease, it is challenging to diagnose Parkinson’s disease from Parkinson’s plus diseases. PD diagnosis is elicited from either neurological examination, lab tests, or scans of the brain. As a consequence of not having the proper treatment to treat the PD, surgery, or medication is the possible one to improve the health of PD affected people. Several medications procedures are followed. Surgery will be prescribed when the medications are no longer good enough. Deep Brain Stimulation is the type of surgery used presently. In the future, some potential treatments will explore the areas in particularly neural (cell) transplantation, Gene therapy, and

Immunotherapy. In Neural Transplantation, displace the affected and dead brain cells with the new cells. The new cells can develop and increased. The outcome of the research holds the partial result, some peoples are getting improvement in the health and some of not. Gene therapy is another research area; this technique is also having some complications to implement effectively. Research is still on the horizon to cure Parkinson's disease entirely [1].

Computer-Aided Diagnosis (CAD) is rapidly emerging in these days to help people to check the early symptoms on their own with needed reports and data. This paper is one such diagnosis system developed using Artificial Intelligence, Machine Learning, and neural network schemes. For, we employ Stacked Autoencoder and AIRS parallel to extract the raw features and classify PD affected persons from healthy individuals by applying the feature vectors.

The upcoming part of the work is compiled in the following manner. Section 1 of the paper is Introduction has already been discussed. Section 2 discusses the existing works that inspired this paper. Materials and methods will be Section 3 as it mainly concentrates on the technical aspects of the proposed work and employed algorithms. Results, Simulation, and Comparison are done on Section 4 of the paper. Section 5 concludes with the summarization, significance, and importance of the work based on the results and existing works.

II. Literature Survey

In this part, the prevailing literature is conducting and preforming review and on it. The associated works mainly explore diverse feature extraction methods and classification algorithms on the healthy data.James Parkinson who wrote the initial medical depiction for Parkinson's disease in 1817. But it was further processed by Jean-Martin Charcot Parkinson's disease. Jean-Martin dissociated Parkinson's disease from other disorders and is characterized by tremors. It is a neurological progressive disorder. The person whose age is more than 60, Parkinson's disease is common to them. It mainly affects the brain nerve cells due to the deficiency of dopamine secretion. Dopamine acts as a neurotransmitter and helps in the movement of the body parts. Because of aging, the human brain cells start to perish, it will lead to a decrease of dopamine levels. The major symptoms of Parkinson's are difficultly in doing regular/habitual movements, uncontrollable shaking of hands and limbs may encounter memory loss, stiff muscles, sudden temporary loss of control and facial expression changes are recognized. The severity of the disease will be worse if not diagnosed and treated at the early stages.To diagnose the PD, there is a limited diagnostic test are an avail.

To diagnose the motor disorders of PDDaTscan is the only way out. To make the diagnosis ineffective way, machine learning provides an efficient way. In this paper, the voice dataset is used for diagnosing the PD by the use of supervised learning. The dataset found from the UCI machine learning repository. The overall dataset consists of 195 vowel voice records. Among the dataset, 48 voice records from healthy people and 147 from affected persons. 22 features are selected for preprocessing. 10 features are selected based on the Filter-Based Feature Selection algorithm from 22 features. The specific algorithm used was the Pearson Correlation scoring method is implemented to correlate the features with the label. K-fold cross-validation is used to perform training and testing on all data to increase the efficiency of the outcome.The dataset employs the following models Averaged Perceptron, Bayes Point Machine, Decision Forests, Locally-Deep SVM,Boosted Decision Tree Logistic Regression,Boosted Decision Tree, Neural Networks, and SVM. From these models Boosted Decision Tree provide the most accurate result when compared to other models. This paper concludes the voice recordings are feasible to diagnose Parkinson's disease.

Artificial Immune Recognition System is a modern supervised learning algorithm, inspired by the immune system. AIRS provides the best outcome for classification problems. AIRS is the fusion of artificial intelligence and biological inspired computation evolved from the metaphor and the heuristic knowledge of the biological immune system [2--6]. The AIRS is the first AIS procedure used to solve the classification problems. AIRS has somespecialized characteristics such as Self – Regulation, Generalization, Performance, and Parameter Stability. AIRS has many biological terms such as antigens, B-cells, T-cells, clonal selection, etc., the implementation level of AIRS is a very complicated one. The procedure of the AIRS algorithm is needed to prepare a collection of memory cells. Those memory cells are needed to train the data. The developmental process of the AIRS algorithm has the following steps: 1) Construct the data for the training process and the data should be normalized, use Euclidean distance measures for calculating the affinity measures, then select the antigens randomly for memory pool. 2) Training the memory cells by antigens. 3) From the selected memory cells are mutated clones, such clones are moved to ARB (Artificial Recognition Ball). 4) Competing for the limited resources 4. Selecting the memory cells. 5) The classification has to be done by implementing the k nearest neighbor method. The above life cycle of AIRS produces better accuracy in diagnosing the disease early. This paperconcludes, the AIRS has provideda good accurate result when compared to the rest of the classifiers [2].

This paper describes the Parallel AIRS. Parallel AIRS is one of the AIRS algorithms. AIRS 1 and AIRS 2 are the serial versions of the AIRS algorithm[7-9]. Both algorithms relying on a single processor to train the memory cells. But Parallel AIRS has multiple processors, so more than one processor can perform their task-parallel. AIRS 1 and AIRS 2 algorithm have the nine steps [10-13] worked in a single processor. The following steps are done by the parallel AIRS. Step1: From the root read the training data Step 2: Distribute the training the data to np (number of processes). Step 3: Each processor executes from step 1 to step 9 based on the serial processuntil the training data obtained. Now each processor holds the trained memory cells. Step 4: Collect the memory cells from each processor and the memory cells are merged and back to the root (initial stage). Speed up is achieved without any loss of accuracy in the classification. The efficiency of Parallel processor can be stated as E(P) = T(1) / P * P(T), where P is the total number of processors, T(1) is the time for AIRS 1 and AIRS 2, T(P) is the time for the algorithm of a parallel version. The AIRS algorithms (AIRS 1, AIRS 2, and Parallel AIRS) implemented on datasets in the WEKA platform [14]. The classification accuracy of Parallel AIRS shows the best when compared to the other AIRS algorithms.

Autoencoder is one of the unsupervised machine learning algorithmsin a deep neural network. The output values should be equal to input values. It is used to deplete the size of our inputs as a compressed form, by performing the reconstruction the original data is evolved. The architecture of the Autoencoder is of three parts, encoder, hidden layer, and decoder. The encoder compresses the input data into latent space representation [15-20]. It reduces the original dimension of the data. Hidden layers refer to code it holds the compressed input and the decoder, it reconstructs the code from latent space representation to produce the output. The autoencoder is used to extract some specific features from the data and produce the output. So, the autoencoder used as feature extraction. Stacked Autoencoder consists of various sparse encoder layers. Each is placed one after another like a hierarchical format. Each input of successive bottleneck (hidden or internal) layer connected to each bottleneck layer of output [21-22]. The algorithm of stacked autoencoder mainly follows three steps: First, obtained the trained data from the autoencoder. Second, trained data of the previous layer is used as an input to the successive layer, this process will continue until the training should be completed on all input data. Finally, after the completion of training in all internal layers, fine-tuning is attained. This paper employs Stacked Autoencoder to diagnose Alzheimer’s disease (AD), mild cognitive impairment (MCI). By training the data employing the Stacked Autoencoder improves the level of accuracy.

III. Materials and Methods

This segment discusses the process implemented in the paper to discern the best classifier for PD. It explains the dataset, feature extraction algorithms, k – fold cross-validation, and AIRS - Parallel classification algorithm. Fig. 1 represents the complete workflow of the proposed model.

A. Dataset Information

The examining work is started with acquiring the samples of voice recordings of PD affected peoples and healthy peoples from the UCI repository. The dataset is taken from the University of California and the Irvine Machine Learning repository. It contains 20 patient details with healthy people has 20 samples with 10 females and 10 males and the affected people 16 females and 14 males. The finalized version of the dataset holds 1040 instances and 26 attributes. The details about the dataset are given in Table I.

B. Feature Selection

In pre-processing, feature selection is the first step, where the raw data will be analyzed by a particular algorithm and the features will be further reduced based on the quality and clarity of the data. mRMR is the feature selection algorithm used in this work. For comparative analysis, two further feature selection algorithms called Correlation Feature Selection and Genetic Algorithm were employed. In general, the feature selection technique with the least number of features selected will mostly be considered as an optimized one [22-25]. Table III shows the numbers of features by this technique.

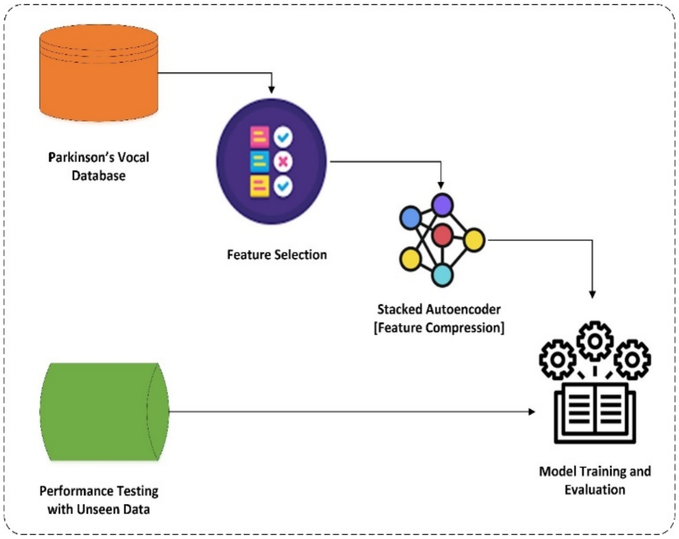

Parkinson’s Vocal Database
Feature Selection
Stacked Autoencoder [Feature Compression]
Model Training and Evaluation
Performance Testing with Unseen Data

Fig. 1. Workflow of the Proposed System.

TABLE I. DATASET DESCRIPTION

Dataset Source	***Department of Neurology, Istanbul University***
Disease Type	Parkinson's Disease
Total Number of Samples	1040
Total Number of Features	26
Classes	Binary Value 0 - Not PD, Value 1 – PD
Dataset Characteristics	Multivariate

Minimum Redundancy - Maximum Relevance (mRMR) is a technique, here used to select the optimum feature subsets. The core mechanism of the algorithms is; it selects the features highly relevant to the necessary classification yet features are mutually having less relevance it implies minimizing redundancy between the feature data. This technique fetches high accuracy with mutually unrelated features having more details about the problem. It helps in the precise classification of the data. The final subset S is identified based on the following equation.

$$mRMR = max_{F_i \notin S}[M(F_i;t) - \frac{1}{|S|}\sum_{F_j \in S} M\,(F_j;F_i)] \qquad (1)$$

C. Feature Extraction

When the input data is too large to process if it is repetitive it can be manipulated into a compressed set of features called feature extraction. Some of the feature extraction techniques are Latent semantic analysis, Partial least squares, principal component analysis, Multifactor dimensionality reduction. Autoencoder is one of feature extraction, it produces the output by eliminating unnecessary interruptions or noise. A stacked autoencoder is one of the methods of the autoencoder. In this paper, stack autoencoder is used as a feature extraction method. The stacked autoencoder receives the input as a voice signal from the big data source. The input is compressed by applying encoder layer this can be done by several layers and stored the values in the hidden layer, In the hidden layer, the data has to be trainedonce the training is done with it the output is then reconstructed from the hidden layer making use of decoder layer and it produces the output, the output should be equal to the input. Here, the number of feature 22 is given as an input and it performs the compression and it extracts the output with 8 features, here the remaining 14 features are considered as a noisy signal and eliminate those signals. This autoencoder model employs a cross-entropy loss function, and it suits well for this binary classification task. The parameters of the SAE are given in Table II. The general equation of the cross-entropy is represented below.

$$L\ (x, \bar{x}) = -\sum_{i=1}^{d}[x_i log\overline{x_\imath} + (1 - x_i)\log(1 - \bar{x}_i)] \qquad (2)$$

D. K- Fold Cross-Validation

K fold Cross Validation is also known as Rotation Estimation. It is one of the statistical methods used in machine learning to evaluate the skill of the particular model. This method holds a single variable "k". k refers to the total number of the groups the data has to be split for validation purposes. It has a simple procedure to work with k fold cross-validation. It randomly shuffles the input dataset, then it split into 10 groups (if k=10). Acquire one group for testing the data, the remaining group is undergone training the data by applying the model on it. Once it is trained then the group moves to test the data. After the completion of the test data, evaluate the score of the test data set. The evaluated score has been reserved and eliminate the models. Based on the value of the score, the model skills have to be analyzed.

TABLE II. PARAMETERS OF STACKED DENOISING AUTOENCODER

Parameters	***Values***
Epochs	100
Learning Rate	0.05
Momentum	0.40
Activation Function	Sigmoid
Reconstruction Error	Cross-Entropy Measure

E. AIRS-Parallel

In this system, AIRS Parallel is used as a classifier to diagnose the disease effectively. It is one of the AIRS methods. The training data set has to be given as an input to the AIRS Parallel after the completion of k fold cross-validation runs. It divides the dataset into many processors. Each processor holds some dataset randomly. Each processor performs the serial version of the AIRS process on the dataset. After the process completion on each processor, it gathers memory cells from each processor. Performing the merging operation on the memory cells into a single pool of memory cells. The memory cells in the pool are further divided into classes. In each class, an affinity pairwise calculation has been performed between the memory cells. If the affinity is less between the two memory cells then the affinity threshold scalar is a product by affinity threshold (aff(mc1, mc2) < afft * affts). As a result, only one of the memory cells is retained in the last pool. The outcome of this algorithm provides better accuracy to diagnose Parkinson's disease. The model skills have to be analyzed. In Table IV, the parameters of the artificial immune algorithm AIRS is given.

Algorithm: Stacked Autoencoder

Input: *S_i= Signal data, f_{sam} = Factor Sampling, n = Total number of Autoencoder, Target signal ($T_{s)}$ = f_{sam} sample signals, w= Weight, I_n = Internal node, H_e = High Epchos, SR = Sparsity Regularization*

```
$I_n$ =   Choose the number of SAE layers
$H_e$= Choose the number of SAE layers
w= Choose the number of SAE layers
SR= Choose the number of SAE layers
Train the Stack Autoencoder
Autoren = trainAutoencoder(2- n)   /* Train the Autoendcoder of n */
Test the input vs output
(a) Test the input signal (S) against Target Signal
(b) Calculate the error from the predicted output as a Target Signal
        if |err| ≥ level of tolerance
              repeat step 15
else
Finish training
```

Output:

Compressed Vectors

TABLE III. NUMBER OF FEATURES SELECTED BY EACH FEATURE SELECTION TECHNIQUE

Raw Feature Count	***mRMR***	***GA***	***CFS***
26	13	17	20

TABLE IV. PARAMETERS OF AIRS-PARALLEL ALGORITHM

Parameters	***Values***
Affinity Threshold Scalar Factor	0.2
Pool Size (Initial value)	1
Clonal Rate	10
Hyper Mutation Rate	2
Initial Pool Size	1
Stimulation Value	0.9
Iteration Number	1000

IV. EXPERIMENTAL RESULTS

In this section, the classification performance of the proposed combination will be evaluated and compared with the existing techniques. In summary, initially, the training dataset contains 26 featured vocal datasets obtained from 20 PD affected patients and 20 healthy individuals. It contains various kinds of 26 sound recordings of the voluntary individuals, in turn, forms 1040 overall voice recordings. The sound recording consists of sustained vowels, words, numbers, and small sentences. The test dataset consists of 6 voice samples that have been recorded from 28 PD affected patients. These 6 voice samples contain only sustained vowels 'a' and 'o' every three times and it has a total of 168 voice recordings. The dataset is obtained from the UCI Machine Learning repository. To narrow down the dataset for more accurate prediction and with comparatively reduced run time, the dataset will be pre-processed before the classification stage. In, pre-processing of the feature dataset, the 26 features have been reduced to 13 feature subsets. Furthermore, the selected 13 feature subset has been reduced to eight feature vectors deploying stacked autoencoder by performing compression and dimensionality reduction mechanisms. The extracted features have been estimated through the K-fold cross-validation technique to evaluate its predictive accuracy utilizing the existing dataset [26]. Here 5 folds were used to test and train the model to predict the accuracy. Fig. 2 represents the number of features selected by different feature selection methods.

For comparison of the proposed with the existing techniques, quality metrics need to be employed to determine the accurate performance analysis of the proposed work and its significance. For the reason that 4 major metrics were used to evaluate the proposed Stacked Auto encoder-AIRS Parallel technique. The main goal is to attain better disease classification accuracy to prove the importance of this work. The metrics are accuracy, specificity, sensitivity, and the confusion matrix plot. The parameters of the AIRS Parallel algorithm for Parkinson's disease need to be disclosed earlier. Table I represents the parameters used in the proposed algorithm with values

A. Performance Evaluation

Also referred to as an error matrix, it contains a table used to express the performance of the classifier on a test dataset for true known values. A confusion matrix has actual information and predicted information has been classified by applying the classification algorithm [27]. Based on the available data in the matrix, the performance of the model will be analyzed. The following table represents the confusion matrix for a binary classifier and the next table represents the outcome confusion matrix of the proposed work.

True Positive (TP): Detected as a patient diagnosed with PD by medical experts.

True Negative (TN): Detected as normal and categorized as healthy by medical experts.

False Positive (FP): Detected as patient and categorized as healthy by the medical experts.

False Negative (FN): Detected as normal who diagnosed with PD by medical experts.

Accuracy decides the overall performance of the system by classifying the PD affected individuals from the healthy ones and the accuracy was determined in percentage, higher the percentage, higher the accuracy [28-30]. The classification accuracy for the datasets of this study was calculated using the below equation.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

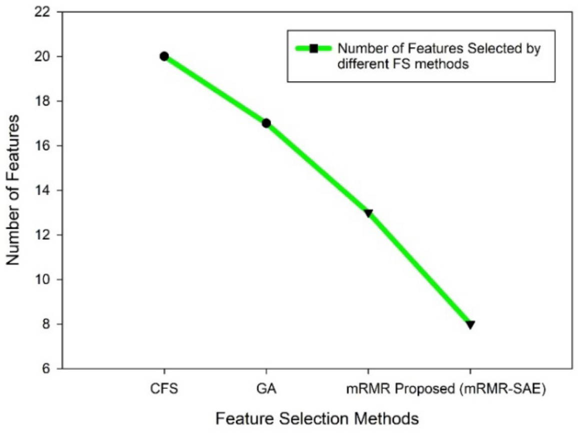


Fig. 2. Number of Features Selected by different Methods.

The sensitivity (sen) and specificity (spec) are calculated from the following equations.

$$Sen = \frac{TP}{TP + FN} \tag{4}$$

$$Spec = \frac{TN}{TN + FP} \tag{5}$$

The overall significance and importance of the model will be exposed only by comparing the results with the existing model's performances. Seeing, AIRS, and AIRS 2 algorithm with stacked auto encoder's performances will be taken to compare the results with the proposed combination. The results of AIRS – Stacked Autoencoder, AIRS 2 – Stacked Autoencoder, and the proposed AIRS Parallel –Stacked Autoencoder were presented as the Table V for comparison.

From the results given in Table V, it is evident thatthe proposed model outperforms the compared combinations in terms of Classification Accuracy, Specificity, and Sensitivity. Also, the proposed model can be further compared with the previous work, CFS-ACO with SVM classifier. Table VI shows the comparison values.

In Fig. 3, the scores of different validation metrics attained by different immune algorithms on the selected features is plotted. It is visible the proposed work comparatively classifies the disease better than CFS-ACO- SVM combination [31]. Insensitivity, the previous work seems to perform better than the proposed work but it gives a better percentage of results in terms of accuracy and specificity. This section highlighted the peculiarity of the work and from the experimentation results; the numbers in percentage clearly show the need for the work for better progress in the future.

TABLE V. COMPARISON OF RESULTS ATTAINED UNDER DIFFERENT AIRS ALGORITHMS (%)

Classifiers	***Acc***	***Sen***	***Spec***
AIRS	0.85	0.82	0.87
AIRS 2	0.9	0.86	0.92
AIRS Parallel	0.97	0.94	0.99

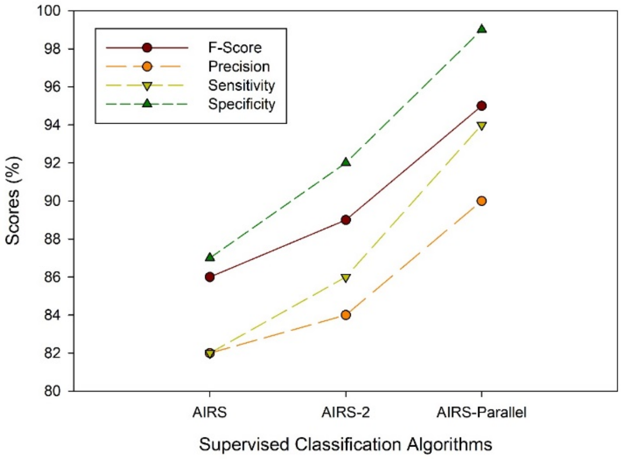


Fig. 3. Performance of Classifiers under different Evaluation Metrics.

TABLE VI. COMPARISON OF RESULTS FROM EXISTING STUDIES AND PROPOSED WORK (%)

Classifiers	**Acc**	**Sen**	**Spec**
CFS - ACO – SVM	0.95	0.96	0.98
Proposed	0.97	0.94	0.99

V. COMPARATIVE ANALYSIS OF FEATURE SELECTION TECHNIQUES BASED ON QUALITY METRICS

To benchmark, the performance of the proposed mRMR-SAE model, two other feature selection methods have been employed over mRMR. The performance is tested with SAE having the same configuration of the proposed algorithm. The techniques are shortly briefed below.

A. Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) evaluates and selects the feature subsets from the given data using a unique selection process. The feature selection was done based on acquiring an effective feature subset, it having more correlation with the classification and less or uncorrelated to the existing features [32].

B. Genetic Algorithm

Genetic Algorithm (GA) is a nature-inspired, search based selection technique derived from Charles Darwin's evolution theory. GA resembles the ideology of nature by selecting the fittest individuals for procreation of the forthcoming generation. GA has five main phases for a successful selection process. They are Initial Population, Fitness Function, Selection, Crossover, and Mutation. Each phase plays a significant role in GA for an optimal selection, thus resulting in healthy offspring reproduction. The performance of CFS and GA with stacked autoencoder is represented in Table VII and the accuracy of feature selection methods on AIRS-P is given in Fig. 4.

TABLE VII. PERFORMANCE OF AIRS-PARALLEL UNDER DIFFERENT FEATURE SELECTION TECHNIQUES (%)

FS Techniques	**Acc**	**Sen**	**Spec**
CFS-SAE	0.88	0.87	0.89
GA-SAE	0.92	0.93	0.92
Proposed	0.97	0.94	0.99

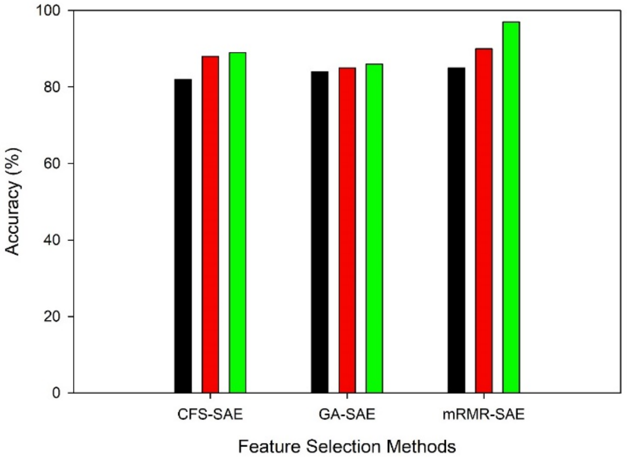


Fig. 4. Accuracy of Feature Selection Methods under different Classifiers.

VI. CONCLUSION

In this paper, the voice and speech recordings of PD affected and healthy individuals are analyzed with different statistical feature selection methods and neural network models. The 26 feature instances are pre-processed by deployingmRMR and Stacked Autoencoder - a neural network-based auto encoder technique used to reduce the noise in the data and compress the information of the data to reduce the number of attributes present in the original dataset. After dimensionality reduction of the dataset, the classification ability of the compressed vectors was evaluated with the K-fold cross-validation technique. Finally, the 8 feature vectors will be classified by the AIRS Parallel algorithm. The result of the proposed work was compared with AIRS, AIRS 2, and CFS-ACO-SVM combination. From the comparison, we can visibly conclude, the proposed AIRS Parallel with Stacked Autoencoder technique comparatively outperforms the employed techniques in all given quality metrics with 97% accuracy. It denotes the importance of this classification system for PD. Any Artificially Intelligence machine learning system will not be able to attain a 100% classification accuracy rate. But, the run time and other aspects of the system can be improvised in the future works. As the next step to this diagnosis/classification model, a Computer-Aided Diagnosis system can be developed, inspired by this proposed model to get a better classification accuracy rate with less run time and memory space usage.

VII. CONFLICT OF INTEREST

Authors declare no conflict of interest.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Christopher G. Goetz, “The History of Parkinson's disease: Early Clinical Descriptions and Neurological Therapies,” September 2011.

[2] Jason Brownlee, “Artificial Immune Recognition System (AIRS) A review and Analysis”, 2005.

[3] Andrew Watkins, Jon Timmis, Lois Boggess, “Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm,” December 1, 2002.

[4] C. M. Bishop, “Neural Networks for Pattern Recognition. Oxford University,” 1995.

[5] L.N.de Castro and J. Timmis, “Artificial Immune Systems: A New Computational Intelligence Approach. Springer,” 2002.

[6] S. Hofmeyr and S. Forrest, “Arichitecture for an aritifcial immune system,” Evolutionary Computation,” 2000.

[7] Watkins, A. and Timmis, Jon, “Exploiting Parallelism Inherent in AIRS, an Artificial ImmuneClassifier,” 2004.

[8] Basheer, Shakila; Bivi, S Mariyam Aysha; Jayakumar, S; Rathore, Arpit; Jeyakumar, Balajee. Machine Learning Based Classification of Cervical Cancer Using K-Nearest Neighbour, Random Forest and Multilayer Perceptron Algorithms,Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, May 2019, pp. 2523-2527(5).

[9] Watkins A., Bi X., Phadke, “AParallelizing an immune-inspired algorithm for efficient pattern recognition,” 2003.

[10] Goodman D., Boggess L., Watkins A, “Artificial immune system classification of multiple-class problem,”2002.

[11] Heung-Il Suk,Seong-Whan Lee, Dinggang Shen, “Latent feature representation with stacked auto-encoder for AD/MCI diagnosis”, November 14, 2013.

[12] Sultana, H Parveen; Shrivastava, Nirvishi; Dominic, Dhanapal Durai; Nalini, N; Balajee, J.M. Comparison of Machine Learning Algorithms to Build Optimized Network Intrusion Detection System, Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, May 2019, pp. 2541-2549(9).

[13] Jae-Neung Lee, Yeong-Hyeon Byeon and Keun-Chang Kwak, “Design of Ensemble Stacked Auto-Encoder for Classification of Horse Gaits with MEMS Inertial Sensor Technology,” August 2018.

[14] Pierre Baldi, “Autoencoders, Unsupervised Learning, and Deep Architectures,”2012.

[15] G.E. Hinton and R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks. Science,” 2006.

[16] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” 2010.

[17] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research,” February 2010.

[18] Karthikeyan T., Sekaran, K., Ranjith D., Vinoth Kumar V., & Balajee J M. (2019). Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques. International Journal of Web Portals, 11(2), 41–52.

[19] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, “Exploring strategies for training deep neural networks. Journal of Machine Learning Research,” January 2009.

[20] G. An, “The effects of adding noise during backpropagation training on a generalization performance. Neural Computation,” 1996.

[21] Ian H. Witten and Eibe Frank, “Data Mining: Practical machine learning tools with Java implementations, San Francisco: Morgan Kaufmann,” 2000.

[22] Donald Goodman, Lois Boggess, and Andrew Watkins, "An Investigation into the Source of Power for AIRS, an Artificial Immune Classification System," 2003.

[23] Jon Timmis and Mark J. Neal,” A Resource Limited Artificial Immune System for Data Analysis Research and Development in Intelligent Systems XVII,” December, 2000.

[24] Hossein, R.; Ajmal, M.; Mubarak, S., “Learning a deep model for human action recognition from novel viewpoints,” 2018.

[25] Potapov, A.; Potapova, V. “Peterson, M. A feasibility study of an autoencoder meta-model for improving generalization capabilities on training sets of small sizes,” 2016.

[26] G. E. Hinton, A. Krizhevsky & S. D. Wang, “Transforming Auto-encoders.

[27] Tingxi Wen, Zhongnan Zhang, “Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals,” May 2018.

[28] L. Zhang, F.-N. Yuan, J.-T. Shi, X. Xia, G. Li, “Theories and Applications of Auto-Encoder Neural Networks: A Literature Survey,” January 2019.

[29] Weixing Chen, Chaochen Cui, Xiaojing Li. “Bearing Fault Diagnosis Based on Improved Denoising Auto-encoders,” January 2020.

[30] Abass Olaode, “Local Image Feature Extraction using Stacked-Autoencoder in the Bag-of-Visual Word modelling of Images”.

[31] S. Wang, Z. Ding and Y. Fu, "Feature Selection Guided Auto-Encoder," 2017.

[32] K. Kamalakannan, Dr. G. Anandharaj. (2020). Deep Feature Selection from the Vocal Features for Effective Classification of Parkinson ’s disease. International Journal of Advanced Science and Technology, 29(08), 1661-1672.

Gender Diversity in Computing and Immersive Games for Computer Programming Education: A Review

Chyanna Wee[1], Kian Meng Yap[2]
Department of Computing and Information Systems
Sunway University, Bandar Sunway, Malaysia

***Abstract*—This paper provides a review of the current state of the gender gap in computer science and highlights how immersive games can mitigate this issue. Game-based learning (GBL) applications have been shown to successfully incite motivation in students and increase learning efficiency in both formal and non-formal educational settings. With the rise of GBL, researchers have also used virtual reality to provide pupils with a more immersive learning experience. Both GBL and virtual reality techniques are also used for computer programming education. However, there is a paucity of applications that utilize these techniques to incite interest in computer science from a female perspective. This is a cause for concern as immersive games have been proven to be capable of inciting affective motivation and fostering positive attitudes towards specific subjects. Hence, this review summarises the benefits and limitations of GBL and virtual reality; how males and females respond to certain game elements; and suggestions to aid in the development of immersive games to increase female participation in the field of computer science.**

Keywords—Computer science education; game-based learning; gender; virtual reality

I. Introduction

The projected number of degrees awarded to females was estimated to increase from 39 percent in 1966 to 61 percent in 2019 [1]. Despite the increasing number of degrees awarded to females, only about 35 percent of them are enrolled in Science, Technology, Engineering, and Mathematics (STEM) courses [2]. To make matters worse, the dropout rate for female students is 23 percent higher than for males in STEM courses [3]. The gender gap is most prominent in the field of computer science, where estimates of the ratio of male to female enrolment at tertiary level are approximately 70:30 as of 2019 [4]. This may be attributed to the gender gap in computer science advance placements (AP) test-takers where there were 52,574 males to just 17,111 females in 2019 [5]. Collectively, all this data means that females make deliberate decisions when it comes to not pursuing a career in the computer science field. Although there are several contributing reasons for this, the main factor stems from stereotypical representations of STEM-related fields that implicitly deter females and give them an incorrect perception of what the field is made of [6].

Due to low female enrolment rates in computer science, there is a lack of diversity in the computing workforce. According to data provided by the United States Department of Labour's Bureau of Labour Statistics (BLS), computer-related jobs are projected to rise by 12.5 per cent from 2014 to 2024 [7]. However, there is still a substantial difference between the number of female employees and the number of male employees in computing professions, with approximately 1,226,000 females compared to 3,578,000 males [8]. This is also prevalent in the field of academia, where only 15 per cent of professorial computing positions are held by females [8]. Ultimately, this implies that the consequences of low female enrolment rates in computing courses reflect upon the workforce. This is a cause for concern as diversity is crucial to stimulate and trigger innovative solutions to problems in the field.

II. Related Works

In recent years, the game-based learning (GBL) technique is widely used in educational contexts to foster knowledge acquisition, skill acquisition, and engagement in students [9]. To further enhance student participation in the learning process, GBL is also frequently combined with virtual reality and augmented reality. Previous reviews ([10], [11], [12]) have covered the benefits of the use of games for programming education. There have also been reviews ([13], [14]) that highlighted the use of virtual reality in education. More recently, there are also reviews ([15], [16]) that discussed the prevalent gender gap in computer science and virtual reality-based research. However, there is a relative paucity of reviews that explored the possibility of utilizing immersive games to increase female participation in the field of computer science.

Therefore, this review aims to foster more research in the development of immersive learning applications to reduce gender gaps in computing. Fig. 1 provides an overview of the topics discussed in this review, starting with why diversity is important in computing and factors that influence female participation in computing. Other major topics can be found in the "Results" section in this review where GBL and virtual reality are explored concerning their application in education and computer programming education to determine if the utilization of these techniques may aid in increasing female participation in the field of computer science. Furthermore, an overview of how males and females perceive game elements are also discussed to further aid in promoting the development of more female-centric immersive games on computing education.

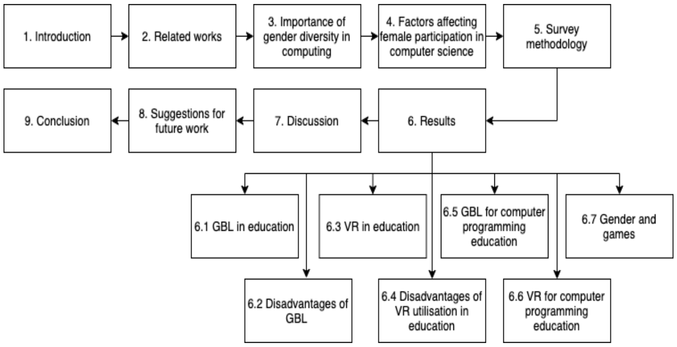

1. Introduction
2. Related works
3. Importance of gender diversity in computing
4. Factors affecting female participation in computer science
5. Survey methodology
9. Conclusion
8. Suggestions for future work
7. Discussion
6. Results
6.1 GBL in education
6.3 VR in education
6.5 GBL for computer programming education
6.7 Gender and games
6.2 Disadvantages of GBL
6.4 Disadvantages of VR utilisation in education
6.6 VR for computer programming education

Fig. 1. Overview of the Review Paper.

III. Importance of Gender Diversity in Computing

With technological advancements in recent years, more jobs in the field of computing are expected. With 598,700 job openings annually from 2018 to 2028, compared to 290,200 jobs for non-computing STEM subjects, the demand for computing graduates has increased and will continue to grow for several years to come [17]. Since this is the case, the enrolment rates for computing courses have also increased by 5.4 percent in 2019 since the previous year [18]. Even so, just 27 percent of females are working in computer-related careers [19]. The gap is more significant in academia, where it is estimated that it would take about 118 years to close the gender gap between authorships and publications related to computer science [20]. This is a cause for concern because diversity will bring different perspectives to the table in terms of understanding and designing solutions for a wide variety of individuals. Take car crash dummies, for example, that are modelled based on a man's body. According to a study by the University of Virginia [21], females have a 47 percent greater chance of facing serious injuries compared to males due to the inaccurate representation of the female body through the dummies. This issue persists in technology where females are more likely to suffer from virtual reality headset-induced motion sickness at 77 percent compared to 33 percent for males [22]. The authors conclude that this is due to the difference in physiological structure of both genders and suggest that manufacturers address aspects of design that are related to one's ability to stabilize their bodies. Ultimately, if these issues continue to persist, the accessibility of new and upcoming technology will only be limited to only part of the population, further perpetuating the already wide gender gap in the field of computer science.

IV. Factors Affecting Female Participation in Computer Science

Throughout history, the field of computer science has been inherently dominated by males. This is largely due to traditional perceptions of how certain genders are deemed more "suitable" for the pursuit of certain fields. For example, the fields designed to be better suited for females include nursing, teaching, and homemaking. Undoubtedly, this propels them to pursue the previously mentioned careers as parental approval directly influences one's career choices [23]. This, in turn, increases the gender gap and further decreases the sense of belonging of female students in the field of computer science. Because the sense of belonging is a strong motivator for the interest of female students in the field of computer science, this creates a never-ending cycle [24]. The lack of female representation also accelerates the lack of interest in females to pursue computer science. In a study conducted by Hussain et al, female role models can help remove the negative consensus related to computer science [25].

Furthermore, how computer science classes are conducted can also be unappealing to females. In a study conducted by Giannakos et al, poor teaching, excessive workload, and boring classes also contributed to high dropout rates [26]. Syllabi that include real-world problems instead of abstract ones are crucial in making the class seem more interesting [27]. This can be achieved by addressing relevant topics of interest to females, such as developing applications, that may help people or animals. Other than that, grouping strategies that foster cooperation and discussion can also be effective [28]. This is because females prefer to work in groups and have a lower level of confidence in themselves and their abilities [29]. As for their male counterparts who have a higher level of confidence in their skills, they often prefer to work on their own [29]. However, because there are often more males than females in computer classes, collaboration can also be an issue. In combination with stereotypical perceptions, females may feel shunned and are therefore unable to participate in class [30]. According to Hoegh and Moskal, the most likely step in eliminating the declining numbers in computer science enrolment is to understand the students' attitude to the field [31]. Despite having equal access to computers as their counterparts, females generally have a negative attitude towards computing and are found to be less concerned about anything that involves computing [32]. Consequently, Jo Tondeur et al found that females also have a less positive attitude compared to males towards computers in general [33]. The societal perceptions of computing-related careers can also increase cynicism in females. For example, students are made to believe that those with computing careers have to face computers all day long and that it will make them seem unsocial [34]. This problem stems mainly from a lack of understanding on what professionals do in different fields of computing. Researchers at Carnegie Mellon University found that many students picture computer science students to be intense hackers, which can also be unappealing to those that have aspirations to do more than technical work [35]. In response to this, changes were introduced in the curriculum to emphasize that there is more to computer science than just being a "hacker". Since then, from just seven female students from a total of 95 students in 1995, it was reported that 105 female students were enrolled into the computer science course from a total of 211 students in 2018 [36].

Lastly, the lack of exposure at an early age also contributes to the gender gap issue. Most computer games and early educational software have been developed and designed predominantly with boys in mind [37]. This, in turn, could explain why males today are more familiar with computers and seem to have a head start in classes. This, coupled with the stereotyping that has been mentioned, may intimidate females and lead to early termination of the computer science class or course in higher levels of education. This can also cause a decrease in the confidence of one's technical skills and abilities, which in turn, further inhibits female participation in the field of computer science. However, in a study conducted by Chen, although males had more motivation than females due to prior experience, males were outperformed during the performance evaluation [38]. This means that given the right instructional simulations, even learners without computing experience can outperform those with computing experience effectively. Consequently, according to LaBouliere et al, females may also face a dip in confidence if they are unable to observe their finished product [39]. Hence, to address these concerns, immersive games can be developed to foster interest as well as motivation in computing.

V. Survey Methodology

Now as we understand the importance of diversity in computing and the factors that affect female participation in computing, the aim of this review which is “to foster more research in the development of immersive learning applications to reduce gender gaps in computing” becomes more important. To support the aim of this review, the following research questions were constructed.

RQ1: What are the benefits of GBL and VR on computing education and how can this reduce the gender gap in computing?

RQ2: What are proposed game elements that can be employed in immersive learning applications to reduce the gender gap in computing?

To answer the research questions posed above, this review will be conducted systematically according to the steps as detailed in sections IV A to C.

A. *Definition of Keywords*

To determine how GBL and VR can promote female participation (RQ1), keywords and search terms such as “Game-based learning AND Education” and “Virtual reality AND Education” were first used to first identify how immersive techniques affect learning in general. Keywords such as “Game-based learning AND Computer education” and “Virtual reality AND Computer education” are then used to identify how both GBL and VR affect computing education. Furthermore, to determine elements that should be employed in immersive learning applications in promoting female participation (RQ2), the keyword “Gender AND Games” were used.

B. *Literature Sources*

The search for literature is done through Google Scholar to ensure that there are no biases towards certain databases. It is estimated that almost sixty percent of systematic reviews published fail to retrieve ninety-five percent of available and relevant literature [40]. Hence, the use of a search engine such as Google Scholar would ensure the review includes a more comprehensive sample of literature. To ensure that only the most relevant literature is collected, searches were limited to twenty pages deep and were sorted by relevancy.

C. *Inclusion and Exclusion Criteria*

In this review, literature was also analysed through specific inclusion and exclusion criteria as listed below.

Inclusion criteria

- Literature relevant to keywords specified.
- Literature that is written in the English language only.

Exclusion criteria

- Literature that is duplicated.
- Literature with an abstract only.

Moreover, to ensure that the sources were relevant and suitable for this review, literature was also further analysed through the steps detailed below:

- Analysis of title and abstract.
- Analysis of introduction and conclusion.
- Analysis of entire article.

VI. Results

Fig. 2 below shows the breakdown of the number of studies selected for each research question during each step of the screening process. Articles that were found to have satisfied all the criteria are discussed in the sections below.

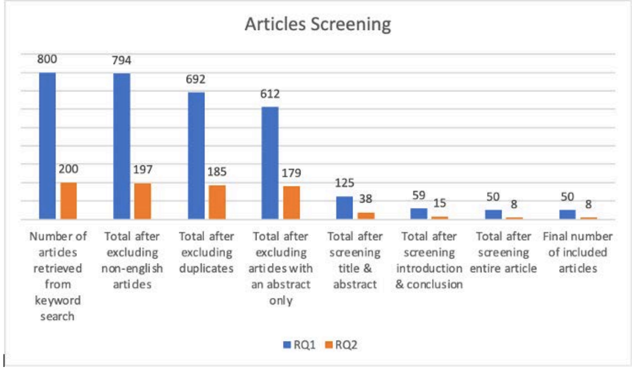


Fig. 2. Article Screening Overview.

A. *Game-based Learning in Education*

Contrary to more conventional forms of teaching, those who use educational games or simulations for learning have exhibited positive feedback on knowledge acquisition [41]. Educational games are considered a more organic way of learning because those engaged in gaming from an early age acquire cooperation and thinking skills [42], [43]. For instance, Li et al [44] developed a quiz game resembling “Who wants to be a millionaire” to engage students to learn the Chinese language. The feedback from elementary school students who played the game is mostly positive since students have enjoyed learning through games. Consequently, Yan and Tam [45] also observed the same results after employing a game based on Chinese history. About 81 percent of the participants liked to learn through games and 70 percent conveyed that they gained knowledge from the game. Hence, by using educational games in a classroom environment, students can be more engaged and encouraged to actively participate in lessons because GBL focuses on students instead of teachers [46]. This, in turn, will improve students' success from a learning perspective [47], [48], [49], [50].

Apart from engaging students, inciting motivation is also important to ensure that students continue to enquire and participate in the learning process. Cezar et al [51] proposed a Role-Playing Game (RPG) in which participants are required to solve calculus problems to advance through the game. About 83.3 percent of the participants felt motivated to finish the game. Similarly, according to a survey conducted by Yong et al [52] on the feasibility of digital games in mathematics education, 51 percent agreed that mathematics could be more interesting with the use of games. Furthermore, 48 percent agreed that digital games could help with the learning process. Consequently, Lo and Lin [53] proposed a digital board game where the player is allowed to move forward if the mathematical question raised to them is answered correctly. The findings from this study showed a significant difference in attitude before and after the games were played. By incorporating games into the learning process, students will be exposed to interactive learning experiences, increasing interests towards a particular subject matter.

The use of games in the learning process can also be used for the achievement of motor, spatial and cognitive skills. Hogle et al [54] found that those who were trained using a laparoscopic simulator showed an improvement in operative performance and depth perception compared to those who did not use the simulator. Similarly, Stefanidis et al [55] also showed positive results in which participants achieved suture proficiency by using a simulator. In the case of secondary tasks, longer training times are needed to achieve higher proficiency. Furthermore, research done by Green and Bavelier [56] has shown that frequent gamers are more proficient when it comes to tracking simultaneously two more items than those who often don't play. This is also seen in another experiment conducted by Feng et al [57], where the results have shown that when female players were exposed to an action game, both spatial and attentional skills were increased. Barlett et al [58] found that cognitive skills such as memory and auditory perception can also be improved despite playing for a short time only. This fact is supported by Piaget [59] who states that play is integral and evolves with the different stages of cognitive development.

B. *Disadvantages of Game-based Learning*

The advantages of GBL, however, depend very much on the design of the game itself [60]. The game must be designed in such a way as to force the player to achieve learning objectives while not being bogged down by the objectives of the game itself. Simply put, this is a contradiction that is difficult to overcome, considering that games are played only for fun, while GBL requires the players to accomplish external goals that are not "fun" [61]. Another limitation to GBL is that players can get bogged down with the idea of the game itself that they cannot think about what they've learnt [62]. For example, students may tend to focus on how their scores compare to that of their classmates rather than the learning objectives that are set for them to achieve. Other than that, GBL applications are usually infused with narratives to engage users. However, this can lead to confusion on the part of the user, as it is often difficult to distinguish between fiction and non-fiction. This phenomenon is prevalent in the results of the research conducted by Huizenga et al [63], in which pupils participate in a game to learn the history of medieval Amsterdam. The team incorporates fictitious narratives to further engage the pupils. This resulted in some confusion as to the actual historical facts and fictional elements of the game.

C. *Virtual Reality in Education*

Virtual reality is also prominent in tertiary education, where it is used in a variety of fields. Dinis et al [64] proposed a software program that allows students to design their three-dimensional virtual environment. By doing so students would be able to explore and apply what they have built up in civil Engineering situations. Likewise, Kharvari and Hohl [65] used virtual reality methods for reproducing the virtual form of a building for architecture students to study the precedents of a particular structure without the need for a field visit. The fact that virtual reality is capable of replicating how lightning behaves in a room also favours lightning education as proposed by Boyles et al [66]. Students would be able to interact with different lightning scenarios that are unlikely to exist in the real world.

Furthermore, intending to increase the engagement of students in the field of history, Zhang et al [67] suggested the use of the popular building game, "Minecraft" to build exact replicas of historical monuments. With this, students would be able to map the virtual experience to the knowledge they have gained in a traditional classroom setting. Hence, with the ability to directly interact with the virtual environment, students would be more willing to participate in lessons. By the same token, Lugrin et al [68] developed a virtual reality-based museum where visitors can experience interactive exhibitions without the need to travel to the actual museum. Since virtual reality is capable of effectively reproducing real-life scenarios, Caluya and Santos [69] proposed an authoring tool to study clouds for weather reporting purposes. Depending on the behaviour and appearance of the clouds, the participants in this study are exposed to different sky conditions and how these two factors relate to each other.

Moreover, virtual reality technology is widely used in medical training where live biomedical samples are not always available for training purposes. For instance, Seo et al [70] developed a virtual experience focusing on canine skeletal systems in which participants were able to freely observe and assemble the corresponding limbs. Similarly, de Mauro et al [71] proposed a simulated neurosurgical microscope to train and educate brain surgeons. The simulation focuses mainly on surgical procedures, as well as brain tissue-mimicking to help surgeons determine normal and low-grade glioma tissues. Comparably, Rajeswaran et al [72] developed a virtual experience to instruct medical professionals on endotracheal intubation surgery. Dong et al [73] proposed the integration of virtual reality and haptic feedback to simulate the preparation of cryogenic samples. With haptic feedback, biomedical professionals would be able to learn the best way to handle samples, as the equipment involved in Cryo-Electron Microscopy is expensive and delicate.

In summary, virtual environments can be developed to provide access to inaccessible places and tools for educational purposes. The utilisation of virtual reality in education also makes the learning process more interactive, which can increase student participation.

D. Disadvantages of Virtual Reality Utilisation in Education

Despite its advantages, the use of virtual reality in the context of education also has limitations. According to Mathur [74], despite being able to provide access to inaccessible medical resources, specialised virtual reality systems in the medical field are still prohibitively expensive for widespread adoption. To counteract this, low-poly graphics could be used to build an application because they require less processing power and can make the game run smoothly. This is especially essential for virtual reality games developed for education, so as not to deteriorate the learning experience. Furthermore, because virtual reality applications use Head Mounted Device (HMD) conventional input devices, such as mouse and keyboard, are no longer usable and are instead replaced by controllers. This can be seen in a virtual experience designed by Lei Wei et al [75] that uses Microsoft Kinect controllers. The controllers have been reported to be inaccurate and counter-intuitive, which has undermined the experience of the participants. Virtual reality education systems may also not be sufficiently engaging for users to continue to use them. For example, applications developed by Hsiao et al [76] and Hsieh et al [77], who developed a virtual campus for educators to teach online and a virtual museum respectively, reported a lack of engagement from participants.

E. Game-based Learning for Computer Programming Education

To make programming more engaging and improve learning, GBL is also often integrated into programming lessons. For instance, Mathrani et al [78] employed the "LightBot" game which required participants to control an animated robot to the end of a maze with the use of blocks representing commands that are similar to code. With a total of twenty participants, the consensus regarding the game was positive in which they found the game to be interesting, fun, and ranked the game as effective in instilling programming concepts like conditionals, functions, and recursion. Olsson and Mozelius [79] tested a memory game that involves matching values to particular data types and a syntax error "bombing" game on sixty-five total participants. Results from the study indicated that both games were successful in effectively helping students practice data types and Python syntax. More specifically, a total of forty-nine students found the games helpful for learning.

Wong and Yatim [80] developed a game called "The Odyssey of Phoenix" to aid in the learning of Object-Oriented Programming (OOP) concepts. This is done by mapping game processes to concepts. For example, the concept of inheritance is mapped to the crafting element in the game where resources required by both the main and nose gear can be shared since they both belong to the gear category. To measure the effectiveness of the game, pre and post-tests results were compared and analysed from 214 first-year students. Results from the study showed improvements in terms of learning effectiveness, leading to the conclusion that GBL is a great tool for knowledge acquisition. Papadakis and Kalogiannakis [81] developed a quiz platform based on multiplayer online role-playing games (MMORPG) where students can complete programming-based quests and challenges. By conducting surveys on thirty participants from a high school in Greece, the study concluded that the participants were engaged and motivated when completing the challenges presented to them. The authors concluded that GBL is useful in promoting learning outcomes for typically strenuous and difficult subjects. Furthermore, Oyelere et al [82] developed a "MobileEdu-puzzle" game for programming education. The game works by making students arranging misarranged lines of code. Over fifty-one students who participated in the study, seventy percent expressed that they were able to effectively learn programming with the experience.

These findings make GBL an ideal tool to motivate and increase participation when it comes to learning how to program. According to Santos et al [83], students appreciate the challenges that come with programming-based games as well as the achievement system to keep them motivated. Furthermore, Taylor et al [84] found that the development of computational thinking skills can also be enhanced by incorporating block-based programming into GBL applications. However, students with prior programming experience have found games with increasing difficulty levels to be tiring. It is therefore important to consider the participants' knowledge of programming when developing a programming-based game. Various approaches discussed in this section are summarised in Table I.

TABLE I. SUMMARY OF GAME-BASED LEARNING APPLICATIONS FOR COMPUTER PROGRAMMING EDUCATION

Author(s)	**Methodology**	**Outcome**
Mathrani et al [78]	Programming a robot with command blocks	Improved learning effectiveness and engagement
Olsson and Mozelius [79]	Memory game to teach data types	Improved learning effectiveness
Wong and Yatim [80]	Mapping OOP concepts to in-game activities	Improved learning effectiveness
Kalogiannakis [81]	Programming based quiz game	Improved learning effectiveness and engagement
Oyelere et al [82]	Programming based puzzle game	Improved learning effectiveness

F. Virtual Reality for Computer Programming Education

Alternatively, virtual reality can also be used to make programming classes more immersive and efficient. Pears et al [85] argued that visualisation is necessary to reduce the distance between the programming language and the student's mental models. For example, Vincur et al [86] combined both VR and game elements with block-based programming to introduce basic programming concepts called "Cubely". Cubely consists of cubes that represent programming concepts that can be arranged to form blocks of code. In other words, students must build code from code blocks to control the character to overcome challenges. Results from this study showed that over nineteen participants, eighteen preferred using Cubely due to the ease of use over typical code bootcamps available online.

Tanielu et al [87] developed a VR experience called "OOPVR" to reduce the abstractness of OOP concepts with analogies. To help students better understand the relationship between classes and objects, OOPVR uses a blueprint of a house to represent a class where many houses can be built from the same blueprint, signifying that multiple objects can be instantiated from a class. Various analogies were done in a similar manner throughout OOPVR to represent concepts like encapsulation, methods, and instances. To evaluate if the house analogy was effective in visualizing OOP concepts, the authors analysed results from a total of seventeen participants. Compared to the pre-questionnaires, post-questionnaires revealed that the participants showed higher confidence when it comes to visualizing OOP concepts. Bouali et al [88] developed a VR game called "Imikode" to help students familiarise themselves with OOP concepts. The system allows students to create virtual worlds with code. For instance, to instantiate an object, the command "fox = new Fox()" is used, creating a fox in the virtual world. The authors, however, have not yet done tests to determine the effectiveness of the system.

Consequently, Chen et al [89] developed a VR game that allows students to create levels to challenge their peers. Before starting the game, the student playing the role of the level creator will write codes in the virtual environment to place robot characters around the environment that will act as obstacles. Students who then play the game are required to acquire hints scattered around the virtual environment while overcoming the obstacles set up by the previous student. In a sense, students will be able to judge the effectiveness of their own code depending on how the system reacts. The authors claim that students of age nine to thirteen who tried the game have provided positive feedback in terms of engagement and learning effectiveness. Segura et al [90] developed a VR game called "VR-OCKS" that requires students to use code functions represented with blocks in the game to complete puzzles. To test the effectiveness of the system, twenty participants that have played VR-OCKS and another twenty participants who didn't play the game were recruited. Each group of twenty participants was further halved to complete challenges synonymous with the puzzles presented in VR-OCKS in "Kodu" and "Blockly", two popular systems that utilise visual programming to help students learn programming. Results indicated that those who played VR-OCKS before were able to complete twenty-five percent more levels than those who didn't. A summary of the approaches discussed in this section is shown in Table II.

G. Gender and Games

Throughout the years, many studies have been carried out regarding games and gender. The results of these studies have shown that gender preferences are a key motivator when it comes to playing specific game types [91], [92]. Considering how GBL has proven to be an effective tool for education, gender preference considerations are important in ensuring equal participation. Thus, the purpose of this section is to shed light on factors that appeal to a specific gender when it comes to games, specifically digital games. Lastly, game characteristics that are disliked by females will also be presented in this section.

To determine the preferred game design characteristics of different genders, Spieler and Slany [93] compiled programs developed by participants on the Pocket Code platform using a visual programming environment. In terms of game genre and themes, it is found that both females and males used the adventure genre in their games. However, males prefer the space theme whereas females prefer the nature-based themes. As for game mechanics, both genders used experience points as rewards and challenges in the form of missions in which there is an achievement system. The presence of leaderboards in male-created games is a key difference between the genders. This fact is supported by Hassouneh and Brengman [94] who argued that males value achievement more than females. Consequently, although challenges in games remain the main motivator for both genders, young men are more motivated by games that engage players to reach higher levels as well as beating the game [95]. When it comes to game aesthetics, the narrative, sensation, and fantasy elements are used by both genders but are more apparent in female-created games. The exploration element is, however, exclusive to games created by female pupils, where challenges of the game will lead to the discovery of new parts of the virtual world. This statement also coincides with the argument of Zhou et al [96], where games that include exploration and experimentation elements are generally preferred by female players.

TABLE II. SUMMARY OF VIRTUAL REALITY APPLICATIONS FOR COMPUTER PROGRAMMING EDUCATION

Author(s)	Methodology	Outcome
Vincur et al [86]	Code blocks to solve challenges	Users preferred this approach over traditional online bootcamps
Tanielu et al [87]	Mapping OOP concepts to in-game activities	Users showed higher confidence in understanding OOP concepts
Bouali et al [88]	Forming worlds virtually with code	No findings
Chen et al [89]	Level creation with code	Improved learning effectiveness and engagement
Segura et al [90]	Solving challenges with programming blocks	Improved learning effectiveness

TABLE III. SUMMARY OF FINDINGS REGARDING GENDER AND GAMES

Author(s)	Findings
Spieler and Slany [93]	Space and nature themes preferred by male and females correspondingly.
Hassouneh and Brengman [94]	Males value achievement more than females.
Lucas and Sherry [95]	Males focuses on reaching higher levels and beating the game.
Zhou et al [96]	Females prefer exploration and experimentation.
DeCamp [97]	Violence is preferred by males.
Hanh [98]	Females do not appreciate negative portrayals of female characters.

Most games portray female characters as weak individuals who are often rescued by males to attract players. As a result, females do not enjoy or avoid games because they do not appreciate the negative portrayals of female characters [97]. Similarly, violence in games is also one of the main factors contributing to a low number of female gamers. Whereas males are more interested in violence in games, females are less interested in violence [98]. In essence, the low rate of female participation in computer science means that there are more male game developers, who in turn, design and develop games that mainly attract males. This, in turn, lowers the number of female gamers that creates a negative cycle. Table III shows a summary of all the findings from the studies mentioned.

VII. Discussion

Inaccurate stereotypes that are instilled by society regarding computing have negatively affected females in many ways. Stereotypical depictions of computing foster negative attitudes towards the field, perpetuating the already large gender gap. This issue also affects females who are already in the field where they are often deemed as less capable compared to their male peers. This lowers their confidence levels and sense of belonging which often leads to student dropouts. Ultimately, this is the root of the issue that must be immediately resolved.

Furthermore, it is worth discussing how the use of GBL and virtual reality or combinations of both techniques can help solve issues related to the gender gap in computing (RQ1). In section VI A, it is seen that the utilisation of games in education is capable of increasing student motivation. When lessons are presented in the form of games, students are more willing to participate in lessons. As seen in section VI C, virtual reality enables students to experience unconventional situations that may be inaccessible in real life due to costs and safety reasons. Compared to learning from a textbook, the unconventional presentation of lessons in the virtual environment can also lead to increased participation. It is also important to note that the combination of both GBL and virtual reality techniques may solve some limitations related to one or both approaches. For example, some virtual learning experiences ([76] and [77]) were reported to be non-engaging. Considering that game-based applications are capable of engaging students, this problem can potentially be fixed by employing game elements into the virtual experience. Moreover, both GBL and virtual reality applications specifically developed for computing education have reported outcomes such as increased learners' satisfaction and general improvements to the learning process (Table I and Table II). Seeing how females often lose interest in computing due to boring and abstract lessons, the use of immersive games can be used to simulate how code can be used to solve real-world problems that are more interesting. Immersive games can also foster cooperation with multiplayer modes that are beneficial in increasing the sense of belonging of females in computing. Thus, this makes GBL and virtual reality or the combination of both techniques suitable for increasing female participation in the field of computer science.

Unfortunately, outcomes such as increasing female participation remain scarce in immersive programming games-based research. Existing applications highlighted in this review are mainly focused on learning syntax and programming concepts. While existing solutions are effective in retaining those, who are already enrolled in computing-based courses, it does not provide the necessary "appeal" to initiate female interest towards the field. Hence, we require solutions that are specially catered to increasing female participation as it is found that females are more likely to play games that are designed with them in mind [99]. To aid in fostering more research in this area, the "Gender and games" section aims to help future researchers employ game elements that are more appealing to a female audience (RQ2). As seen in Table III, some of the game elements that should be implemented to foster female interests when developing immersive games pertaining to computing education include implementing exploration, experimentation, space, and nature themes. Other than that, the application developed should not feature any violence, negative portrayals of female characters and should not focus too much on achievements.

VIII. Suggestions for Future Work

There are multiple takeaways from the literature reviewed in this paper. To increase the sense of belonging of females in the field of computing, immersive games developed especially for this purpose should focus on promoting confidence. Since the relationship between confidence and sense of belonging is directly proportional, future work should be done by keeping this fact in mind [100]. This may be done by accessing potential factors that demote confidence and using elements of immersive games to mitigate the issue. For example, [39] observed that females face a dip in confidence when they are unable to immediately observe the results of their work. A potential solution for this is to develop a system that allows users to instantly see the effects of their code writing as seen in Fig. 3. In both game and virtual environments, this can be achievable by changing or slightly altering in-game environments to match user-written code.

To ensure female students do not lose motivation when learning computer programming, retention is important. Hence, it would be worthwhile to develop frameworks that can effectively balance the "fun" and "learning" components of game-based applications. This is important in educational contexts as the system should be able to foster motivation to learn without being too distracting to users. In other words, in-game goals should not overpower learning outcomes. Fig. 4 shows a general flow diagram of how to determine the game elements that motivate but do not distract users to balance the "fun" and "learning" elements mentioned previously. Essentially, researchers should use the flowchart depicted in Fig. 4 to determine, first, if a specific game element incites motivation, then to determine whether that element is also non-distracting before deciding to implement that element in their system. As far as virtual reality is concerned, future works should be aimed at developing experiences for the masses. Regarding software, the use of low-poly graphics should be explored in such a way as to allow accessibility for those with lower-end peripherals. This is necessary to ensure that female students would be able to utilise these tools, regardless of their socio-economic backgrounds.

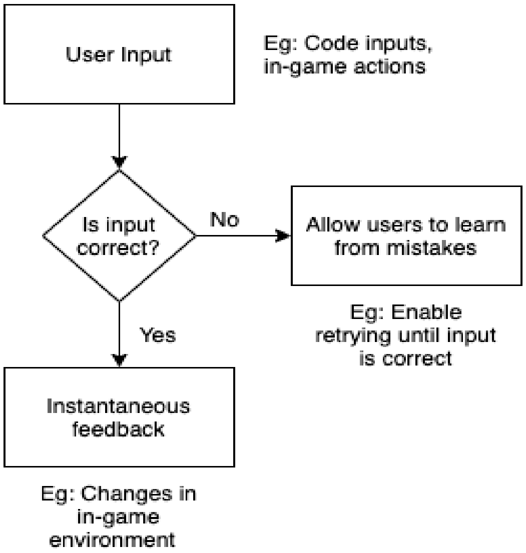

User Input
Eg: Code inputs, in-game actions
Is input correct?
No
Allow users to learn from mistakes
Eg: Enable retrying until input is correct
Yes
Instantaneous feedback
Eg: Changes in in-game environment

Fig. 3. Flowchart of a System that Allows for Instant Feedback.

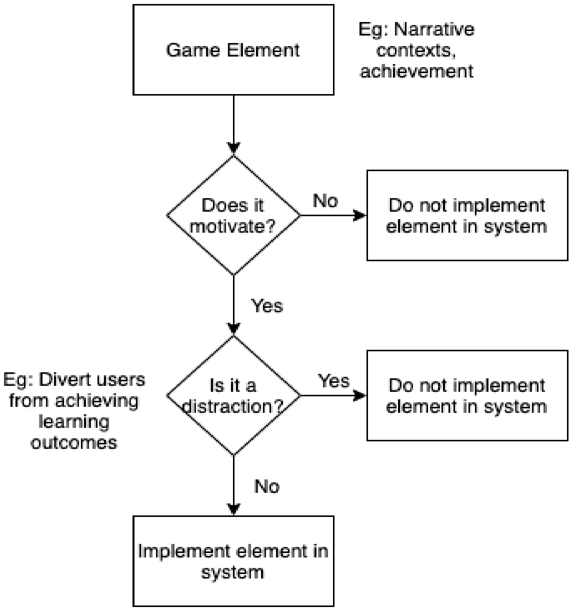

Game Element
Eg: Narrative contexts, achievement
Does it motivate?
No
Do not implement element in system
Yes
Eg: Divert users from achieving learning outcomes
Is it a distraction?
Yes
Do not implement element in system
No
Implement element in system

Fig. 4. Decision Flowchart Determines whether Game Elements are Capable of Inciting Motivation without being Distracting.

To aid female students who do not have prior programming knowledge, further research can be done to identify the issues faced by students when learning introductory programming. Specifically, determining why it is difficult for students to understand the concepts of variables, arrays, loops, and if statements. As it stands, most research concerns factors such as lack of intrinsic motivation, perception, and socio-economic backgrounds as a reason for students not to do well in computer programming. Hence, a deeper insight into how students perceive programming concepts without the use of immersive games is necessary to develop appropriate solutions.

Finally, to further increase female participation in the field of computer science, more research should be done to determine how different genders respond to the elements of the game in an educational context. For example, one can continue to determine how males and females react to a high scoreboard in an educational immersive game. Consequently, it is also possible to find out how the high score board impacts the confidence levels in both genders. By doing so, it will be possible to decide whether to include certain elements when developing the system. Specifically, this is useful for developing educational systems for fields of study with large gender gaps, as in the case of females in computing.

IX. Conclusion

This paper provides a review that highlights reasons affecting female participation in computing with a focus on how stereotypical perceptions of computers prevent females from seeking careers in computer science. Additionally, the advantages of GBL, such as the ability to foster intrinsic motivation and knowledge acquisition, are also discussed in detail, along with their disadvantages in educational settings. Existing game-based applications for computer education are also highlighted, looking at how they work and how they compare to each other. This review also sheds some light on existing virtual-game-based applications and how virtual reality is used in education, specifically in the computing field. Lastly, a summary of how males and females respond to different game elements and their preferences is also discussed.

Along with this, discussions and several suggestions have been made regarding further research that can be done. Suggestions include a flow-chart to increase female self-efficacy and sense of belonging to the use of computing immersive games, developing frameworks that balance “fun” and “learning” elements in immersive games, identifying how students struggle with understanding programming concepts and analysing how different genders react to certain game elements. The purpose of this review is to encourage more research in the development of immersive learning applications to reduce gender gaps in computing.

References

[1] W. Farrell, ‘The Education of Our Sons: A Look at the Educational State of Boys’, 2018. http://whitehouseboysmen.org/wp-content/uploads/2018/03/The-Education-of-Our-Sons.pdf

[2] ‘Women in STEM | Percentages of Women in STEM Statistics’, 2019. https://www.stemwomen.co.uk/blog/2019/09/women-in-stem-percentages-of-women-in-stem-statistics

[3] I. Isphording and P. Qendrai, ‘Gender Differences in Student Dropout in STEM’, IZA Institute of Labor Economics, 2019.

[4] K. Nolan, A. Mooney, and S. Bergin, ‘An Investigation of Gender Differences in Computer Science Using Physiological, Psychological and Behavioural Metrics’, in Proceedings of the Twenty-First Australasian Computing Education Conference on - ACE ’19, Sydney, NSW, Australia, 2019, pp. 47–55. doi: 10.1145/3286960.3286966.

[5] ‘AP Program Participation and Performance Data 2019’, 2019.

[6] A. Robinson and M. A. Pérez-Quiñones, ‘Underrepresented middle school girls: on the path to computer science through paper prototyping’, in Proceedings of the 45th ACM technical symposium on Computer science education - SIGCSE ’14, Atlanta, Georgia, USA, 2014, pp. 97–102. doi: 10.1145/2538862.2538951.

[7] S. Fayer, A. Lacey, and A. Watson, ‘STEM Occupations: Past, Present, And Future’, U.S. Bureau of Labor Statistics, 2017.

[8] ‘NCWIT Scorecard: The Status Of Women In Computing’, National Center for Women & Information Technology, 2018.

[9] M. H. Hussein, S. H. Ow, L. S. Cheong, M.-K. Thong, and N. Ale Ebrahim, ‘Effects of Digital Game-Based Learning on Elementary

Science Learning: A Systematic Review', IEEE Access, vol. 7, pp. 62465–62478, 2019, doi: 10.1109/ACCESS.2019.2916324.

[10] T. Mitamura, Y. Suzuki, and T. Oohori, 'Serious games for learning programming languages', in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea (South), Oct. 2012, pp. 1812–1817. doi: 10.1109/ICSMC.2012.6378001.

[11] M. A. Miljanovic and J. S. Bradbury, 'A Review of Serious Games for Programming', in Serious Games, vol. 11243, S. Göbel, A. Garcia-Agundez, T. Tregel, M. Ma, J. Baalsrud Hauge, M. Oliveira, T. Marsh, and P. Caserman, Eds. Cham: Springer International Publishing, 2018, pp. 204–216. doi: 10.1007/978-3-030-02762-9_21.

[12] J. P. da Silva and I. F. Silveira, 'A Systematic Review on Open Educational Games for Programming Learning and Teaching', Int. J. Emerg. Technol. Learn., vol. 15, no. 09, p. 156, May 2020, doi: 10.3991/ijet.v15i09.12437.

[13] S. Kavanagh, A. Luxton-Reilly, B. Wuensche, and B. Plimmer, 'A systematic review of Virtual Reality in education', Themes in Science and Technology Education, vol. 10, no. 2, pp. 85–119, 2017.

[14] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, 'A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda', Computers & Education, vol. 147, p. 103778, Apr. 2020, doi: 10.1016/j.compedu.2019.103778.

[15] N. Johnson, J. Garcia, and K. Seppi, 'Women in CS: Changing the Women or Changing the World?', in 2019 IEEE Frontiers in Education Conference (FIE), Covington, KY, USA, Oct. 2019, pp. 1–8. doi: 10.1109/FIE43999.2019.9028562.

[16] T. C. Peck, L. E. Sockol, and S. M. Hancock, 'Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research', IEEE Trans. Visual. Comput. Graphics, vol. 26, no. 5, pp. 1945–1954, May 2020, doi: 10.1109/TVCG.2020.2973498.

[17] 'Bureau of Labor Statistics Employment Projections', U.S. Bureau of Labor Statistics, 2019. [Online]. Available: https://www.bls.gov/emp/tables/emp-by-detailed-occupation.html

[18] 'Current Term Enrollment Estimates', National Student Clearinghouse Research Center, 2019.

[19] B. Khan, C. Robbins, and A. Okrent, 'The State of U.S. Science and Engineering 2020', SCIENCE & ENGINEERING INDICATORS, 2020.

[20] L. Wang, G. Stanovsky, L. Weihs, and O. Etzioni, 'Gender Trends in Computer Science Authorship', Allen Institute for Artificial Intelligence, Seattle, Washington, USA, 2019.

[21] J. Forman et al., 'Automobile injury trends in the contemporary fleet: Belted occupants in frontal collisions', Traffic Injury Prevention, vol. 20, no. 6, pp. 607–612, Aug. 2019, doi: 10.1080/15389588.2019.1630825.

[22] J. Munafo, M. Diedrick, and T. A. Stoffregen, 'The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects', Exp Brain Res, vol. 235, no. 3, pp. 889–901, Mar. 2017, doi: 10.1007/s00221-016-4846-7.

[23] J. Taylor, M. Harris, and S. Taylor, 'Parents Have Their Say...About Their What a Career Center Can Do For You College-Age Children's Career Decisions', Nace Journal, vol. 64, no. 2, pp. 15–21, 2004.

[24] J. L. Smith, K. L. Lewis, L. Hawthorne, and S. D. Hodges, 'When Trying Hard Isn't Natural: Women's Belonging With and Motivation for Male-Dominated STEM Fields As a Function of Effort Expenditure Concerns', Pers Soc Psychol Bull, vol. 39, no. 2, pp. 131–143, Feb. 2013, doi: 10.1177/0146167212468332.

[25] A. J. Hussain, L. Connell, H. Francis, D. Al-Jumeily, P. Fergus, and N. Radi, 'An Investigation into Gender Disparities in the Field of Computing', in 2015 International Conference on Developments of E-Systems Engineering (DeSE), Duai, United Arab Emirates, Dec. 2015, pp. 20–25. doi: 10.1109/DeSE.2015.17.

[26] M. N. Giannakos et al., 'Identifying dropout factors in information technology education: A case study', in 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, Apr. 2017, pp. 1187–1194. doi: 10.1109/EDUCON.2017.7942999.

[27] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, 'Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study', J. Res. Sci. Teach., p. n/a-n/a, 2010, doi: 10.1002/tea.20363.

[28] D. Baker, 'What Works: Using Curriculum and Pedagogy to Increase Girls' Interest and Participation in Science', Theory Into Practice, vol. 52, no. 1, pp. 14–20, Jan. 2013, doi: 10.1080/07351690.2013.743760.

[29] P. Kuhn and M.-C. Villeval, 'Are Women More Attracted to Cooperation Than Men?', National Bureau of Economic Research, Cambridge, MA, w19277, Aug. 2013. doi: 10.3386/w19277.

[30] W. Khalil, S. Nayab, T. Naeed, S. Khan, and S. Khalil, 'Female representation in computer science and information technology', in 2015 International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, Dec. 2015, pp. 1–10. doi: 10.1109/ICICT.2015.7469574.

[31] A. Hoegh and B. M. Moskal, 'Examining science and engineering students' attitudes toward computer science', in 2009 39th IEEE Frontiers in Education Conference, San Antonio, TX, USA, Oct. 2009, pp. 1–6. doi: 10.1109/FIE.2009.5350836.

[32] D. Gürer and T. Camp, 'An ACM-W literature review on women in computing', SIGCSE Bull., vol. 34, no. 2, p. 121, Jun. 2002, doi: 10.1145/543812.543844.

[33] Jo Tondeur, Sarah Van de Velde, Hans Vermeersch, and Mieke Van Houtte, 'Gender Differences in the ICT Profile of University Students: A Quantitative Analysis', DiGeSt. Journal of Diversity and Gender Studies, vol. 3, no. 1, p. 57, 2016, doi: 10.11116/jdivegendstud.3.1.0057.

[34] S. Cheryan, A. N. Meltzoff, and S. Kim, 'Classrooms matter: The design of virtual classrooms influences gender disparities in computer science classes', Computers & Education, vol. 57, no. 2, pp. 1825–1835, Sep. 2011, doi: 10.1016/j.compedu.2011.02.004.

[35] A. Fisher and J. Margolis, 'Unlocking the clubhouse: the Carnegie Mellon experience', SIGCSE Bull., vol. 34, no. 2, pp. 79–83, Jun. 2002, doi: 10.1145/543812.543836.

[36] C. Frieze and J. L. Quesenberry, 'How computer science at CMU is attracting and retaining women', Commun. ACM, vol. 62, no. 2, pp. 23–26, Jan. 2019, doi: 10.1145/3300226.

[37] W. IFIP TC9/WG9.1 International Conference on Women and Computerization, E. Balka, and R. Smith, Women, work, and computerization: charting a course to the future : IFIP TC9 WG9.1 Seventh International Conference on Women, Work, and Computerization, June 8-11, 2000, Vancouver, British Columbia, Canada. 2000. Accessed: Jan. 09, 2020. [Online]. Available: https://doi.org/10.1007/978-0-387-35509-2

[38] M.-P. Chen, 'The Effects of Prior Computer Experience and Gender on High School Students' Learning of Computer Science Concepts from Instructional Simulations', in 2010 10th IEEE International Conference on Advanced Learning Technologies, Sousse, Tunisia, Jul. 2010, pp. 610–612. doi: 10.1109/ICALT.2010.173.

[39] J. J. LaBouliere, A. Pelloth, C.-L. Lu, and J. Ng, 'An exploration of the attitudes of young girls toward the field of computer science', in 2015 IEEE Frontiers in Education Conference (FIE), Camino Real El Paso, El Paso, TX, USA, Oct. 2015, pp. 1–6. doi: 10.1109/FIE.2015.7344265.

[40] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco, 'Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study', Syst Rev, vol. 6, no. 1, p. 245, Dec. 2017, doi: 10.1186/s13643-017-0644-y.

[41] S. Erhel and E. Jamet, 'Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness', Computers & Education, vol. 67, pp. 156–167, Sep. 2013, doi: 10.1016/j.compedu.2013.02.019.

[42] H.-Y. Sung and G.-J. Hwang, 'A collaborative game-based learning approach to improving students' learning performance in science courses', Computers & Education, vol. 63, pp. 43–51, Apr. 2013, doi: 10.1016/j.compedu.2012.11.019.

[43] M. D. Kickmeier-Rust and D. Albert, 'Micro-adaptivity: protecting immersion in didactically adaptive digital educational games: Micro-adaptivity in digital educational games', Journal of Computer Assisted Learning, vol. 26, no. 2, pp. 95–105, Mar. 2010, doi: 10.1111/j.1365-2729.2009.00332.x.

[44] K. H. Li, S.-J. Lou, T.-F. Cheng, and H.-Y. Tsai, 'Application of Game-based Learning (GBL) on Chinese Language Learning in Elementary School', in 2012 IEEE Fourth International Conference On Digital Game

And Intelligent Toy Enhanced Learning, Takamatsu, Japan, Mar. 2012, pp. 226–230. doi: 10.1109/DIGITEL.2012.61.

[45] C. H. C. Yan and F. Tam, ‘Learning Chinese History through Digital Game’, in 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, Kaohsiung, Taiwan, Apr. 2010, pp. 156–160. doi: 10.1109/DIGITEL.2010.49.

[46] W. R. Watson, C. J. Mong, and C. A. Harris, ‘A case study of the in-class use of a video game for teaching high school history’, Computers & Education, vol. 56, no. 2, pp. 466–474, Feb. 2011, doi: 10.1016/j.compedu.2010.09.007.

[47] M. Ebner and A. Holzinger, ‘Successful implementation of user-centered game based learning in higher education: An example from civil engineering’, Computers & Education, vol. 49, no. 3, pp. 873–890, Nov. 2007, doi: 10.1016/j.compedu.2005.11.026.

[48] K. Harris and D. Reid, ‘The Influence of Virtual Reality Play on Children’S Motivation’, Can J Occup Ther, vol. 72, no. 1, pp. 21–29, Feb. 2005, doi: 10.1177/000841740507200107.

[49] M. Papastergiou, ‘Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation’, Computers & Education, vol. 52, no. 1, pp. 1–12, Jan. 2009, doi: 10.1016/j.compedu.2008.06.004.

[50] R. Silva, R. Rodrigues, and C. Leal, ‘Play it again: how game-based learning improves flow in Accounting and Marketing education’, Accounting Education, vol. 28, no. 5, pp. 484–507, Sep. 2019, doi: 10.1080/09639284.2019.1647859.

[51] V. Cezar, P. Garcia, V. Botelho, and E. Miletto, ‘Towards an RPG Game to Teach Calculus’, in 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Maceió, Brazil, Jul. 2019, pp. 116–118. doi: 10.1109/ICALT.2019.00037.

[52] S.-T. Yong, P. Gates, A. Chan, C.-S. Lee, R. Matthews, and K.-M. Tiong, ‘Exploring the Feasibility of Computer Games in Mathematics Education’, in 2019 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Subang Jaya, Malaysia, Oct. 2019, pp. 1–6. doi: 10.1109/HAVE.2019.8921018.

[53] J.-J. Lo and F.-M. Lin, ‘A Study of 2nd Grade Students’ Attitude on a Mathematics Game’, in 2012 IEEE Fourth International Conference On Digital Game And Intelligent Toy Enhanced Learning, Takamatsu, Japan, Mar. 2012, pp. 111–113. doi: 10.1109/DIGITEL.2012.29.

[54] N. J. Hogle, W. D. Widmann, A. O. Ude, M. A. Hardy, and D. L. Fowler, ‘Does Training Novices to Criteria and Does Rapid Acquisition of Skills on Laparoscopic Simulators Have Predictive Validity or Are We Just Playing Video Games?’, Journal of Surgical Education, vol. 65, no. 6, pp. 431–435, Nov. 2008, doi: 10.1016/j.jsurg.2008.05.008.

[55] D. Stefanidis, M. W. Scerbo, C. Sechrist, A. Mostafavi, and B. T. Heniford, ‘Do novices display automaticity during simulator training?’, The American Journal of Surgery, vol. 195, no. 2, pp. 210–213, Feb. 2008, doi: 10.1016/j.amjsurg.2007.08.055.

[56] C. Green and D. Bavelier, ‘Enumeration versus multiple object tracking: the case of action video game players’, Cognition, vol. 101, no. 1, pp. 217–245, Aug. 2006, doi: 10.1016/j.cognition.2005.10.004.

[57] J. Feng, I. Spence, and J. Pratt, ‘Playing an Action Video Game Reduces Gender Differences in Spatial Cognition’, Psychol Sci, vol. 18, no. 10, pp. 850–855, Oct. 2007, doi: 10.1111/j.1467-9280.2007.01990.x.

[58] C. P. Barlett, C. L. Vowels, J. Shanteau, J. Crow, and T. Miller, ‘The effect of violent and non-violent computer games on cognitive performance’, Computers in Human Behavior, vol. 25, no. 1, pp. 96–102, Jan. 2009, doi: 10.1016/j.chb.2008.07.008.

[59] J. Piaget, Play, dreams, and imitation in childhood. London: Routledge & K. Paul, 1967. Accessed: May 21, 2021. [Online]. Available: http://books.google.com/books?id=-9F-AAAAMAAJ

[60] M. Kickmeier-Rust, D. Schwarz, D. Albert, D. Verpoorten, J. Castaigne, and M. Bopp, ‘The ELEKTRA project: Towards a new learning experience’, M3-Interdisciplinary Aspects on Digital Media & Education, pp. 19–48, 2006.

[61] M. Pohl, M. Rester, and P. Judmaier, ‘Interactive Game Based Learning: Advantages and Disadvantages’, in Universal Access in Human-Computer Interaction. Applications and Services, vol. 5616, C. Stephanidis, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 92–101. doi: 10.1007/978-3-642-02713-0_10.

[62] D. Gibson, C. Aldrich, and M. Prensky, Eds., Games and Simulations in Online Learning: Research and Development Frameworks. IGI Global, 2007. doi: 10.4018/978-1-59904-304-3.

[63] J. Huizenga, W. Admiraal, and S. Akkerman, ‘Learning history by playing a mobile city game’, in Proceedings of the 1st European Conference on Game-Based Learning (ECGBL) October 2007, University of Paisley, Paisley, Scotland, 2007, pp. 127–134.

[64] F. M. Dinis, A. S. Guimaraes, B. R. Carvalho, and J. P. P. Martins, ‘Virtual and augmented reality game-based applications to civil engineering education’, in 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, Apr. 2017, pp. 1683–1688. doi: 10.1109/EDUCON.2017.7943075.

[65] F. Kharvari and W. Hohl, ‘The Role of Serious Gaming using Virtual Reality Applications for 3D Architectural Visualization’, in 2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Vienna, Austria, Sep. 2019, pp. 1–2. doi: 10.1109/VS-Games.2019.8864576.

[66] M. Boyles, J. Rogers, K. Goreham, M. A. Frank, and J. Cowan, ‘Virtual Simulation for Lighting &#x00026; Design Education’, in 2009 IEEE Virtual Reality Conference, Lafayette, LA, Mar. 2009, pp. 275–276. doi: 10.1109/VR.2009.4811052.

[67] G. Zhang, ‘Virtual Simulation for History Education’, in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, Mar. 2019, pp. 1646–1651. doi: 10.1109/VR.2019.8797734.

[68] J.-L. Lugrin et al., ‘A Location-Based VR Museum’, in 2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Wurzburg, Sep. 2018, pp. 1–2. doi: 10.1109/VS-Games.2018.8493404.

[69] N. R. Caluya and M. E. C. Santos, ‘Kantenbouki VR: A Virtual Reality Authoring Tool for Learning Localized Weather Reporting’, in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, Mar. 2019, pp. 866–867. doi: 10.1109/VR.2019.8798216.

[70] J. H. Seo et al., ‘Anatomy builder VR: Applying a constructive learning method in the virtual reality canine skeletal system’, in 2017 IEEE Virtual Reality (VR), Los Angeles, CA, USA, 2017, pp. 399–400. doi: 10.1109/VR.2017.7892345.

[71] A. de Mauro, J. Raczkowsky, M. E. Halatsch, and H. Worn, ‘Virtual Reality Training Embedded in Neurosurgical Microscope’, in 2009 IEEE Virtual Reality Conference, Lafayette, LA, Mar. 2009, pp. 233–234. doi: 10.1109/VR.2009.4811031.

[72] P. Rajeswaran, T. Kesavadas, P. Jani, and P. Kumar, ‘AirwayVR: Virtual Reality Trainer for Endotracheal Intubation-Design Considerations and Challenges’, in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, Mar. 2019, pp. 1130–1131. doi: 10.1109/VR.2019.8798249.

[73] J. Dong, J. Zhang, X. Ma, P. Ren, Z. C. Qian, and Y. V. Chen, ‘Virtual Reality Training with Passive Haptic Feedback for CryoEM Sample Preparation’, in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, Mar. 2019, pp. 892–893. doi: 10.1109/VR.2019.8797918.

[74] A. S. Mathur, ‘Low cost virtual reality for medical training’, in 2015 IEEE Virtual Reality (VR), Arles, Camargue, Provence, France, Mar. 2015, pp. 345–346. doi: 10.1109/VR.2015.7223437.

[75] Lei Wei, Hailing Zhou, A. K. Soe, and S. Nahavandi, ‘Integrating Kinect and haptics for interactive STEM education in local and distributed environments’, in 2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Wollongong, NSW, Jul. 2013, pp. 1058–1065. doi: 10.1109/AIM.2013.6584234.

[76] I. Y. T. Hsiao, I. Y. S. Li, Y. J. Lan, J. J. S. Huang, and S. J. H. Ynag, ‘Development of a virtual campus on Second Life: A case study of NCU wonderland’, in Workshop Proceedings of the 18th International Conference on Computers in Education, Putrajaya, Malaysia, 2010, pp. 241–247.

[77] P. H. Hsieh, Y. H. Wu, and F. M. Ma, ‘A study of visitor’s learning needs and visit satisfaction in real and Second Life museums’, in Workshop Proceedings of the 18th International Conference on Computers in Education, ICCE 2010, Putrajaya, Malaysia, 2010, pp. 248–255.

[78] A. Mathrani, S. Christian, and A. Ponder-Sutton, ‘PlayIT: Game Based Learning Approach for Teaching Programming Concepts’, Educational Technology & Society, vol. 19, no. 2, pp. 5–17, 2016.

[79] M. Olsson and P. Mozelius, ‘Learning to Program by Building Learning Games’, in Proceedings of the the 11th European Conference on Game-Based Learning ECGBL 2017, 2017, pp. 448–454.

[80] Y. S. Wong and M. H. M. Yatim, ‘A Propriety Multiplatform Game-Based Learning Game to Learn Object-Oriented Programming’, in 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), Yonago, Japan, Jul. 2018, pp. 278–283. doi: 10.1109/IIAI-AAI.2018.00060.

[81] S. Papadakis and M. Kalogiannakis, ‘Evaluating the effectiveness of a game-based learning approach in modifying students’ behavioural outcomes and competence, in an introductory programming course. A case study in Greece’, IJTCS, vol. 10, no. 3, p. 235, 2019, doi: 10.1504/IJTCS.2019.102760.

[82] S. S. Oyelere, F. J. Agbo, I. T. Sanusi, A. A. Yunusa, and K. Sunday, ‘Impact of Puzzle-Based Learning Technique for Programming Education in Nigeria Context’, in 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Maceió, Brazil, Jul. 2019, pp. 239–241. doi: 10.1109/ICALT.2019.00072.

[83] A. L. dos Santos, M. R. de A. Souza, M. Dayrell, and E. Figueiredo, ‘Exploring Game Elements in Learning Programming: An Empirical Evaluation’, in 2018 IEEE Frontiers in Education Conference (FIE), San Jose, CA, USA, Oct. 2018, pp. 1–9. doi: 10.1109/FIE.2018.8658505.

[84] S. Taylor et al., ‘Position: IntelliBlox: A Toolkit for Integrating Block-Based Programming into Game-Based Learning Environments’, in 2019 IEEE Blocks and Beyond Workshop (B&B), Memphis, TN, USA, Oct. 2019, pp. 55–58. doi: 10.1109/BB48857.2019.8941222.

[85] A. Pears et al., ‘A survey of literature on the teaching of introductory programming’, SIGCSE Bull., vol. 39, no. 4, p. 204, Dec. 2007, doi: 10.1145/1345375.1345441.

[86] J. Vincur, M. Konopka, J. Tvarozek, M. Hoang, and P. Navrat, ‘Cubely: virtual reality block-based programming environment’, Association for Computing Machinery, pp. 1–2, 2017.

[87] T. Tanielu, R. ’Akau’ola, E. Varoy, and N. Giacaman, ‘Combining Analogies and Virtual Reality for Active and Visual Object-Oriented Programming’, in Proceedings of the ACM Conference on Global Computing Education - CompEd ’19, Chengdu,Sichuan, China, 2019, pp. 92–98. doi: 10.1145/3300115.3309513.

[88] N. Bouali, E. Nygren, S. S. Oyelere, J. Suhonen, and V. Cavalli-Sforza, ‘Imikode: A VR Game to Introduce OOP Concepts’, in Proceedings of the 19th Koli Calling International Conference on Computing Education Research - Koli Calling ’19, Koli, Finland, 2019, pp. 1–2. doi: 10.1145/3364510.3366149.

[89] J. Chen, M. R. Zargham, M. Rajendren, and J. Cheng, ‘Coding VR Games’, in Int’l Conf. Frontiers in Education: CS and CE, 2019, pp. 123–127.

[90] R. J. Segura, F. J. Pino, C. J. Ogáyar, and A. J. Rueda, ‘VR-OCKS: A virtual reality game for learning the basic concepts of programming’, Comput Appl Eng Educ, vol. 28, no. 1, pp. 31–41, Jan. 2020, doi: 10.1002/cae.22172.

[91] L. Vermeulen and J. Van Looy, ‘“I Play So I Am?” A Gender Study into Stereotype Perception and Genre Choice of Digital Game Players’, Journal of Broadcasting & Electronic Media, vol. 60, no. 2, pp. 286–304, Apr. 2016, doi: 10.1080/08838151.2016.1164169.

[92] B. Manero, J. Torrente, C. Fernandez-Vara, and B. Fernandez-Manjon, ‘Investigating the Impact of Gaming Habits, Gender, and Age on the Effectiveness of an Educational Video Game: An Exploratory Study’, IEEE Trans. Learning Technol., vol. 10, no. 2, pp. 236–246, Apr. 2017, doi: 10.1109/TLT.2016.2572702.

[93] B. Spieler and W. Slany, ‘Game Development-Based Learning Experience: Gender Differ- ences in Game Design’, Sophia Antipolis, France, 2018, pp. 616–625.

[94] D. Hassouneh and M. Brengman, ‘A motivation-based typology of social virtual world users’, Computers in Human Behavior, vol. 33, pp. 330–338, Apr. 2014, doi: 10.1016/j.chb.2013.08.012.

[95] K. Lucas and J. L. Sherry, ‘Sex Differences in Video Game Play:: A Communication-Based Explanation’, Communication Research, vol. 31, no. 5, pp. 499–523, Oct. 2004, doi: 10.1177/0093650204267930.

[96] Z. Zhou, X.-L. Jin, D. R. Vogel, Y. Fang, and X. Chen, ‘Individual motivations and demographic differences in social virtual world uses: An exploratory investigation in Second Life’, International Journal of Information Management, vol. 31, no. 3, pp. 261–271, Jun. 2011, doi: 10.1016/j.ijinfomgt.2010.07.007.

[97] S. Hahn, ‘Researching the Gender Divide of Digital Games: How to Overcome the Virtuous Cycle of the Games Industry’, Acta Ludica, vol. 2, pp. 6–24, 2018.

[98] W. DeCamp, ‘Who plays violent video games? An exploratory analysis of predictors of playing violent games’, Personality and Individual Differences, vol. 117, pp. 260–266, Oct. 2017, doi: 10.1016/j.paid.2017.06.027.

[99] C. Stewart-Gardiner, G. Carmichael, J. Latham, N. Lozano, and J. Greene, ‘Influencing middle school girls to study computer science through educational computer games’, Journal of Computing Sciences in Colleges, vol. 28, no. 6, pp. 90–97, 2013.

[100] A. Nguyen and C. M. Lewis, ‘Competitive Enrollment Policies in Computing Departments Negatively Predict First-Year Students’ Sense of Belonging, Self-Efficacy, and Perception of Department’, in Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland OR USA, Feb. 2020, pp. 685–691. doi: 10.1145/3328778.3366805.

Fairness Embedded Adaptive Recommender System: A Conceptual Framework

Alina Popa
The Doctoral School of Marketing
The Bucharest Academy of Economic Studies, Bucharest, Romania

***Abstract*—In the current fast paced and constantly changing environment, companies should ensure that their way of interacting with user is both relevant and highly adaptive. In order to stay competitive, companies should invest in state-of-the-art technologies that optimize the relationship with the user using increasingly available data. The most popular applications used to develop user relationship are Recommender Systems. The vast majority of the traditional recommender system considers recommendation as a static procedure and focus on a specific type of recommendation, being not very agile in adapting to new situations. Also, when implementing a Recommender System there is the need to ensure fairness in the way decisions are made upon customer data. In this paper, it is proposed a novel Reinforcement Learning-based recommender system that is highly adaptive to changes in customer behavior and focuses on ensuring both producer and consumer fairness, Fairness Embedded Adaptive Recommender System (FEARS). The approach overcomes Reinforcement Learning's main drawback in recommendation area by using a small, but meaningful action space. Also, there are presented two fairness metrics, their calculation and adaptation for usage with Reinforcement Learning, this way ensuring that the system gets to the optimal trade-off between personalization and fairness.**

Keywords—Algorithmic fairness; reinforcement learning; recommender systems; system adaptability

I. Introduction

In the current constantly changing business environment, companies are required to respond appropriately to challenges that appear and adapt quickly to customers new needs and expectations in order to stay “top of mind” with prospects and clients. In retailing, a family of applications called Recommender Systems (RecSys) can help businesses stay relevant to their customers by leveraging the existing data about users and/or different items in order to help users find the right item for them [1]. One disadvantage of the current approaches of RecSys like Collaborative filtering or Content-based recommendations is that these strategies consider only the two elements, users and items when delivering recommendations, making impossible to detect important patterns that include other elements and to adapt it to the context or changing environment. Also, each of the recommendation approaches has its own limitations. Items recommended through Content-based filtering are always similar to the items previously bought or consumed by the user [2], while Collaborative filtering provides a good solution only under static scenarios when there are many users that bought or consumed the same product [3]. Hybrid recommender systems combine two or more recommendation strategies in different ways to benefit from their complementary advantages [4] and overcome the limitations of individual components. Another limitation of RecSys, regardless of strategy is the assumption that user's underlying preferences remains unchanged, thus the recommendation procedure is a static process [5, 6].

One of the best-known approaches that allows to include adaptability in a system is Reinforcement Learning (RL) [7-9]. There is a series of publications that explore the usage of RL in the area of RecSys. Out of which there are those that focus on user-item interaction sequence or user's browsing history and use it to create a state that later is fed to the RL model [10-15]. A different approach is to use user and item sets which are obtained from bi-clustering as environmental states [6]. An earlier paper is using both user information and item information vectors and refers to it as context [16]. Important work on integrating negative influence of irrelevant recommendations is done by using negative rewards [12, 13, 15, 17].

Machine Learning technologies are increasing their presence in our daily lives and have serious implication and influence in a lot of areas. Because these systems are used in many sensitive situations and can lead to life-changing, high stake decisions, it is crucial that algorithms' outputs do not reflect discriminatory behavior. Which is why, alongside with developing new ML use-cases, researchers have done efforts to formalize fairness and sources of bias in ML [18-27], propose fairness evaluation metrics [18, 19, 28-33] and processing methods to mitigate undesirable biases in relation to the proposed definitions [19, 28, 31, 34-37].

It is worth noting that based on literature review, there are no works as per knowledge of the author that explore simultaneously the following elements of an efficient and fair recommender system: 1) focus on customer relationship development 2) adaptability to new situations 3) optimization for long term customer engagement using negative rewards where appropriate and 4) awareness of both consumer and provider fairness.

Thus, in this paper, it is presented the design of a Fairness Embedded Adaptive Recommender System (FEARS) that has its main aim the development of customer relationship. The conceptual framework is combining multiple recommendation strategies through leveraging extensive information about user, items and context. The recommendation strategies are combined and used as the action space by a RL Agent. This way, the system has the ability to automatically learn the optimal policy through trial-and-error; by recommending and receiving reinforcements from user's feedback. This will allow the system to quickly adapt to the changing needs of the customers and will try to come up with a more long-term recommendation strategy to build a fruitful relationship with the client. Not least, the rewards of the RL engine are defined in such a way that both consumer and provider fairness is being ensured.

II. Related Work

In this section, firstly it is described the basic problem of Recommender Systems. Next, each of the recommender strategies specifics, latest advances and limitation are presented. Then, it is introduced the Reinforcement Learning practice and an analysis of it's to date usage and limitations in the recommendation area.

Also, an exploration of the different fairness formalizations for ML and recommendation systems is done in order to give an overview on how these definitions are translated into implementation.

A. *Recommender Systems (RecSys)*

In the human decision-making process, obtaining recommendations from trusted sources is a critical component. Usually, this role is played my family, friends or subject-matter experts. The goal of a recommender system is to create and give relevant recommendations of items or products to users. Depending on the structure of the learning system, traditionally there are distinguished following types of systems[4]:

- Collaborative Filtering: In this type of systems, a user is recommended items based on the previous ratings of the users that bought/ used the product.
- Content-based Filtering: These systems recommend items that are similar to items the user has liked in the past.
- Hybrid approaches: These methods try to combine both collaborative and content-based approaches into one in order to overcome the individual limitations of each of the approaches.

Currently, the architecture of RecSys and their evaluation on real-world problems is an active area of research.

B. *Collaborative Filtering*

Collaborative filtering (CF) systems collect user feedback in the form of ratings or ranks and makes recommendations to the active user based on items that other users with similar preferences liked in the past [38].

The aim of any recommendation system is to suggest elements that are relevant to users by extracting latent variables [39-42].

Latest advances in the field include using graph encoding, Stochastic Shared Embeddings, large-scale Pairwise Collaborative Ranking, Sequential Recommendation Via Personalized Transformer [43] that mainly solve the problem of scaling to massive datasets, learn user and item embeddings and think about the problem as a sequence of actions, not one-shot recommendations. The user and items embeddings are mainly a different way to refer to latent variables and a series of work leverage the power of Neural Networks to try and learn them [44-46].

C. *Content-Based Filtering*

The intuition behind a content-based recommendation is to suggest to a customer a product similar to those the user has previously purchased. Method tries to extract similar objects. There are two main types of measures used to estimate this relationship: measures of distance and measures of similarity between objects [47].

Most of the advances in the content-based recommendation area is based on finding best ways to represent an item through a vector, or, in other words, get their embeddings [48-53].

D. *Recommendation of Complementary Products*

*W*hen recommending complementary products, the system tries to leverage the transactions history of customers [54, 55] through Association Methods like APRIORI.

The APRIORI algorithm remained essentially unchanged since its introduction to the research community, although there are sporadic efforts to extend it [56, 57].

E. *Reinforcement Learning (RL)*

Reinforcement Learning is an area of machine learning that has been inspired by behavioural psychology. The field focuses on how a software agent (hereinafter agent) should take actions and how to interact with an environment so as to maximize a total reward function.

An agent can interact with the environment and learn through trial and error, just like humans and animals. Every action that the agent performs in an environment influences the future state of the agent. Also, each action is rewarded with a reward, and this is the only response the learner receives [58]. The mechanism that generates the reward and the transition from one state of the agent to another refers to the dynamics of the environment [47].

The agent's goal is to maximize his total long-term reward in the way he responds to his environment. This can happen if an agent explores the environment and tries to learn its dynamics.

Formally, the environment is a mathematical model known as the Markov Decision Process (MDP) encountered primarily in dynamic programming. The difference between the classical methods of dynamic control and RL is that the latter does not know the MDP model and can be used if these processes are very complex and other methods are unfeasible [59]. The basic MDP model contains the following components:

- A set of environmental states S1, ..., Sn ∈ S: These can refer to the inherent characteristics of the agent or objects that surround and interact with it.

- A set of actions that the agent can take, A1, ..., Am ∈ A: These refer to all possible actions that the agent controls.
- Transition function from one state to another: Being a Markov process, the next state of the system depends only on its previous state and the action taken, not on the whole history.
- The reward function represents the value of the reward obtained after acting with At in St.

An agent is a computer program that is able to observe and interact with the environment defined by the MDP. The agent perceives the environment as a set of observations that define a state. The agent interacts with the environment in a feedback loop pattern by following the steps below:

1) The agent observes the characteristics of the environment that define the current state, St.

2) The agent chooses an action from the set of possible actions, At, with which it responds to the environment in the current state St.

3) The agent enters a waiting state until the characteristics of the environment change with the St + 1 state and the agent receives the Rt+1 reward.

4) Steps 1-3 are then repeated.

The agent's behaviour or the way he interacts with the environment is described by a function called action policy or simply policy [60]. It specifies the actions to be taken when the agent is in a certain state. The agent's learning goal is to find a policy that maximizes the total reward.

F. Recommender Systems using Reinforcement Learning

As previously mentioned, in the literature there are already RecSys that include an RL engine. It is useful to formalize the problem of RL in the RecSys area and see the differences in the approach of the different research.

As mentioned above, formally, the RL problem can be defined as a mathematical MDP model. For that it is needed to specify the States, Actions and Rewards.

States are defined in different ways in the existing literature. They can reflect a mapping of previous user-item interactions into a hidden state [15], user’s recommendation and ad browsing history [13], previous items that a user clicked [12], the sequence of visited and recommended items [10] or a more detailed interaction sequence that contains clicking, purchasing, or skipping, leaving [14]. An interesting approach is to define states as the cluster resulted from the co-clustering or biclustering of users and items [6] or to extend the state to include user demographics [5]. Efforts are as well invested in how to best represent the state in a RL RecSys [61]. Currently, and as to the knowledge of the author, in the current literature there is no approach where the recommendation context, user demographics, behavioral patterns and recent browsing/interaction history is taken into account in the state definition.

Actions are mostly defined as selecting an item to be recommended from the whole discrete action space which contains the candidate items [12, 14, 15] or even whether to give a recommendation or not, and if yes, what would be the item to recommend [13]. There are authors that consider recommending a list of items [5, 11, 61]. One of the most different approaches it to recommend items from neighboring clusters to the user-items one [6]. As mentioned in multiple articles [62, 63], RL in RecSys has a common issue of efficiency that comes from the fact that the action space is too large, consisting of all candidate items, and thus huge amount of interaction data is required for learning an optimal policy.

The reward function is heavily dependent on user feedback and actions he takes, for example user can click or purchase a recommended item and receive a positive reward or to skip it and get a different reward value [10-13, 15]. Reward can consist of immediate user feedback, but as well as a longer-term objective [14]. Most of the rewards are not deterministic and depend very much on how the user is reacting, but there are also formulations when this is seen deterministically as the Jaccard distance between the user vectors of the time t and t+1 state [6].

It is important to note the research direction as well towards using negative rewards. This can help the learning agent into searching for a policy that would be appropriate for overcoming the information fatigue [12, 13, 15, 17].

G. Fairness in Machine Learning

In the same way as people, algorithms are vulnerable to biases that exist in data and can lead to an unfair decision or outcome. More than 20 types of biases in ML were extracted, categorized and explained by researchers [24, 26] in order to motivate and accelerate the process of mitigating them.

Putted simple, in the context of high stakes decision-making, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics that are considered sensitive variables. Thus, a fair algorithm is one whose decisions are not skewed towards a particular group of people. GDPR, UK Equality Act, Fair Housing Act and Equal Credit Opportunity Act define protected classes such as race, gender, age or disability and state the fairness and equality principles [64].

The most simple and straightforward definition of fairness is “fairness through unawareness”: “A ML model is said to achieve fairness through unawareness if protected attributes are not explicitly used in the prediction process”.

Although these variables are not used in developing the ML model, this doesn’t mean that the information cannot be retrieved from other variables. Chiappa & Isaac [65] emphasize that fairness should be expressed both in terms of sensitive variables, but also considering corelated or proxy variables. Not considering these proxy variables has been shown to increase the risk of discrimination [27].

Mehrabi et al. [66] distinguish three different types of fairness definitions: 1) Individual Fairness where the system should give similar outputs to similar individuals, 2) Group Fairness where ML system treats different groups equally and 3) Subgroup Fairness which intends to obtain the best properties of the group and individual notions of fairness.

Group fairness equal treatment can be in turn defined through [66, 37]: 1) Equal Opportunity where the probability of a person from a positive outcome class of being assigned a positive outcome should be equal for both protected and unprotected group members, 2) Demographic Parity where the likelihood of a positive outcome should be the same regardless of whether the person is in the protected group or not, 3) Disparate Impact considers the ratio between unprivileged and privileged groups likelihood of a positive outcome. Disparate impact uses the “not less than 80%” rule to define if a process has disparate impact or not.

Main approaches for tackling unfairness are differentiated into three groups: 1) pre-processing, 2) in-processing and 3) post-processing [37]. Pre-processing methods are extracting representations from the data in order to remove undesired biases [27]. Then, this unbiased data is used for model development. Some of the methods in this family are adversarial learning, causal methods, relabeling, perturbations, resampling, reweighing, transformation and variable blinding [27]. The in-processing methods are constraining a model to produce fair outputs by including fairness into the learning mechanism like adversarial learning, bandits, constraint optimization, regularization or reweighing [27]. The post-processing methods are working with model outputs to make them fair using calibration, thresholding and transformation approaches [27].

H. *Fairness in Recommender Systems*

As presented in a previous section, RecSys make recommendations to support decision making by studying user behavior and historical patterns. Because it is widely applied in various fields like recommending music, books, people to hire or jobs, an impartial view of the system towards any of the involved sides can be detrimental [67, 68].

Since these systems use past data, they are also inheriting the 1) Historical Bias, which is the already existing bias in the world [26], 2) Representation Bias when the used sample from a population is not representative for the whole population [26] and 3) Social Bias when other people’s actions or generated content affect another person’s opinions [69]. Alongside with these biases, the system itself displays: 1) Popularity Bias when items that are more popular tend to be exposed more [70], 2) Algorithmic Bias when the bias is not present in the input data and is added later by the algorithm in the way it works [69], 3) Presentation Bias when the way items are presented impacts the attraction those items get (e.g., users can click only what they see, thus, items presented more often will get more clicks) [69] and 4) Ranking Bias when top ranked items are perceived as more interesting and thus, receive more traffic [71]. In Fig. 1 it is shown how different types of biases feed each other in a RecSys.

Considerations of fairness have been actively studied in the context of recommender systems. Burke [72] introduced the multisided view of fairness in recommender systems. In the case of recommendations, the system is facilitating a transaction between parties [73]. Fairness towards all the involved parties is important and a balanced point should be found. Burke et al. [74] divide stakeholders of any given recommender system into three categories: 1) Consumers are the individuals that receive recommendations 2) Providers are those that stay behind the recommended items or products and gain from consumer’s choices and 3) System is the platform itself that tries to match providers with consumers and by doing this in a successful way is gaining benefits.

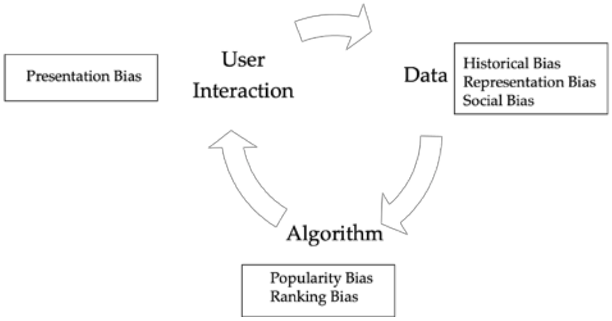


Fig. 1. Different Types of Biases that Appear in a Recommender System. Source: Adaptation after Fig. 2 [115].

Recommender system’s objective for a consumer is to give the best items for his needs, through personalization, in such a way that these items are not constraining him in getting a higher overall utility compared to people from other groups, thus in a fair way. For a provider, recommender system needs to ensure that his items get sufficient exposure and that items are shown to the consumers that have the highest probability of buying or consuming them, thus in a relevant way. Platform’s utility is also important, because this is the initial motivation of having the recommender system in place. A key issue that arises in recommender systems is the tension between a personalized view of recommendation delivery and fairness objectives [74, 75].

Provider’s fairness in recommender systems is typically defined for the objects or subjects to be ranked. It has been explored and formalized as:

1) the bound of the number of items related to each of the protected attributes that are allowed to appear in the top k positions of the ranking [76],

2) a sufficient presence of items belonging to different groups [77],

3) a consistent treatment of similar items [77],

4) a proper representation of items from both protected and unprotected groups [77].

5) exposure, disparate exposure and group fairness disparity; all three proportional to the merit of the item defined as relevance to the query [78],

6) pairwise fairness that expresses the likelihood of a clicked item being ranked above another relevant unclicked item is the same across both groups [79],

7) pairwise statistical parity represents that if two candidates from different groups are compared, then on average each group has an equal chance of being top ranked [79],

8) set-based fairness at discrete points in the ranking with logarithmic discount that emphasize the fact that fairness at top ranks is more important than at lower ranks [80], *9)* difference in acceptance rates measures whether a relevant item from the advantaged and disadvantaged class are accepted at the same rates [81].

In the area of consumer fairness, there are the following metrics that can be used:

1) value unfairness which occurs when one group of users is consistently given higher or lower predictions than their true preferences [82]. Value unfairness becomes large when predictions for one group are consistently overestimated and predictions for the other group are underestimated.

2) absolute unfairness, which measures inconsistency in absolute estimation error across different user groups [82]. This means that the advantaged group has the unfair advantage of good recommendations, while the other groups have poor recommendations.

3) non-parity unfairness is computed as the absolute difference between the overall average ratings of disadvantaged users and those of advantaged users for recommended items [82].

4) balanced neighborhood that expresses the fact that recommendations for all users are generated from neighborhoods that are balanced with respect to the protected and unprotected classes [74].

5) disparate impact of recommendation [74, 37].

Overall, the consumer fairness is less represented in the literature.

There are efforts in the area of mitigating bias and ensuring fairness in recommender systems by using regularization terms [82, 83, 84], reinforcement learning [78, 85] and neighborhood balancing [74].

Although the literature of methods is rich in methods to mitigate unfairness, not all of them are applicable to the dynamic nature of recommender systems. Ge et al. [85] show that by enforcing fair decisions through static fairness criteria metrics, the system leads to unexpected unfairness in the long run and that fairness cannot be defined in a static setting without considering the long-term impact and evolution. Same as with the need to bring adaptability into the recommender system results, the need to have a dynamic view over fairness can be solved by using reinforcement learning.

The practice of using Reinforcement Learning to ensure fairness is an emerging research area [27]. In terms of implementation, fairness dimension can be given to the RL agent as a reward that can be positive in the case of fair outputs and negative otherwise [86, 87]. Other approaches construct the problem as a Constrained Markov Decision Process [85].

III. Proposed Approach

In this section, it is proposed a conceptual framework for the Fairness Embedded Adaptive Recommender System that aims to balance between personalization and fairness for long-term customer engagement. Firstly, the objectives of the recommendation system are introduced with possible solutions. Then, based on this, a novel architecture for this type of systems is proposed. Next, it is described how fairness dimension can be introduced into the RL engine and how the personalization-fairness trade-off can be solved.

A. Fairness Embedded Adaptive Recommender System

In the present paper, the objective is to create a conceptual design of a recommender system that holds a series of requirements:

- System is focusing on customer relationship development.
- System is incorporating an adaptivity functionality.
- System is optimizing for long term customer engagement.
- System is ensuring consumer and provider fairness.
- System is using a small action and state space in the RL engine.
- System is solving the personalization and fairness trade-off.

B. System Overview

Once converted, the relationship with a new customer must be developed for it to become profitable. In simple terms, this means understanding and covering client's needs.

The objective of the application is to extract consumer preferences and use this knowledge to find the most appropriate products and / or content that will be recommended through communication and interaction with the customer. Same time, the recommended content should bring to user the maximum utility and give him equal opportunities compared to people from other social groups.

C. System Components

The framework used follows next steps: 1) Database creation 2) Preprocessing 3) Recommendations generation 4) Recommendation’s combination 5) Reinforcement Learning Engine 6) User Recommendation 7) User Feedback incorporation.

1) Database creation: The application starts by setting up data sources (Fig. 2, 1). The information considered mandatory in a recommender system application is a) consumer data, b) their past interactions, c) data on provider’s items characteristics, d) items’ reviews and e) metadata about the current browsing session.

2) Preprocessing: The next step is to prepare the tables in the form in which they will be used in different recommendation components. This means that a series of tables having different structures will be created: a) The Customer-Item Matrix Table contains information about the items purchased or consumed by a consumer during the analysis period b) The Transaction-Item matrix is typically stored in transactional format where a transaction contains several rows, c) The items characteristics table contains all the tangible and intangible characteristics of an item as well as statistics about how many times it was recommended, clicked, bought from recommendation lists etc.

3) Recommendation Components: Once all the main tables are prepared (Fig. 2, 2), three recommendation components are developed, and their results are merged and combined to create the action space for the RL Engine (Fig. 2, 3-4).

4) User-oriented collaborative filtering recommendation (Fig. 2, 3.1): The method starts from the assumption that similar users have similar preferences [88] and reflects the real situation when recommendations from friends are more effective. Model tries to explain the Customer-Item Matrix Table using a set of latent factors. Latent structures are automatically deduced from the matrix, as long as the number of factors is specified [88]. Once the factors are discovered, the model associates the belonging of an item to a factor and the user's inclination towards the same factor. For each customer, the model will recommend products that have values close to "1" in the reconstructed matrix and have not been purchased in the past.

5) Content-based recommendation (Fig. 2, 3.2): In the case of new users or items, because of no prior history, the recommendation strategy is using the Items characteristics table. For example, one can use all available information about the tangible and intangible properties of the item, including embeddings extracted from unstructured data.

6) Complementary item recommendation (Fig. 2, 3.3): In this step, rule sets are extracted from transactions using association algorithms. This is extremely useful as it emphasizes the context of using/consuming the initial item. This as well brings completeness to customer's need by saying "if you want to use this, do not forget about that".

Following these recommendation strategies, the outcomes are combined between them in order to create the action space for the RL model (Fig. 2, 4). As it can be seen, there is also the "Random Recommendation" component (Fig. 2, 3.4) and "No Recommendation "(Fig. 2, 3.5) that will bring exploration and novelty into the recommendation landscape as well as will keep the system from harming stakeholders through unfair decisions.

7) Reinforcement Learning Engine: The next task of the system is to choose the most appropriate action for a particular use. In other words, the question that needs to be answered is: "For this user, what is the best action to take? Recommendation of a product that corresponds to the latent structures of the user? An item similar to what user consumed/liked before or a complementary item?". The answer can be as well that the best action is no action.

The solution that could combine user information, past behavior and interaction in order to choose the best action is to use a RL engine (Fig. 2, 5). RL problem is defined as a MDP system.

The set of environmental states is represented by the finite clusters over the vector space extracted from the characteristics of the environment:

- The socio-demographic characteristics of the user.
- User past behavior and interaction with focus on indicators like diversification, appetite for novelty, previously liked items.
- The details of the period in which the browsing and recommendation is made. This can include time of the day or year, browsing device, browsing session time etc.

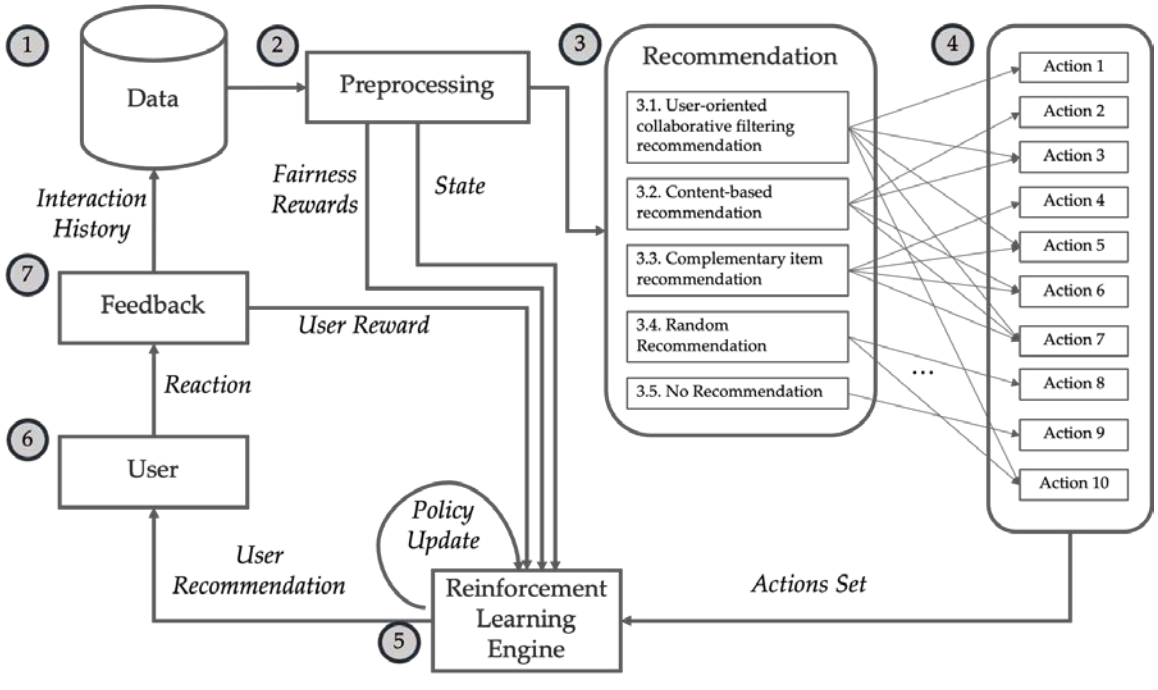


Fig. 2. Design of a Fairness Embedded Adaptive Recommender System.

The set of actions represents all possible actions that the recommendation system controls. As defined before, the set of actions is represented by the individual recommendations or combinations of them (Fig. 2, 4).

The advantages of formulizing the actions like this are: 1) the low complexity given by action space: instead of having all the candidate items, 2) keep the personalization as a key focus, 3) include novelty for consumer and fairness for provider (random product action and no recommendation action).

The reward function is conditioned by the user's response to the recommendation received (Fig. 2, 6-7), but also by fairness rewards of the system.

The reward value at time t+1, Rt+1, after the agent takes At in St is compounded out of three terms:

- Reward coming from user response
- Reward coming from fairness value towards consumer
- Reward coming from fairness value towards provider

Next, the individual terms are defined, because they are the key for embedding fairness and consumer relationship in the system.

User response rewards are linked to the action that user is taking after seeing the recommended item. There can be distinguished the following actions that a user is taking in response, each with a different associated reward. The associated reward will be decreasing as per user feedback reaching a negative value if user ends the communication with company:

- User clicks on the recommended product, and buys/consumes it.
- User clicks on the recommended product, but does not buy/consume it.
- User is adding the item into wish list/buy latter/favorites lists.
- User marks the recommendation as inappropriate.
- User closes the recommendation/searching session without taking any other action.

Consumer fairness reward is linked to a particular item recommended and is known in advance; thus, it is deterministic. Consumer fairness reward is calculated for each item based on Disparate Impact formula as defined in [37]. This particular formula was used because it is easy to adapt and integrate as a reward into the system. Also, the metric encodes the demographical parity idea in a RL workable manner. But the main reason why this particular metric was chosen out of all presented previously is its unique advantage of being deterministic, thus known before taking any action. Considering the act of recommending a particular item being a treatment, this means that the ratio between the likelihood of a positive outcome (presentation of an item) for different groups should ideally be close to 1:

$$Disparate\ Impact(item)$$

$$= \frac{P(recommend\ item=1\mid group=protected\ group)}{P(recommend\ item=1\mid group=privileged\ group)} \quad (1)$$

$$= \frac{P(recommend\ item=1\mid group=protected\ group)}{P(recommend\ item=1\mid group=privileged\ group)} \quad (1)$$

The Disparate Impact for an item is updated every time an item is being recommended using equation 1, and although it is item linked, is actually expressing the consumer fairness. By integrating this metric, it is ensured that individuals with different backgrounds are treated the same, thus they receive same type of items and content recommendation. A good example of a situation that was not fair towards consumer is the study that shows that female users of Google had a lower chance of being recommended and presented hiring ads for high-paying executive jobs [67].

In order to integrate the Disparate Impact into the RL engine, first is needed to adapt it to express the undesirable practice of having either positive or negative bias. Although most of the metrics are focusing on supporting the unprivileged group, this can lead to a turn of the situation, such that the focus here is on not having any type of bias and treating individuals from different groups equally. The equation of calculating the consumer reward is given by equation 2.

$$R_{t+1}consumer\ fairness\ (item) = -|Disparate\ Impact(item) - 1| \quad (2)$$

For the provider fairness, it is desired to ensure that exposure is a function of relevance. For this purpose, difference in acceptance rates is being used (DAR) [81] and is calculated as the pairwise difference of the ratio of true positives divided by the predicted positives for each class. The latest is also called Precision in binary classification evaluation [47] and reflects the fraction of relevant cases. The reason why it was decided to go with this particular metric, is the fact that it is the only metric out of the presented ones that can be used in other recommendation settings besides ranking.

By incorporating DAR into the RL engine there will be ensured that items are presented proportionally to how relevant they are.

$$Relevance\ (item) = P(y = 1\ |\hat{y} = 1) \quad (3)$$

In order to actually use DAR as rewards, for each item, the relevance as defined in equation 3 is calculated. The updated relevance is simulating the situation on recommending an item, while it will not be clicked on. Secondly, DAR (equation 4) is calculated using the newly updated relevance for the item, while maintaining the relevancies for other items ceteris paribus. There will be a DAR value associated with each of the items that can potentially be recommended. Finally, this value is taken with a negative sign (equation 5) in order to count for provider fairness. If there are big differences between items relevance, they will sum up and bring to a high negative reward, thus the agent will try to choose an item that leads to a smaller DAR. The overall reward function is given by the equation 6.

$$Difference\ in\ acceptance\ rates\ (item) = \frac{\sum_{i=0}^{I}\sum_{j=0,j\neq i}^{I} dist\left(Relevance(item_i), Relevance(item_j)\right)}{I(I-1)} \quad (4)$$

$$R_{t+1}provider\ fairness\ (item) = -|Difference\ in\ acceptance\ rates(item)| \quad (5)$$

$$(S_t, A_t) = P(R_{t+1}user\ response|\ S_t, A_t = item) + R_{t+1}provider\ fairness(item) + R_{t+1}consumer\ fairness(item)$$

$$R: S\ x\ A \rightarrow\ R \quad (6)$$

For solving this problem formulation and extracting the optimal policy, one can use Temporal Difference Methods [89] as they are appropriate for continuous tasks having discrete state and action spaces or Deep Q-Learning Networks [90] if one wants to include the reward values into the state.

IV. Results and Discussion

In the current paper, it was presented the design of a Fairness Embedded Adaptive Recommender System (FEARS) which allows to create a fruitful relationship with the client by optimizing for long-term goals while making sure to keep both the consumer and provider fairness.

The gaps in the current conceptual practice were presented in the related work chapter. The desired functionalities of the system were stated in the proposed approach section and in the Table I it is shown how requirements were translated into implementation solutions along with potential issues and limitations that should take the form of further work.

The system was designed to use holistic user information including his socio-demographics and past buying behavior patterns. This was integrated into the system by encoding it into the state of the RL MDP model. As per author knowledge, this is a novel addition to the RecSys using RL approach. Another important detail is the inclusion of the recommendation context expressed through browsing time metadata (ex. time of the year, hour of the day). The way states are used in the RL engine, by clustering the initial vectors, allows overcoming the most common RL limitation of non-efficiency and non-convergence.

Another way to ensure optimal policy convergence is to take control over action state space. This was stated as a clear problem in the literature reviewed and by using an elegant approach of defining actions as recommendation strategies and combination of those, the action space is downside from the number of all items to a maximum of 11 actions.

Although the individual recommendation components are traditional and straightforward, together they are covering the whole scene of consumer interest: similar, complementary, high interest or random products. The way these strategies are combined, namely through an RL engine, brings both adaptivity and ensures reaching long term objectives into the system. A detail that emphasizes the customer relationship health and importance is the practice of using negative rewards into RL component. This means that system will try to optimize for users to be recommended products that they are likely to buy but also play quite safe and not causing information fatigue that can lead to termination of relationship. Another addition is the fair view both towards consumer, but also with respect to items providers. Stated as a clear problem in the reviewed literature, the conflict between personalization, consumer fairness and exploration is solved by introducing not only consumer satisfaction rewards, but also fairness specific rewards.

In this paper, it was presented a conceptual framework that can be adapted to a large range of use cases, from e-commerce companies to both news, article and media items recommendation. Another set of application area may consist of those where the decisions and recommendations are linked to life-changing, high stake situations like hiring, job recommendation or financial lending.

The approach tries to overcome limitations of both individual traditional recommendation systems as well as RL usage in the RecSys by having an integrated view over consumer, a focus on the long-term engagement and a strong enforcing of a sustainable and fair recommendation practice.

TABLE I. System Requirements and Implementation Solutions

System Requirements	*Functionality Implementation*	*Potential issues/ limitations with the solution*
System is focusing on customer relationship development	Usage of an RL engine that has the reward function linked to customer relationship goals.	The reward function is not reflecting accurately the desired objective.
System is incorporating an adaptivity functionality	Usage of a RL engine that takes complex states in account when recommending an item.	High complexity and search space that comes with all the additional information.
System is optimizing for long term customer engagement	RL engine is using negative rewards where appropriate in order to decrease information fatigue and optimize for long term objectives.	Negative rewards are too small in comparison with positive rewards affecting their efficiency
System is ensuring consumer and provider fairness	RL engine reward function is containing fairness metrics and outputs a higher reward in case of fair recommendations.	N/A
System is using a small action and state space in the RL engine	Action space is represented by individual or combinations of recommendation strategies. State space is discretized by using clustering techniques.	Oversimplification of the action and state spaces that could lead to pattern loss
System is solving the personalization and fairness trade-off	Reward function contains both recommendation relevance metrics and fairness metrics.	Inappropriate balance between the two objectives

Future research should involve implementing the approach and use it in real-world situations for evaluating the degree in which it reaches its multisided objectives. Also, different streams of work linked to potential issues presented in Table I should be carried.

V. Conclusions

Major contributions of this paper are presented as follows:

- A reinforcement learning based framework FEARS for better recommendations that focus on both revenues and relationship with the customer was introduced.
- The framework has a holistic view over customer and recommendation landscape ensuring a highly personalized, relevant and positive user interaction.
- Two relevant, adapted fairness metrics are defined and a way to compute them is presented.
- The relevant fairness metrics are embedded into the system as corresponding rewards.
- A RL problem definition was given that overcomes the common RL in RecSys issue of non-efficiency by using a limited, but relevant action space and discretized and clustered state space.

Overall, the system has all the necessary levers to overcome limitations of individual components, solve the personalization-fairness conflict, ensure long-term customer engagement and avoid the typical RL issue.

Same time, the framework should be tested in real-world situations or simulated data and appropriate design changes should be made. This is a conceptual starting point for developing FEARS.

Acknowledgment

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Sammut, C., & Webb, G. I. (2017). Encyclopedia of machine learning and data mining. Springer.

[2] Mladenic, D. (1999). Text-learning and related intelligent agents: a survey. IEEE intelligent systems and their applications, 14(4), 44-54.

[3] Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In Proceedings of the 1st ACM conference on Electronic commerce (pp. 158-166).

[4] Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. Intelligent Data Analysis, 21(6), 1487-1524.

[5] Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., Zhang, Y. (2018). Deep reinforcement learning based recommendation with explicit user-item interactions modeling. arXiv preprint arXiv:1810.12027.

[6] Choi, S., Ha, H., Hwang, U., Kim, C., Ha, J. W., & Yoon, S. (2018). Reinforcement learning based recommender system using biclustering technique. arXiv preprint arXiv:1801.05532.

[7] Maqbool, S. D., Ahamed, T. I., & Malik, N. H. (2011, December). Analysis of adaptability of Reinforcement Learning approach. In 2011 IEEE 14th International Multitopic Conference (pp. 45-49). IEEE.

[8] Mabu, S., Tjahjadi, A., & Hirasawa, K. (2012). Adaptability analysis of genetic network programming with reinforcement learning in dynamically changing environments. Expert Systems with Applications, 39(16), 12349-12357.

[9] Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. Nature Machine Intelligence, 1(3), 133-143.

[10] Taghipour, N., & Kardan, A. (2008, March). A hybrid web recommender system based on q-learning. In Proceedings of the 2008 ACM symposium on Applied computing (pp. 1164-1168).

[11] Zhao, X., Zhang, L., Xia, L., Ding, Z., Yin, D., & Tang, J. (2017). Deep reinforcement learning for list-wise recommendations. arXiv preprint arXiv:1801.00209.

[12] Zhao, X., Zhang, L., Ding, Z., Xia, L., Tang, J., & Yin, D. (2018, July). Recommendations with negative feedback via pairwise deep reinforcement learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1040-1048).

[13] Zhao, X., Gu, C., Zhang, H., Liu, X., Yang, X., & Tang, J. (2019). Deep Reinforcement Learning for Online Advertising in Recommender Systems. arXiv preprint arXiv:1909.03602.

[14] Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., & Yin, D. (2019, July). Reinforcement learning to optimize long-term user engagement in recommender systems. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2810-2818).

[15] Xin, X., Karatzoglou, A., Arapakis, I., & Jose, J. M. (2020, July). Self-Supervised Reinforcement Learning for Recommender Systems. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 931-940).

[16] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010, April). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web (pp. 661-670).

[17] Munemasa, I., Tomomatsu, Y., Hayashi, K., & Takagi, T. (2018, March). Deep reinforcement learning for recommender systems. In 2018 international conference on information and communications technology (icoiact) (pp. 226-233). IEEE.

[18] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

[19] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413.

[20] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

[21] Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159). PMLR.

[22] Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.

[23] Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 49-58).

[24] Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13.

[25] Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019, January). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 99-106).

[26] Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002.

[27] Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. In Recent Trends in Learning From Data (pp. 155-196). Springer, Cham.

[28] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD

international conference on knowledge discovery and data mining (pp. 259-268).

[29] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016, December). The case for process fairness in learning: Feature selection for fair decision making. In NIPS Symposium on Machine Learning and the Law (Vol. 1, p. 2).

[30] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Roth, A. (2017). A convex framework for fair regression. arXiv preprint arXiv:1706.02409.

[31] Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. Advances in Neural Information Processing Systems 30. Pre-proceedings, 30.

[32] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

[33] Creager, E., Madras, D., Jacobsen, J. H., Weis, M., Swersky, K., Pitassi, T., & Zemel, R. (2019, May). Flexibly fair representation learning by disentanglement. In International Conference on Machine Learning (pp. 1436-1445). PMLR.

[34] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.

[35] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining (pp. 797-806).

[36] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.

[37] Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. arXiv preprint arXiv:2010.04053.

[38] Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In Recommender systems handbook (pp. 1-34). Springer, Boston, MA.

[39] Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. Recommender systems handbook, 77-118.

[40] Cheng, W., Shen, Y., Zhu, Y., & Huang, L. (2018). Explaining Latent Factor Models for Recommendation with Influence Functions. arXiv preprint arXiv:1811.08120.

[41] Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009.

[42] Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. Procedia Computer Science, 49, 136-146.

[43] Wu, L. (2020). Advances in Collaborative Filtering and Ranking. arXiv preprint arXiv:2002.12312.

[44] Wang, X., Jin, H., Zhang, A., He, X., Xu, T., & Chua, T. S. (2020, July). Disentangled Graph Collaborative Filtering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1001-1010).

[45] Bonet, E. R., Nguyen, D. M., & Deligiannis, N. (2020). Temporal Collaborative Filtering with Graph Convolutional Neural Networks. arXiv preprint arXiv:2010.06425.

[46] Li, X., Zhang, M., Wu, S., Liu, Z., Wang, L., & Yu, P. S. (2021). Dynamic graph collaborative filtering. arXiv preprint arXiv:2101.02844.

[47] Maimon O., Rokach L. (2010), „Data Mining and Knowledge Discovery Handbook”, Springer.

[48] Grad-Gyenge, L., Kiss, A., & Filzmoser, P. (2017, July). Graph embedding based recommendation techniques on the knowledge graph. In Adjunct publication of the 25th conference on user modeling, adaptation and personalization (pp. 354-359).

[49] Chen, T., Hong, L., Shi, Y., & Sun, Y. (2017). Joint text embedding for personalized content-based recommendation. arXiv preprint arXiv:1706.01084.

[50] Shi, C., Hu, B., Zhao, W. X., & Philip, S. Y. (2018). Heterogeneous information network embedding for recommendation. IEEE Transactions on Knowledge and Data Engineering, 31(2), 357-370.

[51] Grbovic, M., & Cheng, H. (2018, July). Real-time personalization using embeddings for search ranking at airbnb. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 311-320).

[52] Zhao, Z., Zhang, X., Zhou, H., Li, C., Gong, M., & Wang, Y. (2020). HetNERec: Heterogeneous network embedding based recommendation. Knowledge-Based Systems, 204, 106218.

[53] Palumbo, E., Monti, D., Rizzo, G., Troncy, R., & Baralis, E. (2020). entity2rec: Property-specific knowledge graph embeddings for item recommendation. Expert Systems with Applications, 151, 113235.

[54] Talwar, K. S., Oraganti, A., Mahajan, N., & Narsale, P. (2015). Recommendation System using Apriori Algorithm. Int. J. Sci. Res. Dev, 3(01), 183-185.

[55] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. în Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

[56] Wang, C., & Zheng, X. (2020). Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. Evolutionary Intelligence, 13(1), 39-49.

[57] Yao, F., Li, A., & Wang, Q. (2020, November). Bi-Apriori-Based Association Discovery via Alarm Logs. In International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy (pp. 621-632). Springer, Cham.

[58] Kantardzic, M. (2019). Data mining: concepts, models, methods, and algorithms, 3rd Edition, John Wiley & Sons.

[59] Rebala, G., Ravi, A., & Churiwala, S. (2019). Reinforcement Learning Algorithms. în An Introduction to Machine Learning (pp. 213-241). Springer, Cham.

[60] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.

[61] Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., He, X. (2020). State representation modeling for deep reinforcement learning based recommendation. Knowledge-Based Systems, 205, 106170.

[62] Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., ... & Boutilier, C. (2019). Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. arXiv preprint arXiv:1905.12767.

[63] Zhou, S., Dai, X., Chen, H., Zhang, W., Ren, K., Tang, R., Yu, Y. (2020, July). Interactive recommender system via knowledge graph-enhanced reinforcement learning. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 179-188).

[64] Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019, January). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the conference on fairness, accountability, and transparency (pp. 339-348).

[65] Chiappa, S., & Isaac, W. S. (2018, August). A causal bayesian networks viewpoint on fairness. In IFIP International Summer School on Privacy and Identity Management (pp. 3-20). Springer, Cham.

[66] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

[67] Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Proceedings on privacy enhancing technologies, 2015(1), 92-112.

[68] Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. Management Science, 65(7), 2966-2981.

[69] Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.

[70] Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. Scientific reports, 8(1), 1-7.

[71] Lerman, K., & Hogg, T. (2014). Leveraging position bias to improve peer recommendation. PloS one, 9(6), e98914.

[72] Burke, R. (2017). Multisided fairness for recommendation. arXiv preprint arXiv:1707.00093.

[73] Abdollahpouri, H., Burke, R., & Mobasher, B. (2017, July). Recommender systems as multistakeholder environments. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (pp. 347-348).

[74] Burke, R., Sonboli, N., & Ordonez-Gauger, A. (2018, January). Balanced neighborhoods for multi-sided fairness in recommendation. In Conference on Fairness, Accountability and Transparency (pp. 202-214). PMLR.

[75] Lee, E. L., Lou, J. K., Chen, W. M., Chen, Y. C., Lin, S. D., Chiang, Y. S., & Chen, K. T. (2014, August). Fairness-aware loan recommendation for microfinance services. In Proceedings of the 2014 international conference on social computing (pp. 1-4).

[76] Celis, L. E., Straszak, D., & Vishnoi, N. K. (2017). Ranking with fairness constraints. arXiv preprint arXiv:1704.06840.

[77] Castillo, C. (2019, January). Fairness and transparency in ranking. In ACM SIGIR Forum (Vol. 52, No. 2, pp. 64-71). New York, NY, USA: ACM.

[78] Singh, A., & Joachims, T. (2019). Policy learning for fairness in ranking. arXiv preprint arXiv:1902.04056.

[79] Narasimhan, H., Cotter, A., Gupta, M., & Wang, S. (2020, April). Pairwise fairness for ranking and regression. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 5248-5255).

[80] Yang, K., & Stoyanovich, J. (2017, June). Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (pp. 1-6).

[81] Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. arXiv preprint arXiv:1505.05723.

[82] Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. arXiv preprint arXiv:1705.08804.

[83] Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 643-650). IEEE.

[84] Kiswanto, D., Nurjanah, D., & Rismala, R. (2018, October). Fairness aware regularization on a learning-to-rank recommender system for controlling popularity Bias in E-commerce domain. In 2018 International Conference on Information Technology Systems and Innovation (ICITSI) (pp. 16-21). IEEE.

[85] Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Zhang, Y. (2021). Towards Long-term Fairness in Recommendation. arXiv preprint arXiv:2101.03584.

[86] Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Fair algorithms for infinite and contextual bandits. arXiv preprint arXiv:1610.09559.

[87] Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., & Parkes, D. C. (2017). Calibrated fairness in bandits. arXiv preprint arXiv:1707.01875.

[88] Kotu, V., Deshpande, B. (2018). Data science: concepts and practice. Morgan Kaufmann.

[89] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[90] Riedmiller, M. (2005, October). Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In European Conference on Machine Learning (pp. 317-328). Springer, Berlin, Heidelberg.

Online Learning Acceptance Model during Covid-19: An Integrated Conceptual Model

Qasem Kharma[1], Kholoud Nairoukh[2], AbdelRahman Hussein[3]
Mosleh Abualhaj[4], Qusai Shambour[5]
Faculty of Information Technology
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Because of Covid-19, many countries shutdown schools in order to prevent spreading the virus in their communities. Therefore, schools have opted to use online learning technologies that support distance learning for students. As consequences, Ministry of Higher Education and Scientific Research encourages higher education institutes to adopt blended learning in their programs. However, different students react in different ways to online learning. Some students were able to make productive use of online learning strategies more than others. A conceptual model based on 15 variables was constructed based on UTAUT2, TAM, and other models to investigate and to study the factors that affect students' acceptance of online learning. 29 hypotheses were investigated to study the relationships among the variables that affect online learning acceptance and online learning community building in Al-Ahliyya Amman University. The collected responses were analyzed using a structural equation modeling (SEM) approach. SPSS and AMOS were used to analyze the data.**

Keywords*—*Online learning; technology acceptance; learning assistance; learning community building assistance

I. Introduction

Since Covid-19 (Coronavirus) started spreading, many countries decided to shutdown schools and move to online learning solutions [1, 2]. Moving to distance learning because of the virus became necessary in order to keep providing education to students. Different distance learning paradigms were implemented to support distance learning such as live-streaming lectures [3], the use of Massive Open Online Courses (MOOC) [4], and the use of worksheets and animations [5]. Additionally, some educational gamification [6, 7] websites and computer applications were adapted to support online learning. Unfortunately, not all students accept the online learning in place of face-to-face classes.

Based on the online learning feedback during Covid-19, Ministry of Higher Education and Scientific Research developed a plan to adopt online and blended learning in the higher education institutes. As consequences a three-year action plan (2021-2023) was developed to manage adapting synchronized and asynchronized online learning to be blended in higher education. The plan included training faculty members from private and public university. However, moving from traditional lecture-led classes to online learning would cause to move from lecture-centric to learning-centric paradigms. Therefore, students need also to be training and to adapt to the new learning paradigms.

Although the members of Generation Z, who are familiar with various technologies accessing the internet, are the majority of students, it is very important to study the factors that affect their intention to use their technical knowledge in digital learning [8]. It is very important to know the factors that are affecting users' adoption of technologies in order to achieve the objectives of the proposed technology. Many models, such as the Unified Theory of Acceptance and Use of Technology (UTAUT) [9, 10], were proposed to explain users' behavior in utilizing technologies. However, the factors that affect the use of a technology are vary based on the environment [11].

In this research, a model is constructed using 15 variables to investigate its effects on the online learning in Al-Ahliyya Amman University. The model was developed based on UTAUT2 [10], the conceptual TAM (Technology Acceptance Model) [12-15], and a conceptual model proposed by Vululleh [16]. Based on the defined variables, the research model was constructed, and 29 hypotheses were investigated and analyzed using a structural equation modeling (SEM) approach in SPSS and AMOS.

II. Literature Review

One of the widely applied models to study the factors that affect using a technology is UTAUT [9]. It was developed by investigating eight different models. The eight models are the model of PC utilization (MPCU) [17], social cognitive theory (SCT) [18], Technology Acceptance Model (TAM) [19], innovation diffusion theory (IDT) [20], the combined theory of planned behavior/technology acceptance model (C–TPB–TAM) [21], and the motivational model (MM) [22]. As a result, UTAUT was formulated based on four factors: performance expectancy (PE), effort expectancy (EE), social influence (SI), and facilitating conditions (FC). Also, UTAUT considered four moderators: gender, age, experience, and voluntariness of use. In 2012, UTAUT was extended to UTAUT2 [10] by adding hedonic motivation, price value, and habit factors to the model. Later, the Values-Enhanced Technology Adoption (VETA) model [23] extended UTAUT2 by adding new variables to UTAUT2 variables. Although UTAUT and UTAUT2 are widely used to study the acceptance of technology, they need to be adapted to the context of eLearning.

Islam [13] investigates the effects of learning management systems (LMS) on achieving learning outcomes by developing a model based on TAM. The model suggests that LMS may assist students in learning and building a collaborative network in hybrid courses which combine both face-to-face education and online learning. Learning assistance and perceived community building assistance were added to TAM. The model was formulated using five factors: Perceived Usefulness, Perceived Ease of Use, e-Learning Use, Perceived Learning Assistance, and Perceived Community Building Assistance. The model did not consider the design of online courses nor the factors which can affect the learner community building.

Since online learning is usually web browser-based or based in mobile applications, a model developed by Liu, et al. [12] added new variables that may affect the acceptance of online learning to TAM. The new variables are related to the technology design and the user experience with technology. The model used the following constructs: Perceived Usefulness, Perceived Ease of Use, and Intension to Use (from TAM). The following constructs were developed: Online Course Design, User-Interface Design, Previous Online Learning Experience, and Perceived Interaction. The model integrates the effects of online course design and interface design with TAM. However, the model did not consider the effects of the variables such as hedonic motivation, price value, community building and habit factors.

Vululleh [16] studied students' acceptance of e-learning technology in Liberia. The research model was based on TAM, which proposes that users' Behavioral Intention to adopt a technology is controlled by their beliefs of Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). Additionally, the model extended TAM by adding two factors, Social Influence (SI) and Quality of Life (QL). Therefore, the conceptual model studied the effects of four factors: PU, PEOU, SI, and QL on the Behavior Intention (BI) to use e-learning. Like Liu, et al. [12] and Islam [13], Vululleh [16] proposed a model based on TAM. Factors such as design, interface and community building were not considered.

Some models studied the differences of acceptance in different environment. For instance, Lee [14] examines the difference between Korean and American students' perceptions of online education. The model was based on TAM using the two factors Perceived Usefulness (PU) and Perceived Ease of Use (PEOU), and adding the two factors Perception of Online Learning Acceptance and Student Satisfaction (OLAS) and Online Education Support Service Quality (PSQ). The model found that the (PSQ) directly influenced behavioral intention toward online learning acceptance and student satisfaction. However, the impact of service quality on the behavioral intension toward e-learning in different cultures was limited. The model investigates the direct and indirect effects of PSQ on PLAS; however, more variables such as design, interface and community building need to be considered.

The model proposed by Ratna and Mehra [15] examines the behavioral intention of the university students in India for using e-learning. The model was based on TAM using the factors of Gender, Student Major, and Monthly Family Income. It examines the applicability of TAM among university students in India and found that Perceived Ease of Use (PEOU) has a significant effect on Perceived Usefulness (PU), Attitude (ATT), Behavioral Intention (BI) and Actual Use (AU). Also, PU has a significant effect on ATT, BI, and AU. Moreover, ATT has a significant effect on Intention to Use (INT). Finally, this model found that BI has a significant effect on AU. Like the aforementioned models, this model is also based on TAM and limited to study the effects of four variables only.

III. Model Development

The proposed model is based on 15 variables that were utilized in several studies [9, 10, 12-16, 19, 24, 25] to investigate the factors that affect the adoption of technologies:

1) Performance Expectancy is how technology benefits users to perform certain activities [10], and Perceived Learning Assistance (PLA) is how an online learning component can help the individual's learning [13]. Therefore, in our research model, these two factors are combined into one factor.

2) Perceived Community Building Assistance (PCBA) refers how the online learning helps individuals to have social interactions with others [13]. Liu, et al. [12] defines Perceived Interaction as interpersonal interaction and human-system interaction. The interpersonal interactions include interactions with peers and instructors. In our conceptual model, we opt to include Perceived Community Building Assistance and interpersonal interaction into one factor.

3) Online Course Design (OCD) refers to the types and quality of the materials that can be included in the online course [12]. Since the online course is web browser-based or based in a mobile application, it can include different types of material such as images, animation, and video. Therefore, the quality of these materials can have great impact on users.

4) When developing software, User-Interface Design (UID) plays crucial role in the success of the software. In online course, User-Interface Design refers the organization and arrangement of the online course content and visual design [12].

5) Effort Expectancy is how ease the use of the technology affects users [10]. Effort Expectancy is also called by other researchers such as Liu, et al. [12] as Perceived Ease of Use. Hence, only one of them (PEU) is included in our conceptual model to find the effects of ease of use of e-learning technology on the intention to use it.

6) Social Influence (SI) is how other people, such as family and friends, influence users' decisions to adopt the technology [10].

7) Facilitating Conditions are the resources and the support that are available to the user to use the technology [9]. These include both physical and environmental factors [26] that help students to learn using online resources. In our model, the Perceived Service Quality (PSQ) variable represents all the factors that help the students to use the online resources.

8) Hedonic Motivation (HM) is defined as the fun, enjoyment, or pleasure that the users of a technology can have [10]. This variable was not in UTAUT [9], but it was added to UTAUT2 [10] and was considered the most important added variable to UTAUT2 [27].

9) Quality of Life (QoL) is how online learning processes can affect the students' faith [26]. This includes saving time and costs when using or downloading the online learning materials.

*10)*Price Value (PV) is the monetary costs assigned with the use of the technology [10].

*11)*Habit and Experience (Habit) is the user's ability to use the technology without the need for training because of learning and previous experience [10].

*12)*Behavioral Intention (BI) measures students' acceptance of online courses in the present and in the future [26].

*13)*Technological Experience (EXP) represents the skills that students need to use online courses. Martinho, et al. [24] classify Technological Experience into two external variables: Base Technological Experience and Advanced Technological Experience. However, since basic computer skills are taught in schools starting from the elementary grades, we opt in the proposed model not to distinguish between Base Technological Experience and Advanced Technological Experience.

*14)*Perceived Usefulness (PU) is the degree to which students believe that online learning will enhance their job performance [13]. This variable was introduced in TAM [19].

*15)*Previous Online Learning Experience (OLE) [12] including using technology, internet, and online learning resources can affect learners' intentions to use online learning.

Based on the aforementioned variables, the following hypotheses were proposed.

H1: OLE will positively affect students' BI to use online courses.

H2: OLE will positively affect students' PU of online courses.

H3: OLE will positively affect students' PEU of online courses.

H4: OCD will positively affect students' PU of online courses.

H5: OCD will positively affect students' PEU of online courses.

H6: OCD will positively affect students' PCBA of online courses.

H7: PSQ will positively affect students' PU of online courses.

H8: PSQ will positively affect students' BI to use online courses.

H9: PSQ will positively affect students' PEU of online courses.

H10: UID will positively affect students' PEU of online courses.

H11: UID will positively affect students' PCBA of online courses.

H12: PU of online courses will positively affect students' BI to use online courses.

H13: PU of online courses will positively affect students' PLA to use online courses.

H14: PU of online courses will positively affect students' PCBA of online courses.

H15: EXP will positively affect students' PU of online courses.

H16: EXP will positively affect students' PEU of online courses.

H17: PEU of online courses will positively affect students' PU of online courses.

H18: PEU of online courses will positively affect students' BI to use online courses.

H19: PEU of online courses will positively affect students' PLA to use online courses.

H20: PEU of online courses will positively affect students' PCBA of online courses.

H21: SI will positively affect students' PU of online courses.

H22: SI will positively affect students' BI to use online courses.

H23: BI to use online courses will positively affect students' PLA to use online courses.

H24: BI to use online courses will positively affect students' PCBA of online courses.

H25: QoL will positively affect students' BI to use online courses.

H26: HM will positively affect students' BI to use online courses.

H27: PV will positively affect students' BI to use online courses.

H28: Habit will positively affect students' BI to use online courses.

H29: PCBA will positively affect students' PLA to use online courses.

IV. Methodology

Quantitative methods were utilized, in the form of non-structured survey questionnaires with closed questions using a 5-point Likert-type scale where 1 is highly disagree and 5 is highly agree. The questionnaire was constructed of 15 variables and 69 items in total. The items included in the questionnaire were collected from prior studies. Prior to the survey, a pilot study was conducted to assure the reliability and validity of the questionnaire. The entire instrument demonstrated acceptable reliability; Cronbach's alpha was 0.962. The questionnaire was randomly distributed to Al-Ahliyya Amman University students. The final accepted questionnaires for analysis were 462 out of 517 from the submitted responses. The data were analyzed using SPSS version 25 and AMOS version 23.

A. Data Analysis

There are nine colleges in Al-Ahliyya Amman University: Faculty of Information Technology, Faculty of Pharmacy, Faculty of Architecture & Design, Faculty of Allied Medical Sciences, Faculty of Arts & Sciences, Faculty of Engineering, Business School, Faculty of Law, and Faculty of Nursing. The responses, which were used in the analysis according to the colleges, was as followings: Faculty of Information Technology (36%), Business School (22.7%), Faculty of Pharmacy (19%), and the other colleges (22.3%) (see Table 1).

Faculty of Engineering and Faculty of Pharmacy are five-year colleges, while the other colleges are four-year colleges. The responses that were used spanned the following student levels: First Year (24%), Second Year (25.8%), Third Year (21.6%), Fourth Year (17.7%), and Fifth Year (10.8%) (see Table 2).

Table 3 illustrates the previous student experience with online courses. 60 students (13%) had never ever studied with any online course, while 402 (87%) students had studied using some online courses. 205 students studied one or two online courses. 165 students studied 3 to 5 online courses. 32 students studied 6 or more online courses.

TABLE I. STUDENT DISTRIBUTION OVER COLLEGES

College	Frequency	Percent
Faculty of Information Technology	167	36.1
Faculty of Pharmacy	88	19.0
Faculty of Architecture & Design	21	4.5
Faculty of Allied Medical Sciences	13	2.8
Faculty of Arts & Sciences	24	5.2
Faculty of Engineering	19	4.1
Business School	105	22.7
Faculty of Law	18	3.9
Faculty of Nursing	7	1.5
Total	462	100.0

TABLE II. STUDENT LEVELS

Level	Frequency	Percent
First Year	111	24.0
Second Year	119	25.8
Third Year	100	21.6
Fourth Year	82	17.7
Fifth Year	50	10.8
Total	462	100.0

TABLE III. STUDENT EXPERIENCE WITH ONLINE COURSES

Online Course	Frequency	Percent
Never studied online course	60	13.0
studied 1-2 online courses	205	44.4
studied 3-5 online courses	165	35.7
Studied more than 6 online courses	32	6.9
Total	462	100.0

B. Hypotheses Testing

The model is consisted of 15 variables that were measrured in a questionnare of 69 items. The variable averages (ranging from 3.5 to 3.9) and standard deviations (ranging from 1 to 1.3) are listed in Table 4. The variables Cronbach's Alpha if Item Deleted was ranged from 0.958 and 0.962. Since total variable Cronbach's Alpha is 0.962, all variables were included in the model and none of them need to be dropped. Many indexes were measured to test the model fitness (see Table 5). Since all the minimum values of the indexes were achieved, the model fits enough to measure the hypotheses. All factor loadings were above 0.6; therefore, all items were used in the analysis.

TABLE IV. THE MODEL VARIABLES

Variable	Mean	Std. Deviation	Cronbach's Alpha if Item Deleted
PU	3.4654	1.04352	.959
PEU	3.4802	1.04912	.959
PLA	3.5415	1.13455	.959
PSQ	3.5584	1.12182	.958
EXP	3.9446	1.24697	.960
PCBA	3.5361	1.10217	.958
OCD	3.5000	1.16716	.958
UID	3.5260	1.24651	.958
OLE	3.6075	1.25583	.959
SI	3.5216	1.14455	.958
QoL	3.5628	1.18976	.958
BIs	3.6129	1.17900	.958
PV	3.4726	1.19596	.962
Habit	3.5476	1.07761	.959
HM	3.5440	1.29309	.960

TABLE V. MODEL FIT RESULTS

Index	Value	Recommended values
Degrees of freedom	2203	> 0 [17]
Discrepancy Chi-Squared	4031	> .05 [17]
Relative Chi-Squared	1.83	< 5.0 [18]
Comparative Fit Index	0.915	≥ 0.9 [18, 19]
Tucker Lewis Index	0.909	≥ 0.9 [17, 20]
Incremental Fit Index	0.915	≥ 0.9 [17, 20]
Root Mean Square of Error Approximation	0.042	≤ 0.08 [17]

The 29 hypotheses were tested using AMOS. 15 hypothses were supported while 14 hypothese were rejected. The result is illustrated in Table 6. H17, which states that Perceived Ease of Use of online courses will positively affect students' Perceived Usefulness of online courses, has high impacts; if Perceived Ease of Use goes up by 1 unit, Perceived Usefulness of online courses goes up by 1.204. In addition, if Perceived Ease of Use goes up by 1 unit, PLA goes up by 0.839 unit. Therefore, Perceived Ease of Use is considered a very important factor in online learning.

On the other hand, the effects of the factors in H7, H14, and H18 have negative effects. In H18, Perceived Ease of Use will negatively affect Behavioral Intentions. When Perceived Ease of Use goes up one-unit, Behavioral Intentions goes down by 0.822. Thus, when the students feel that online courses are too easy to use, they will use them less. In short, H7, H14, and H18 will be stated as following:

H7: Perceived Service Quality negatively affects students' Perceived Usefulness of online courses.

H14: Perceived Usefulness of online courses negatively affects students' Community Building Assistance of online courses.

H18: Perceived Ease of Use of online courses negatively affects students' Behavior Intention to use online.

TABLE VI. HYPOTHESES TESTING

Hypotheses				Estimated	S.E.	C.R.	P	Comments
H1	BI	←	OLE	-0.076	0.089	-0.858	0.391	Not Supported
H2	PU	←	OLE	0.041	0.074	0.559	0.576	Not Supported
H3	PEU	←	OLE	0.003	0.055	0.054	0.957	Not Supported
H4	PU	←	OCD	0.25	0.096	2.595	0.009	Supported*
H5	PEU	←	OCD	0.332	0.109	3.057	0.002	Supported*
H6	PCBA	←	OCD	0.684	0.169	4.034	***	Supported***
H7	PU	←	PSQ	-0.368	0.143	-2.582	0.01	Supported**
H8	BI	←	PSQ	0.362	0.193	1.87	0.062	Not Supported
H9	PEU	←	PSQ	0.63	0.09	6.969	***	Supported***
H10	PEU	←	UID	-0.12	0.086	-1.404	0.16	Not Supported
H11	PCBA	←	UID	-0.162	0.109	-1.484	0.138	Not Supported
H12	BI	←	PU	0.565	0.243	2.33	0.02	Supported*
H13	PLA	←	PU	0.034	0.137	0.248	0.804	Not Supported
H14	PCBA	←	PU	-0.286	0.141	-2.025	0.043	Supported*
H15	PU	←	EXP	-0.013	0.045	-0.286	0.775	Not Supported
H16	PEU	←	EXP	-0.012	0.043	-0.278	0.781	Not Supported
H17	PU	←	PEU	1.204	0.18	6.674	***	Supported***
H18	BI	←	PEU	-0.822	0.401	-2.052	0.04	Supported*
H19	PLA	←	PEU	0.839	0.183	4.586	***	Supported***
H20	PCBA	←	PEU	0.653	0.165	3.963	***	Supported***
H21	PU	←	SI	-0.142	0.095	-1.497	0.134	Not Supported
H22	BI	←	SI	0.377	0.122	3.095	0.002	Supported**
H23	PLA	←	BI	0.222	0.052	4.273	***	Supported***
H24	PCBA	←	BI	0.059	0.066	0.892	0.373	Not Supported
H25	BI	←	QoL	0.388	0.085	4.589	***	Supported***
H26	BI	←	HM	0.023	0.048	0.48	0.631	Not Supported
H27	BI	←	PV	-0.073	0.046	-1.605	0.109	Not Supported
H28	BI	←	Habit	0.315	0.076	4.165	***	Supported***
H29	PLA	←	PCBA	-0.041	0.082	-0.496	0.62	Not Supported

*** Significant p ≤ 0.001, ** Significant p ≤ 0.01, * Significant p ≤ 0.05

V. Limitation and Conclusion

The study was conducted in a private university, Al-Ahliyya Amman University. Online learning has become a necessity after the spread of Covid-19. Therefore, this study can be extended to include responses from students in different educational institutions, including public and private institutions and K-12 institutions. All responses were collected from students. The study can be extended to collect responses from instructors as well. Finally, the moderator affects, such as age, gender, and experience, were not considered in this study because there studies such as Liao, et al. [21] and Kharma [22] found that there were no significant effects of these moderators in online learning models. However, when extending the study to include different categories of educational institutions, there might be impacts of these moderators on the relations in the model.

The study proposed a theoretical model to investigate the effects of 15 variables on online learning acceptance. Based on these 15 variables, 29 hypotheses were tested. Only 14 hypotheses were supported. The final model is shown in Fig 1. Four variables: Perceived Usefulness, Social Influence, Quality of Life, and Habit had positive effects on Behavioral Intention to use the online course. The highest estimated beta value of these was of Perceived Usefulness. In other words, in order to increase students' behavioral intention to use an online course, they should know the importance of the online courses in learning. The role of academic advising can help in increasing students' awareness of the usefulness of online learning and how it can affect the learning process by accessing the online material at anytime and anywhere.

Moreover, the study found that the course design and ease of use of the course affect the online collaborative learning. Therefore, in order to enhance peer-learning and group-learning, the online courses should include in the design simple and easy features that allow students to communicate with their classmates. Additionally, ease of use positively affects students' perceptions of online learning. Hence, the ease of use of online learning components and environment is crucial to increasing students' intent to learn from online learning components and to cooperate with their classmates.

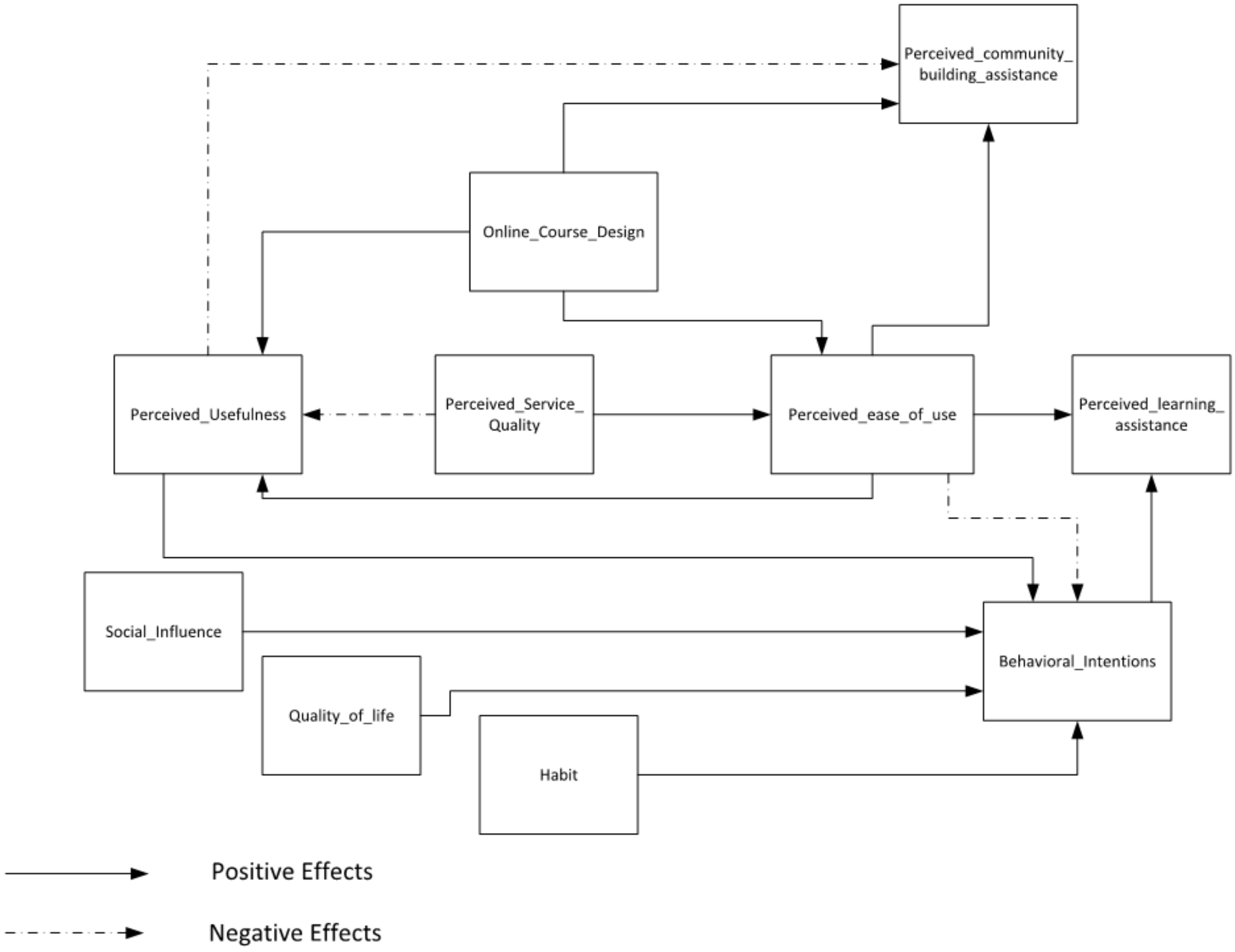


Fig. 1. Final Model.

Three relations had negative effects in the model. The first one is that the more assistance in using the online course, and more physical requirements to access the online course, the less usefulness the online course has. The second is that the more assistance in using the online course and more physical requirements to access the online course, the less collaboration there is with others. Finally, if the students feel that the online course is too easy, they won't be as interested in the online course. Therefore, the physical requirements and the technical assistance should be minimized. For instance, online courses should use good compression to minimize the usage of internet bandwidth and hardware requirements. Additionally, online courses should not be too easy. The students like some challenges in the learning process.

Finally, the effects of Hedonic Motivation were not supported. This finding agrees with Mehta, et al. [23]. College-level students do not use online learning for fun. However, further investigation is needed to study the effects of Hedonic Motivation in different educational system levels, such as K-12 schools.

References

[1] W. Van Lancker and Z. Parolin, "COVID-19, school closures, and child poverty: a social crisis in the making," The Lancet Public Health, vol. 5, no. 5, pp. e243-e244, 2020.

[2] M. Mossa-Basha et al., "Policies and guidelines for COVID-19 preparedness: experiences from the University of Washington," Radiology, p. 201326, 2020.

[3] A. Sandhu, A. Fliker, D. Leitao, J. Jones, and A. Gooi, "Adding Live-Streaming to Recorded Lectures in a Non-Distributed Pre-Clerkship Medical Education Model," in ITCH, 2017, pp. 292-297.

[4] M. Brown, "Why invest in MOOCs? Strategic institutional drivers," The 2018 OpenupEd trends report on MOOCs, pp. 6-9, 2018.

[5] S. Mahtari, M. Wati, S. Hartini, M. Misbah, and D. Dewantara, "The effectiveness of the student worksheet with PhET simulation used scaffolding question prompt," in Journal of Physics: Conference Series, 2020, vol. 1422, no. 1: IOP Publishing.

[6] D. Dicheva, C. Dichev, G. Agre, and G. Angelova, "Gamification in education: A systematic mapping study," Journal of Educational Technology & Society, vol. 18, no. 3, 2015.

[7] J. Majuri, J. Koivisto, and J. Hamari, "Gamification of education and learning: A review of empirical literature," in Proceedings of the 2nd international GamiFIN conference, GamiFIN 2018, 2018: CEUR-WS.

[8] S. F. Persada, B. A. Miraja, and R. Nadlifatin, "Understanding the Generation Z Behavior on D-Learning: A Unified Theory of Acceptance and Use of Technology (UTAUT) Approach," International Journal of Emerging Technologies in Learning (iJET), Generation Z; Digital Learning; UTAUT; Confirmatory Factor Analysis; Behavior Intention vol. 14, no. 05, p. 20, 2019-03-14 2019, doi: 10.3991/ijet.v14i05.9993.

[9] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27, no. 3, pp. 425-478, 2003, doi: 10.2307/30036540.

[10] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," MIS Quarterly, vol. 36, no. 1, pp. 157-178, 2012, doi: 10.2307/41410412.

[11] D. Al-Fraihat, M. Joy, R. e. Masa'deh, and J. Sinclair, "Evaluating E-learning systems success: An empirical study," Computers in Human Behavior, vol. 102, pp. 67-86, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.chb.2019.08.004.

[12] I.-F. Liu, M. C. Chen, Y. S. Sun, D. Wible, and C.-H. Kuo, "Extending the TAM model to explore the factors that affect Intention to Use an Online Learning Community," Computers & education, vol. 54, no. 2, pp. 600-610, 2010.

[13] A. N. Islam, "Students'e-Learning System Usage Outcomes: A Study with a Learning Management System," in Conference on e-Business, e-Services and e-Society, 2011: Springer, pp. 255-268.

[14] J.-W. Lee, "Online support service quality, online learning acceptance, and student satisfaction," The Internet and Higher Education, vol. 13, no. 4, pp. 277-283, 2010.

[15] P. Ratna and S. Mehra, "Exploring the acceptance for e–learning using technology acceptance model among university students in India," International Journal of Process Management and Benchmarking, vol. 5, no. 2, pp. 194-210, 2015.

[16] P. Vululleh, "Determinants of students'e-learning acceptance in developing countries: An approach based on Structural Equation Modeling (SEM)," International Journal of Education and Development using ICT, vol. 14, no. 1, 2018.

[17] J. F. Hair, R. E. Anderson, B. J. Babin, and W. C. Black, "Multivariate data analysis: A global perspective (Vol. 7)," ed: Upper Saddle River, NJ: Pearson, 2010.

[18] R. P. Bagozzi and Y. Yi, "On the evaluation of structural equation models," Journal of the academy of marketing science, vol. 16, no. 1, pp. 74-94, 1988.

[19] B. M. Byrne, Structural equation modeling with AMOS: Basic concepts, applications, and programming. Routledge, 2016.

[20] R. Ho, Handbook of univariate and multivariate data analysis and interpretation with SPSS. CRC press, 2006.

[21] Q. Liao, J. P. Shim, and X. Luo, "Student acceptance of web-based learning environment: An empirical investigation of an undergraduate IS course," AMCIS 2004 Proceedings, p. 377, 2004.

[22] Q. Kharma, "Investigating Students' Acceptance of Online Courses at Al-Ahliyya Amman University," International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019, doi: 10.14569/IJACSA.2019.0100729.

[23] A. Mehta, N. P. Morris, B. Swinnerton, and M. Homer, "The Influence of Values on E-learning Adoption," Computers & Education, vol. 141, p. 103617, 2019/11/01/ 2019, doi: https://doi.org/10.1016/j.compedu.2019.103617.

[24] D. S. Martinho, E. M. Santos, M. I. Miguel, and D. S. Cordeiro, "Factors that Influence the Adoption of Postgraduate Online Courses," International Journal of Emerging Technologies in Learning (iJET), acceptance model; higher education; intention-to-use; online learning. vol. 13, no. 12, p. 123, 2018-12-20 2018, doi: 10.3991/ijet.v13i12.8864.

[25] D. Aijaz Ahmed Arain, Z. Hussain, W. Rizvi, and M. Vighio, "Extending UTAUT2 toward acceptance of mobile learning in the context of higher education," Universal Access in the Information Society, vol. 18, no. 3, pp. 659–673, 07/22 2019, doi: 10.1007/s10209-019-00685-8.

[26] N. Samsudeen Sabraz, "University students' intention to use e-learning systems," Interactive Technology and Smart Education, vol. 16, no. 3, pp. 219-238, 2019, doi: 10.1108/ITSE-11-2018-0092.

[27] K. Tamilmani, N. P. Rana, N. Prakasam, and Y. K. Dwivedi, "The battle of Brain vs. Heart: A literature review and meta-analysis of "hedonic motivation" use in UTAUT2," International Journal of Information Management, vol. 46, pp. 222-235, 2019/06/01/ 2019, doi: https://doi.org/10.1016/j.ijinfomgt.2019.01.008.

Data Analytics in Investment Banks

Basma Iraqi[1]*, Lamia Benhiba[2], Mohammed Abdou Janati Idrissi[3]
Mohammed V University in Rabat
ENSIAS, Rabat
Morocco

***Abstract*—Capital Markets are one of the most important pillars of worldwide economy. They gather skilled finance and IT professionals as well as economists in order to take the best investment decisions and choose the most suitable funding solutions every time. Data analytics projects in Capital Markets can definitely be very beneficial as all optimizations and innovations would have a financial impact, but can also be very challenging as the field itself has always incorporated a research component, thus finding out what could really be of an added value might be a tricky task. Based on a comprehensive literature review, this paper aims to structure the thoughts around data analytics in investment banks, and puts forward a classification of relevant data analytics use cases. Lastly, it also discusses how transforming to a data-driven enterprise is the real change investment banks should aim to achieve, and discusses some of the challenges that they might encounter when engaging in this transformation process.**

Keywords—Capital markets; data analytics; data analytics use cases; data-driven transformation; investment banks

I. Introduction

Capital markets are one of the most important and complex fields in worldwide economy that involves multiple stakeholders, among which investment banks are the main actor. These later play the role of the mediator between enterprises looking for funding and others willing to invest. They advise their clients on the best approaches to collect funds and manage their investments, thanks to a very thorough understanding of the client profile, but also the ups and downs of the markets and the economy.

Faced with troubled economic times, investment banks, like all other financial actors, have been pushed, more than ever, to rethink their products and services and improve their quality by analyzing further all the available data on clients and markets [1]. The interest towards data analytics comes therefore naturally as this is where magic happens with the data. Through the different types of analytics possible, the algorithms and techniques developed, there is opportunity for investment banks to improve their relationship with their clients and enhance their overall revenues.

Nevertheless, starting data analytics projects can be a challenging step as these are usually initiatives where the output is not known from the beginning and the return on investment not determined. Thus, it is important to do thorough research beforehand to identify the types of analytics that could be of interest to the organization, as well as find out similar success stories in order to define the scope and get inspired before starting the experimentation.

Research on data analytics in investment banks have continuously focused on specific use cases' implementations, but the topic has never been addressed from a broader perspective. For the financial services sector, no comprehensive guides are available to help investment banks in approaching data transformation from the beginning, by identifying the problems they can resolve and the types of analytics they can implement in order to choose use cases from.

This paper aims therefore to provide a classification of relevant data analytics use cases for investment banks, and discusses the importance of engaging in a data-driven transformation and some of the challenges that could be faced in this process. The rest of this paper will be organized as follows: Section II presents capital markets and investment banks, Section III outlines the motivations behind adopting data analytics in investment banks and Section IV proposes a classification of the most relevant use cases that can be deployed in this sector based on a review of the literature. In the last section, a discussion is initiated about some of the challenges that investment banks may encounter when initiating the data-driven transformation, in order to provide better guidelines for investment banks in this process.

II. Introduction to Capital Markets and Investment Banks

Capital markets are markets where demand and supply for funding based on different instruments are organized. These instruments are grouped in different asset classes and quoted in different markets [2].

The main objectives of Capital Markets are to secure funding for national economy, structure liquidity and savings, enable firm restructuring, mergers and acquisitions, cover multiple kinds of risks and provide and represent the main references for all asset values [3].

Out of all the activities performed in capital markets, portfolio management is one of the most important ones. It focuses on decision making with regards to investments for individuals or institutions, taking into account the investment policy, objectives, asset allocation, risk and performance [4]. The science behind it enables measuring performance, and consequently making consistent improvements based on continuous learning, controlling risk and ensuring consistency and discipline throughout the process and the portfolio lifecycle.

*Corresponding Author

In order for the different stakeholders in the Capital Markets to decide on the strategies they will work with and manage their portfolios appropriately, they also rely greatly on different economic indicators which represent some of the most important signals and raw data influencing the capital markets' activities. That is because it is these same indicators that explain the motivation behind the need for funding, investing or restructuring among others. These indicators could be of the following three types: (1) Supply indicators, (2) Demand indicators and (3) Inflation, Money and financial indicators [5].

Modeling and predicting how changes in the values of economic indicators impact the investment strategies adopted and how the change in the prices of the different instruments due to economic variations affect the value of the investment portfolio and its risk vulnerability is how Capital Markets function on a daily basis.

In capital markets, investment banks play a very important role. In fact, they are the main intermediary between large corporations or governments that need funding, and the market investors. Based on the financial challenges of each, investment banks advise their clients on the best approaches to raise capital taking into account the current investment climate as well as the client's needs [6]. This also applies to mergers and acquisitions' scenarios where businesses are often looking to restructure by purchasing their competitors.

As such, the main departments of an investment bank are mainly [6]:

- Trading and Sales: the department in charge of executing deals on behalf of the clients, but also the one responsible of trading the bank's own money on the different financial instruments.
- Asset Management: referring to portfolio management for large institutions, insurance companies or pension funds.
- Wealth Management: the function of advising wealthy families and individuals in ways to fulfill their financial long-term objectives and needs by investing in capital markets.
- Securities Services: the department responsible for all services related to assets held or issued by the clients. These services include custody, clearing, settlement and fund administration among others.
- Research: the function of collecting all kinds of information about the market and the different corporations in order to generate recommendation reports on some stock buy or sell.
- In addition to the risk management, internal controls, accounting and IT functions among others.

Given their very diverse missions, investment banks' departments often have opposite views on market trends, look for different opportunities, offer various services to the same clients, and thus collect distinct data inputs as well as analyze and process them differently.

III. Data Analytics in Investment Banks

Analytics is a buzz word. It is used almost everywhere and in very different contexts. There are even many research papers which only aimed at finding the right definition to analytics. One of the best descriptions one could find in the literature regarding analytics is the one provided by CETIS (the Center for Educational Technology and Interoperability Standards) which defines "Analytics [as] the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data" [7].

Actionable insights shed the light on the importance of action in analytics. Analytics' results should go beyond reporting or descriptions to opportunities and conclusions that prompts users for action. That's actually the main difference between Analytics and Analysis, which are the most confused concepts in this regard. Indeed, while "Analytics aim to project the future based on past performance, analysis rather presents a historical overview of past data" which is equivalent to what descriptive or diagnostic analytics could render [8]. An example of an actionable insight is the outcome that goes with A/B testing initiatives of E-commerce websites (such as Amazon or eBay) in which they test variations of their websites in samples of users in order to analyze their sentiments and reactions to the changes made.

The literature distinguishes four types of analytics [9]:

- Descriptive analytics: which answers questions such as: what is happening? What is it about? What does it mean? What are its attributes? About the data and content being analyzed.
- Diagnostic analytics: are answers to the following questions: why did it happen? What are the triggering events? What are the preliminary conditions for each context?
- Predictive analytics: are more focused about the future, thus answering questions like: what is likely to happen? What happened in similar situations throughout the past? What could be the results of such actions or decisions?
- Prescriptive analytics: come after predictive analytics to discuss and suggest potential alternatives fix problems or capitalize on opportunities. Questions answered in prescriptive analytics include: what should I do about it? What can I do prevent it? How can I best exploit it?

The capital markets, and investment banks more particularly, have been one of the first and most important users of data analytics. That is not a surprising finding as the field does not manufacture any physical product and data is really its main asset [10]. In fact, statistics in 2016 have even showed that the field is the number 1 consumer of data services with a share of 13% of the global data analytics use cases implemented [11], and that is because data is used and generated tremendously every second of the day in this field, and that analytics could definitely help making sense of data that could never be processed exhaustively.

In addition, the interest in Data Analytics within investment banks is also the result of a tight conjuncture which led to weak returns compared to earlier times. Indeed, according to a study by Accenture, the financial sector and capital markets more specifically, has witnessed since a few years a stagnation of revenues given the fall of margins and the rising complexity of regulations. Also, the Fixed Income, Currencies and Commodities business, which has historically filled the largest share of revenues, witnessed an important share shrinking for the same reasons. Last, new bank competitors have arisen, making the pressure on old banks even tighter [1].

Data has therefore made it at the center of the banks' analytics projects because it promises [12]:

- A better risk management: investment banking is the sector where resources are heavily invested in risk because the consequences of a bad risk assessment could be devastating. The 2008 financial crisis and its impact on the global economy is the best example to illustrate the high importance of this business line. This applies to all types of risks, whether it be related to [1]: (1) Fraud detection: where big data could be leveraged in order to identify patterns of fraudulent transactions or atypical operations and alert the appropriate personnel in order to investigate further. (2) Liquidity: where analytics could be used to keep track of the short- and long-term liquidity every time, assess the impact of transactions on liquidity in real time and run simulations and stress tests regularly in order to make sure that the necessary funds for banks to function correctly are always available. (3) Credit risk: where internal data about clients and counterparties is enriched with external data from the web, social media and the news in order to get an exhaustive feel of their financial situation and ensure that the hazards are well managed and known if any.
- A pleased and Loyal Customer: this is where sentiment analysis comes to play in order to better understand the needs of the customers and address them correctly. Thanks to all the data available in the web, including the news, social media, research reports and corporate websites, it is now possible to anticipate what the client might or might not appreciate, and direct them to the most suitable products (cross and up selling) at the right time. This not only enables an improved customer loyalty, but also makes attracting prospects a more successful process.
- A secure ecosystem: Analytics enables a very large and thorough monitoring where patterns of incidents and problems are identified through machine learning algorithms, thus making their handling and resolution a much straightforward process.

Since it is usually very expensive to start collecting all relevant data for the bank before making any use of it, all organizations, whether they are in financial services or in another industry, usually proceed in a different way. Companies start by defining a business objective linked to a strategic goal they want to reach thanks to analytics, and then reverse the process to identify the data to capture in order to achieve the goal. This specific application of data analytics is what a use case refers to, and it is the most common approach to address analytics as it offers a scope and a project perspective to the data application and enables the association of KPIs and governance that are very important to the monitoring of data analytics initiatives.

Despite these promising benefits, there are unfortunately fewer Analytics initiatives deployed in the capital markets sector, and especially in investment banks. That is because the challenge of finding interesting use cases is real, and the technical complexity that comes along the execution of these use cases is also more pronounced and different from the challenges that could be faced in other environments. So, what business problems could Data Analytics fix or optimize in investment banks? And what problems could be faced when implementing them?

In fact, based on the literature review of the use of analytics in investment banking, it came out that a brainstorming around business use cases of data analytics in investment banking is actually missing. As much as there is a lot of material when it comes to fields like retail banking, insurance or even industry, there is significantly less work in the investment banking sector given its business complexity and specific characteristics.

The focus of the remainder of this section is to describe data analytics use cases that can be of interest to investment banks, based on a review of research papers, of data analytics use cases worldwide as well as the authors' professional experience. These papers will be discussed following a proposed classification, which puts each use case in one of the 7 categories described below (from A to G) and indicates whether it is a deployed solution or is still the object of ongoing research. The research papers included in this study are the authors' selection of the most relevant papers in each category, based on the assessment of their added value according to the authors' financial expertise. Last, throughout this section, both internal bank data and external economic data will be proven to be equally important in implementing data analytics in investment banks, and the methods and data processing techniques that can be used to generate insight from this data will also be discussed.

A. *Prediction of Financial Market Patterns*

The investment banking sector is very tied to market data and economic indicators. It is based on this data that traders and portfolio managers decide each day whether to buy or sell. Therefore, predicting the market shifts and movements is a very valuable information for the actors to anticipate their investment strategies and act accordingly.

Predicting market patterns can be done based on the history of transactions, current and historical market data as well as indicators of the economic sector. The idea behind this use case is to identify if the current scheme is somehow similar to an old one in order to learn from the mistakes made by then and make the most out of it. This use case can be implemented by labeling the dataset based on sets of conditions and feature engineering in order to output the knowledge needed. Then techniques like clustering to group alike data together or game theory to evaluate how each variable in the dataset influences the model positively or negatively can be used [13].

Another method to predict the patterns of financial markets is to build models which, based on input financial indicators, predict values of the output ones. An example of this approach is the article “Big Data Analytics for Financial Market Volatility Forecast based on Support Vector Machine” [14] which zooms on high frequency data or real time data, which is data that is generated on a very high velocity. This type of data has been proved to be very rich in terms of information, but its treatment has always been difficult given the high number of problems that must be fixed beforehand. In this article, the focus was put on volatility as an important indicator of high frequency data on financial markets, and it suggests predicting the short-term volatility via a model of the jump volatility. The technique used for this purpose is Support Vector Machine. It takes eight input indicators in order to model the price and predict the next day’s price, and thus its short-term volatility. It also applies several kernel functions, which happen to have different prediction results on different markets.

On this category of use cases, JPMorgan Chase has been able to deploy a NLP project for equity investing. Indeed, in order to improve the research capabilities of their portfolio managers on equity, JPMorgan launched a Natural Language Processing project specialized on equity investing. The tool is fed with news articles on equity markets and generates insights on the data such as identifying enterprises which have new products to launch or the ones currently in difficulty. According to JPMorgan, the investments based on this tool have outperformed those based on the NASDAQ50 index [15].

B. Detection of Crisis Situations

Detecting crisis situations is technically a very similar use case to the prediction of market patterns. Indeed, it just focuses on a specific label of the data which is the crisis or the most alarming ones. The how-to is therefore the same, and the only difference consists in shedding light on the weak signals only in order to anticipate any future challenges or problems that may occur and set a backup plan in advance. This use case can be particularly of interest to the risk teams who need to anticipate and adjust their measures in case of problems or disasters.

An example of use cases in crisis detection is the paper “Monitoring Banking System Connectedness with Big Data” [16], where the authors came up with a solution to model connectedness between individual firms based on firm-level information in order to ensure better supervision and monitoring of financial stability since connectedness between economic agents has become one of the most addressed topics in the sector. The methodology applied for this purpose is a decomposition of the outcome measure into multiple components, and then a network construction based on indicators of connectedness in the US banking system. The first step is simply based on linear regression of CAPM framework using the French-Fama model, and the second step is about network construction and analysis. This latter is done through many measures in order to provide the best insight possible to the banking system stability state.

C. Client Risk Profiling

In portfolio management, there are two main inputs which define how the investment strategy should be: The first is the performance level wanted and the second is the risk appetite. Determining how much risk the client is willing to take is yet another challenging task that portfolio managers have to realize. Indeed, a portfolio manager could find cases where a client is unable to evaluate their risk appetite or one who underestimates his willingness to take risks or vice versa.

For such problems, there are multiple solutions. Scoring algorithms can be used which, based on questionnaires, determine the weights and risk scores to assign to every question and answer. An example of algorithms that does so is the self-organizing map, which belongs to the artificial neural network family and is based on unsupervised learning for model training. Another solution to the risk profiling use case is to apply classification models to the client data in order to put them on a risk group based on a set of attributes. Here, algorithms like KNN, Decision Trees or SVM can be used. And last, clustering can also be used on the historical transactions authorized by the client in order to actually find out what other clients they are similar to and adjust their risk profile accordingly [17].

One paper that addressed this topic in a different way is “The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics” [18]. This paper comes up with a new approach to estimate credit scores, which is based on mobile call records and social data. In fact, the paper suggests that the use of mobile data just like social media data can have an interesting impact on the clients' credit scores. The data is modeled as a network where nodes represent people in the log. Labels as well as other information are extracted from by using link-based features as well as a personalized page rank. The methodology is then evaluated statistically and economically. In fact, the tests proved that the combination of call network data with conventional data has given better results in credit scoring. The data by itself (without conventional data) has even rendered satisfying outputs. The paper claims that this methodology is especially interesting in developing countries where historical financial data is rare or inexistent. The only limitation to this approach is the availability of data, especially given the tight regulations in place in many developed countries.

On this family of use cases, CitiBank has been able to deploy a Credit-Sentiment Monitor from news media. This tool, which name is CitiVelocity, is based on the news articles published in Thomson Reuters. The objective behind such a tool is to provide Citi staff with insights on the companies that are mentioned positively and negatively in the press. This information would then be used to evaluate indices of Credit Default Swaps in which these companies are included [19].

D. Product Recommendation for Clients and Prospects

Investments banks, at the opposite of retail banks, do not have huge numbers of clients as the business is mainly B2B (Business To Business), thus a thorough understanding of the client needs and profile is required to maintain a good and continuous relationship with the client, especially for the senior bankers and the client relationship management team, who represent the front interface of the bank with its clients. For that, a 360º vision of the client is necessary. This latter should be composed of the client history, which consists of all client transactions and exchange logs with the bank, its risk profile and its strategic orientations. These could be deduced from the sentiment analysis of the client activity on their website or social media accounts or its interaction with the bank's website.

Recommendation systems in such case are based on collaborative filtering algorithms, and the model used could be matrix factorization, neural matrix factorization or singular value decomposition among others. The rationale behind is to give an implicit score to all products for every client, such that the higher the score, the higher are the chances that the client will be satisfied from the recommendation [20].

The paper "an investment portfolio recommendation system for individual e-commerce users" has come with a new recommendation system implementation, where aside of the investor's risk preference, parameters such as gender, education or location are taken into consideration. The system is built based on the collaborative filtering algorithm and the VaR (Value at Risk) method for risk measurement. The particularity of the recommendation system used is based on the rapprochement with the behavior of other e-commerce investors with similar preferences [21].

Among investment banks, UBS has succeeded in providing a free investment consulting service based on recommendation engines. They implemented a tool called SmartWealth based on machine-learning algorithms which is meant to provide prospects with a free advisory service on online investment in order to increase their chances of becoming clients. In order to realize such a service, UBS asks the client to answer a set of questions, and then the algorithm assigns the client a risk segment and invests its money in an adequate fund [22].

Goldman Sachs had also transformed its whole approach of identifying investment opportunities for the clients based on three aspects: (1) Momentum, where techniques of industry sentiment are applied in order to identify correlations between businesses in the same sector; (2) value, which takes into consideration sector information in addition to financial data in order to define the intrinsic value of the firm; and last (3) profit, which concerns e-commerce companies and enables to evaluate them based on the patterns of web traffic [23].

E. Client Satisfaction follow up

Client satisfaction is also an important brick of the bank client relationship, but also a difficult one as it is not straightforward to evaluate the client satisfaction from the bank services. There are three main sources from which client satisfaction could be inferred: the bank portal, where client complaints are processed, emails and phone calls exchanged with the client and their online posts and articles or interaction with the bank's social media accounts.

Determining whether the client feedback is positive or negative is processed using machine learning and sentiment analysis, where after cleaning the data extracted from the identified sources, embedding is performed, then model training is realized through algorithms like decision trees, neural networks or naïve bayes among others. Such insights could be of interest to all the bank departments, but to the senior bankers more especially who can use the lessons learnt from the client satisfaction measures to better approach them and handle their inquiries.

The fanciest way to follow up on client satisfaction today is to have chatbots. The role of the chatbot is to answer client questions automatically or to log and route their complaints to the relevant staff. The paper "AI-Based chatbot service for financial industry" has suggested an approach based on machine learning in order to automate the communication between the bank and the client regarding product sales or customer support, such that the robot agent keeps accumulating knowledge throughout the process [23].

On the same application of chatbots, J.P.Morgan and Morgan Stanley have all implemented chatbots for their production environments. The beginning is not easy, but the result is worth it. Indeed, thanks to conversational agents, Morgan Stanley for instance, has been able to answer many of the Research-related questions of their clients (analyst profile, research coverage, etc) thanks to well-trained agents who answer efficiently all the questions [25].

F. Simulation and Recommendation of Market Trades and Investment Strategies

For their daily activities, traders and portfolio managers are always handling the trade-off between their stock and the changes in market data and economic indicators. Market data being on movement in real time, it is just a matter of seconds before a good buy is not anymore or the opposite. That is why, the main struggle in such a use case is to constantly have at hand data that is enriched by market data, transformed taking into consideration all the market and economic changes and customized by filtering the economic indicators that have an impact on the positions on hand.

For that purpose, applications for real-time computations of risk metrics, limits or price predictions are all interesting. Machine learning models like K-means for clustering are a good fit to select assets while taking portfolio diversification into consideration. Reinforced learning or deep learning can also be used to control the dynamics of the system in an optimized way [26]. Last, recommendation systems can be applied, such that based on historical transactions, returns and economic factors, recommendations of new operations can be made based on techniques like matrix factorization [27]. A paper that has also worked on the topic is "An equity fund recommendation system by combining transfer learning and the utility function of the prospect theory". Actual transaction data is used to test the technique suggested, and the result is to answer the main question “what to buy?” given inputs of the operations and the market [28].

On the side of investment banks, MAN Group has also used artificial intelligence algorithms in order to define the best investment strategies on her AI-Managed funds. Based on millions of data points including transactional data, the tool developed by MAN identifies patterns to improve and adapt to the market changes. According to Bloomberg, this algorithm is behind more than 50% of MAN profit on its biggest funds [29].

G. *Optimization of Internal and Support Processes*

Finally, besides all the benefits that the main business lines can realize, the support functions can also make huge gains out of business analytics. Audit and Internal Control departments can get their controls automated, IT can have their system and technical incidents and problems automatically detected and resolved, and compliance and legal teams can make use of better fraud detection models among others. And for this family of use cases, all kinds of techniques could be used depending on the problem at hand.

One paper that handled a similar use case is “Automatic Detection of Relationships between Banking Operations using Machine Learning” [30], which solved the problem of matching operations in different systems. In fact, because there are different contexts in which the banks make use of the same operation (such as accounting, risk, internal control), banks end up having the same operation in different formats in multiple systems, and it is impossible for humans to figure out which transactions actually point out to the same operation. The contribution of this article is therefore a framework that enables operation matching thanks to machine learning mechanisms. The framework proposed is divided into 4 stages: Stage 1 is pre-processing, which consists of preparing the data from the data lake by selecting candidate operations, identifying possible records, merging and filtering. Stage 2 is machine learning based on ANN algorithms, and stage 3 is post-processing where a summary and combination of all results of the classifiers is performed in order to generate an exhaustive report of all matching operations. In order to test this framework, the authors applied it to 3 different scenarios, and they all ended up with at least 99,58% accuracy.

BNP Paribas Securities Services automatic trade matcher is definitely one of the best illustrations of this use cases’ category. In fact, BNP Paribas built a tool named Smart Chaser in order to match trades between the Clients and Brokers. The tool, based on predictive analysis, has allowed BNP Paribas to identify patterns of mismatches in order to avoid them beforehand, thus reducing manual intervention on the trades [31].

Table I summarizes all the research and enterprise use cases discussed within this paper:

TABLE I. DATA ANALYTICS USE CASES IN INVESTMENT BANKS

Use Case Family	***Use Case***	***Problem***	***Technique Used***	***In production or ongoing research?***
Prediction of financial market patterns	Big data analytics for financial Market volatility forecast based on support vector machine [14]	Analytics on high-frequency data	Support Vector Machine	Ongoing research
	NLP project for equity investing – JP Morgan Chase [15]	Research in equity investing	Natural Language Processing	In production
Detection of crisis situations	Monitoring Banking System Connectedness with Big Data [16]	Monitoring of firm connectedness	Linear Regression and Network Construction	Ongoing research
Client risk profiling	The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics [18]	Credit scores’ estimation	Network modeling	Ongoing research
	CitiVelocity – CitiBank [19]	Credit-sentiment monitor from news media	Sentiment Analysis	In production
Product recommendation for clients and prospects	An investment portfolio recommendation system for individual e-commerce users [20]	Investment portfolio recommendation	Collaborative filtering algorithm	Ongoing research
	SmartWealth – UBS [22]	Product recommendation for prospects	Machine learning algorithm	In production
	Investment opportunities approach – Goldman Sachs [23]	Investment opportunities definition for clients	Sentiment Analysis	In production
Client satisfaction follow-up	AI-Based chatbot service for financial industry [24]	Customer support	Machine Learning	Ongoing research
	Conversational agents – Morgan Stanley [25]	Client research questions	Machine Learning	In production
Simulation and Recommendation of market trades and investment strategies	An equity fund recommendation system by combining transfer learning and the utility function of the prospect theory [28]	Asset recommendation	Transfer Learning and Utility Function	Ongoing research
	AI-Managed funds – MAN Group [29]	Fund management	Artificial Intelligence	In production
Optimization of internal and support processes	Automatic Detection of Relationships between Banking Operations using Machine Learning [30]	Operation matching in different systems	Machine learning (ANN algorithms)	Ongoing research
	Smart Chaser – BNP Paribas Securities Services [31]	Automatic trade matching	Predictive analysis	In production

Based on this summary, the most important dimensions for an investment bank transformation emerge and structure the business and strategic objectives behind the data analytics. In fact, from a high-level perspective, one can isolate the external environment from the internal one within the bank; and within each environment, another detailed partitioning can be made. On one hand, the external environment includes market and economic data in addition to client information, whether it be their financial results or their perception of their relationship with the bank. On the other hand, in the internal environment side, one can distinguish between three main components: the product, the transaction and the risk metrics. All these dimensions combined give a 360° view of the investment bank's processes and enable a transformation impacting all the key factors. It is then up to each bank and depending on its context and pain points that the focus can be put on one or some of the use case families instead of all of them.

This being said, data today has become the backbone of all innovative and successful banking projects and the key element in order to take advantage of any of the use cases previously mentioned is to switch the whole decision-making process to one that is based primarily on figures and facts generated through analytics. As such, information technology projects within the bank should be strategically aligned with the business objectives, and all means and technological capabilities to achieve this cross-disciplinary cooperation should be deployed. In other words, the turning point for investment banks would rather be to transform to data-driven enterprises where data lies at the core of any process and is the most crucial factor in all business choices. This transformation is the only genuine change that can guarantee that all data sources, whether they are structured or unstructured, are being used and combined in order to create rich and insightful input data, and that analytics are achieving the best possible outputs in a timely manner through very fast and efficient data processing. However, this transformation does not come easily as there are various challenges to overcome, some of which will be discussed briefly in the next section.

IV. Challenges of becoming a Data-driven Investment Bank

A. Use Case Prioritization

One of the first challenges investment banks face when starting analytics use cases is how to prioritize them. In fact, as illustrated by the use cases listed in the above section, there are so many inter-dependencies between the use cases because they mostly rely on the same data: a mix of internal deals and operations with market and economic data. Thus, deciding on what use cases should go first is not only a matter of business priorities but also a matter of technical constraints, mainly related to data availability.

B. Data Availability

This first point leads to the second concern of data analytics projects in the field, which is ensuring that all data is accessible. In fact, all the magic of analytics in capital markets lies in the correct combination of internal and external data, which is not always available in internal databases, and is rather found in data providers' platforms, social networks and websites of regulatory entities, ministries, national agencies and clients.

C. Cloud Integration

Besides, given the large amounts of data investment banks may end up processing due to the broad scope of external data needed for analytics, the nature of the data repository / data lake used can also be a difficult decision to make, and one that significantly impacts the long-run and cost of the analytics initiatives. There is an important trade-off to make between the strict regulations on investment banking and data confidentiality they require, and the big data volumes to process, in which a big chunk is mainly public and already available to everyone on the net.

D. Data Architecture

Another problem to consider is the technical and technological architecture of the environment hosting data analytics use cases. In fact, the peculiarity of the external data needed for investment banks is that it is mostly presented in numerous files (pdf, word or excel) of small or medium sizes, since every publication comes in a separate document. The block storage, that is mostly used in data analytics in all other domains, is not necessarily the best or most straightforward option as it is not initially designed for the storage of small files.

E. Event Sourcing

Last, when it comes to internal data, reproducing and replaying historical events is a must for analytics projects. However, the investment banks mostly rely in their IT departments on vendor software which are developed based on architectures that do not enable such options. Therefore, in order to get the most out of internal data too, a lot of work has to be done on the data within the IT systems in order to transform it to fit into an event-driven architecture, which is the only architecture that can answer all the analytics needs.

III. Conclusion and Future Work

This paper describes why, now more than ever, launching analytics initiatives is important in investment banks. It presents also use cases that could be of interest to them and the processing techniques used by these use cases. It is meant to be a first-hand guideline for investment banks that are willing to engage in the process of data-driven transformation, and raises the challenges and problems they might face in this process, especially in terms of use case prioritization, data availability, cloud integration, data architecture and event sourcing.

As a future work, a deeper investigation of investment banks' data driven transformation will be performed by shedding light on its principles and key aspects and proposing solutions to the different challenges raised in order to complete the guideline. On the challenges' part, a special attention will be given to use case prioritization as it should serve all types of investment banks (those who are willing to transform to data-driven enterprises or those who are just looking for data analytics use cases to implement). This will be done by defining the most significant criteria to consider in this process depending on the bank's profile and strategic orientations and building a model to sort the different alternatives accordingly.

REFERENCES

[1] Jelf, Owen. “Capital Markets Technology 2022.” *Accenture*, 2018, www.accenture.com/t20180124t060525z__w__/us-en/_acnmedia/pdf-69/accenture-capital-markets-technology-2022.pdf.

[2] Morvan, Jérémy. “(PDF) Chapitre 1 - Les Marchés Financiers - ResearchGate.” *Researchgate*, www.researchgate.net/publication/318641258_Chapitre_1_-_Les_marches_financiers.

[3] “Exposé Sur Le Marché Des Capitaux.” *Economie-Monetaire-Et-Marche-Des-Capitaux*, economie-monetaire-et-marche-des-capitaux-23.webself.net/file/si267430/download/Exposé sur le marché des capitaux-fi4489230.pdf.

[4] Bilaus, Bogdan. “Portfolio Management for Institutional Investors.” *Cfasociety*, June 2010, www.cfasociety.org/romania/Files/Analiza Financiara si de Investitii, Etica si Standarde Profesionale, Asociatia Brokerilor, June 2010/Bogdan Bilaus - PM for institutional investors.pdf.

[5] Laharach, Youssef. “Marchés Financiers Et Gestion De Portefeuille.” *Laharach-Youssef*, 2014, www.laharach-youssef.com/medias/files/support-en-pdf-partie-i.pdf.

[6] Epstein, Adam, et al. Schroders, Vault Career Intelligence, 2015, www.schroders.com/en/sysglobalassets/digital/careers/2015-european-guide.pdf.

[7] Taylor-Sakyi, K. (n.d.). Big Data: Understanding Big Data (Tech.).

[8] Shende, Vikram. “Project Analytics to Improve Project and Portfolio Decision Making.” *PMI*, 2017, www.pmi.org.in/conference2017/pdfs/papers-pdfs/theme-3-rapidly-changing-world/21-Project-Analytics-to-Improve-Project.pdf.

[9] Zikopoulos, P. (2012). Understanding big data: Analytics for enterprise class hadoop and streaming data.

[10] IBM® Institute for Business Value, and Saïd Business School at the University of Oxford. “Analytics: The Real-World Use of Big Data in Financial Services .” *IBM*, 2013, www.ibm.com/downloads/cas/E4BWZ1PY.

[11] Trelewicz, Jennifer Q. “Big Data and Big Money: The Role of Data in the Financial Sector.” *InfoQ*, InfoQ, 17 Oct. 2017, www.infoq.com/articles/big-data-in-finance/.

[12] “(PDF) Big Data Analytics Enabled Smart Financial Services ...” Researchgate, 2017, www.researchgate.net/publication/321282806_Big_Data_Analytics_Enabled_Smart_Financial_Services_Opportunities_and_Challenges.

[13] Fan, Weiguo, and Michael D Gordon. “The Power of Social Media Analytics.” *ACM*, 1 June 2014, cacm.acm.org/magazines/2014/6/175163-the-power-of-social-media-analytics/fulltext.

[14] Yang, Rongjun, et al. “Big Data Analytics for Financial Market Volatility Forecast Based on Support Vector Machine.” *International Journal of Information Management*, Pergamon, 13 June 2019, www.sciencedirect.com/science/article/pii/S0268401218313604.

[15] “Innovation with Machine Learning.” Jpmorgan.com, www.jpmorgan.com/insights/research/machine-learning.

[16] Hale, Galina, and Jose A. Lopez. “Monitoring Banking System Connectedness with Big Data.” *Journal of Econometrics*, North-Holland, 16 Apr. 2019, www.sciencedirect.com/science/article/pii/S030440761930082X.

[17] Ertek, Gurdal, and Murat Kaya. “Scoring and Predicting Risk Preferences.” *Researchgate*, 2012, www.researchgate.net/publication/233755791_Scoring_and_Predicting_Risk_Preferences.

[18] Óskarsdóttir, María, et al. “The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion Using Mobile Phone Data and Social Network Analytics.” *Applied Soft Computing*, Elsevier, 9 Oct. 2018, www.sciencedirect.com/science/article/pii/S156849461830560X.

[19] “2020 Annual Report.” Citigroup, Citigroup, 2020, www.citigroup.com/citi/investor/quarterly/2021/ar20_en.pdf.

[20] O. Isinkaye, Folasade, et al. Recommendation Systems: Principles, Methods and Evaluation, Researchgate, 2015, www.researchgate.net/publication/283180981_Recommendation_systems_Principles_methods_and_evaluation.

[21] Li, Xiang, and Chunxia Yu. “An Investment Portfolio Recommendation System for Individual E-Commerce Users .” *Semantics Scholar*, International Conference on Production Research , 2017, pdfs.semanticscholar.org/d5b7/53157004b28d892f18 d508e8e3aa715b2da3.pdf.

[22] Wealth and Asset Management 2022: The Path to Digital Leadership, Oracle, 2017, www.oracle.com/assets/wealth-report-summary-full-report-4010572.pdf.

[23] *The Role of Big Data in Investing*, www.gsam.com/content/gsam/us/en/institutions/market-insights/gsam-insights/gsam-perspectives/2016/big-data/gsam-roundtable.html.

[24] Okuda, T. & Shoda, S.. (2018). AI-based chatbot service for financial industry. Fujitsu Scientific and Technical Journal. 54. 4-8.

[25] Global Digital Wealth Management Report, BCG, 2019, media-publications.bcg.com/BCG-Global-Digital-Wealth-Management-Report-2019-2020-ENG.pdf.

[26] Y. Ng, Andrew, and Stuart Russel. “Algorithms for Inverse Reinforcement Learning.” *Ai.stanford*, ai.stanford.edu/~ang/papers/icml00-irl.pdf.

[27] Moylan, and Anderson. Nonlinear Regulator Theory and an Inverse Optimal Control Problem - IEEE Journals & Magazine, ieeexplore.ieee.org/document/1100365.

[28] Zhang, Li, et al. “An Equity Fund Recommendation System by Combing Transfer Learning and the Utility Function of the Prospect Theory.” *The Journal of Finance and Data Science*, Elsevier, 14 Feb. 2018, www.sciencedirect.com/science/article/pii/S240591881730020X.

[29] “AI Pioneers in Investment Management.” CFA Institute, CFA Institute, www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment -Management.ashx.

[30] González-Carrasco, Israel, et al. “Automatic Detection of Relationships between Banking Operations Using Machine Learning.” *Information Sciences*, Elsevier, 12 Feb. 2019, www.sciencedirect.com/science/article/pii/S0020025519301409.

[31] “BNP Paribas Securities Services Brochure.” BNP Paribas Securities Services, 2021, securities.cib.bnpparibas/app/uploads/sites/3/2021/05/ss-doc-broch-glance.pdf.

Early Prediction of Plant Diseases using CNN and GANs

Ahmed Ali Gomaa[1]
Management Information Systems
Madina Academy High Institute for Management and Technology, Giza, Egypt

Yasser M. Abd El-Latif [2]
College of Computing and Information Technology
Arab Academy for Science
Technology and Maritime Transport
Faculty of Science, Ain Shams University
Cairo, Egypt

***Abstract*—Plant diseases enormously affect the agricultural crop production and quality with huge economic losses to the farmers and the country. This in turn increases the market price of crops and food, which increase the purchase burden of customers. Therefore, early identification and diagnosis of plant diseases at every stage of plant life cycle is a very critical approach to protect and increase the crop yield. In this paper using a deep-learning model, we present a classification system based on real-time images for early identification of plant infection prior of onset of severe disease symptoms at different life stages of a tomato plant infected with Tomato Mosaic Virus (TMV). The proposed classification was applied on each stage of the plant separately to obtain the largest data set and manifestation of each disease stage. The plant stages named in relation to disease stage as healthy (uninfected), early infection, and diseased (late infection). Classification was designed using the Convolutional Neural Network (CNN) model and the accuracy rate was 97%. Using Generative Adversarial Networks (GANs) to increase the number of real-time images and then apply CNN on these new images and the accuracy rate was 98%.**

Keywords—Plants diseases; deep learning; early detection; convolutional neural network; generative adversarial networks

I. Introduction

Quality crop production is an essential feature of any country's economic growth. The agricultural sector provides jobs for many people; in addition, it accounts for a large part of the Gross Domestic Product (GDP) in many countries around the world [1]. For example, it is clear that there is a rapid and wide agricultural development and land reclamation in Egypt, with increased application of technological advances. There has been remarkable increase made in agricultural sector, such as the 1.5 million Acres Project, the El-Alamein project, and the lake and Fayoum projects. The use of modern methods and technologies in agriculture, significantly increase crop production and yield, and increase protection of plants from insect pest infestation and disease infections at all stages of planting, harvesting, and post-harvesting till successful marketing [2]. In a fast-growing world population, although there are many improvements in large production and access to food, food security is threatened by a different set of factors such as a decreased fertility of soil and lands, decreased plant pollination efficiency, insect and other arthropod pests, and plant diseases. Plant diseases are classified as follows: bacterial diseases (bacterial speck, bacterial spot, bacterial canker) [3], fungal diseases (early blight, late blight, sectorial leaf spot, and Anthracnose fruit rot) [4], and viral diseases Tomato Mosaic Virus (TMV) [5]. The early detection and accurate diagnosis of plant diseases at every stage of the plant life cycle and the extent of infection until reaching the most infectious stage or appearnce of severe disease symptoms, and easily classifying them is very important, as shown in Fig. 1.

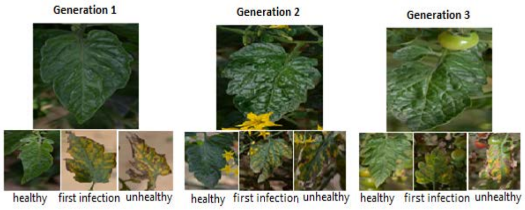


Fig. 1. The Age Stages of the Plant and the Stages of Infection.

Our Deep Learning (DL) model uses leaf images to detect diseases in plants by CNN to extract features from images, such as horizontal edges, vertical edges, Red Green Blue values, etc. A plant disease diagnosis system that uses machine learning techniques can correctly identify diseases plants healthy or unhealthy only [6]. Automatic detection of plant diseases at every age is an important research, analytical, and applied topic because it can help in monitoring large fields of crops in short time with high accuracy. Therefore, disease detection can discover symptoms visually and mechanically in the earliest time they appear on the leaves or other parts of the plant [6], as reported in numerous research publications and reports. CNN is believed to be the best DL neural network for extracting visual features [7]. The CNN-based network can be trained to discover diseases in plants by providing a large number of real-time images. In the case of lacking enough and good quality data or number of images, other techniques such as GANs can be used to generate the needed data for analysis and comparison with real-time data collected from the field. Both healthy and diseased plants and a future training model can be used to predict diseases in plants using plant leaf images [8].

II. Related Work

Recent advances in agricultural technology have led to a demand for a new set of automated, non-destructive methods for detecting plant diseases. Hence, several methods have turned to computer visual and machine learning (ML) techniques to create a rapid method for detecting plant diseases when symptoms appear [9]. Classifying plant diseases can be a very complex task because it depends mainly on published and used classification systems and also by experience of farmers and researchers.

Developing a reliable system that can be applied to many plant classes is a difficult task. To date, most automatic plant disease classification methods have depended on ML algorithms and basic feature engineering. These methods usually focus on specific environments and are suitable for a smaller number of categories, as some small changes in the system can lead to a severe drop in resolution. Recently, CNNs have shown impressive results in many image classification tasks that have allowed researchers to improve the classification of agriculture and plant diseases [10]. CNN is a technology that mixes artificial neural networks (ANNs) and up to date DL strategies [11].

In deep learning, CNN is at the center of spectacular advances. This ANN has been applied to several image recognition tasks for decades and has attracted the eye of the researchers of many countries in recent years; as CNN has shown promising performances in several computer visual and ML tasks [12]. This paper describes the underlying architecture and different applications of the CNN.

In Y. Kawasaki, ET. al. [13], the authors introduce a novel plant disease detection system based on CNN. Using only training images, CNN can automatically acquire the requisite features for classification and achieve high classification performance. A total of 800 cucumber leaf images are used to train CNN using the proposed techniques. Under the 4-fold cross-validation strategy, the proposed CNN-based system (which also extends the training dataset by generating additional images) achieves an average accuracy of 94.9 % in classifying cucumbers into two typical disease classes and a non-diseased class. In this study, the authors proposed a novel plant viral disease detection system using CNN and confirmed its effectiveness. They also asserted that the strategy for training CNN has significantly improved the accuracy of its classification. This work will free system users from paying extra attention to the details of plant shooting conditions.

In Y. Kawasaki, et. al. [13], future the system makes a large contribution in the agricultural field. Data augmentation is an essential part of the training process applied to DL models. The motivation is that a robust training process for DL models depends on large annotated datasets, which are expensive to be acquired, stored and processed. Therefore, a reasonable alternative is to be able to automatically generate new annotated training samples using a process known as data augmentation [14]. A GAN model consists of two important factors: the discriminator (D), and the generator (G). The generator and discriminator have opposite objectives during training. The discriminator is trained toward distinguishing between synthesized and real-time data while the generator is trained to fool the discriminator with synthesized data, as shown in Fig. 2.

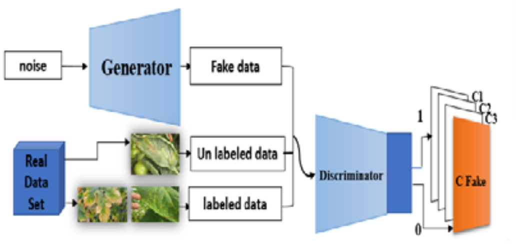


Fig. 2. GANs Architecture.

In D. Farm. [15], the authors propose a synthetic sampling solution is presented at the data level to identify them from small and unbalanced data sets using GANs. The reason for using GANs is the challenges in different fields as they deal with small data sets and volatile amounts of samples per category [16]. As a result, GANs offer an approach that can improve learning regarding data distributions, reduce bias resulting from class imbalance, and change classification. Resolution limits towards more accurate results. The method of [16] was trained on a small dataset of 2789 images of highly perishable tomato plant diseases with a class imbalance in 9 disease categories. Moreover, they evaluated their results in terms of different measures and compared the quality of these results for stratified excellence. GANs are an exciting and quickly changing field, delivering on the deal of generative models in their capacity to generate realistic examples across a range of problem domains. In 2014, conditional GANs was extended to a conditional model if both the generator and discriminator are conditioned on some extra data. They can perform the conditioning by data feeding into both the discriminator and generator as additional input layer [17].

In 2016, the Auxiliary Classifier GAN (AC-GAN) has received much interest due to easy and extensibility to different applications. AC-GAN integrates the conditional information (label) by training the GAN discriminator with an additional classification loss. AC-GAN is able to generate high-quality images and has been extended to different learning problems. However, the difference between the generated samples by AC-GAN going to decrease as the number of classes increases; hence limiting its power on large-scale data [18].

In 2016, the Information Maximizing GAN (Info-GAN) integrated the output of the generator to a component of its input called the hidden codes. Uncovering some successful and unsuccessful configurations for generating images using Info-GAN [19] are shown in Fig. 3.

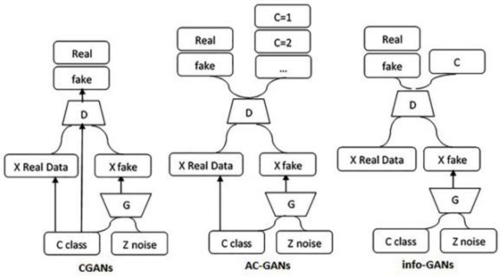


Fig. 3. Types of GANs.

III. Methodology

This methodology works on three steps:

First step: divide the plants into three class's generation (G1–G2–G3) with respect to the age stage.

Second step: each generation contains three Phases (Pi) of plants according to its pre-symptoms (uninfected P1 - early infection P2 – late infection P3).

Third step: Prediction and early detection of diseases to apply it to each generation, as shown in Fig. 4.

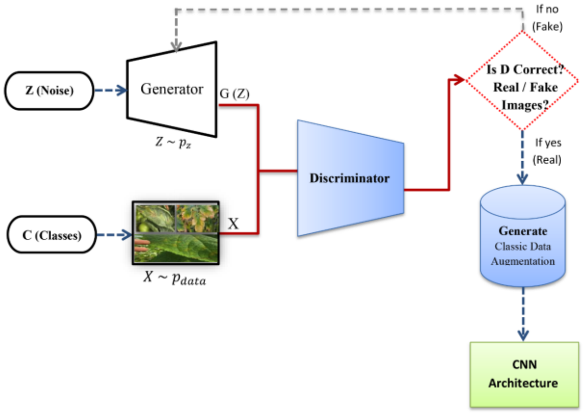


Fig. 4. The Model of Plants Generation/Phases.

Description of the model of plant generation/phases:

- p_{data} Represents all data set.
- Sample a noise set and a real-data set that includes classes (G1, G2, G3), each with size m.
- X represents the real sample belonging to the distribution $X \sim p_{data}$.
- Z denotes a random series belonging to the distributionp_z, which obeys a normal distribution. D and G represent the discriminator and generator respectively.
- The output of the generator is *X* fake (Data).
- The generator to make synthetic samples G (z) extremely approach to the distributionp_{data}.
- To increase the data set apply the discriminator on this data and extract the global polling and hidden layer to get extra real data.
- Global polling layer is inserted in front of the discriminator network's output layer to extract representative features with 512 dimensions.
- The discriminator input is (*X* and G (z)) and they are compared until get the output discriminator.
- The final output of the discriminator is real/fake images.
- When data is fake, the discriminator and generator are trained alternatively. For the training process of the generator, the synthetic samples G (z) is taken into the discriminator and the produced loss value loss G is transmitted back to the generator one more time.
- When the data is real, it is entered into the data augment repository and then after that it is transferred to the CNN structure.
- The final model output is the data augmentation as a shape class to using in CNN mode.

The following results were collected from the use of these images on the CNN, as shown in Table I.

IV. Experiments And Results

In this section, the study concludes the Experimental setup for our synthetic task generates data to detect diseases early:

1) Dataset: These images were collected from agricultural lands and it is a real data set that was used in this work to prove the growth stages of the plant and also increase the data from the original data and determine the stages of plant disease, there is a total of 5400 real images of diseased and healthy plants. These images covered all growth stages of plants and the extent of disease infection.

2) CNNs are proposed to reduce the number of parameters used and adapt the network architecture exactly to visual tasks. CNNs are usually composed of a set of layers that can be grouped by their functionalities; a CNN is typically composed of four types of layers: Convolution Layer, ReLu and sigmoid functions, Pooling, and Fully Connected Layer [20, 21, 22].

TABLE I. Apply CNN on the Groups of Images According to their Generation/Phases

Generation	Phases	Number of images	Loss	Vall_Loss	Accuracy	Vall_Accu	Max_Accu
G1	P1	600	0.0936	2.9398	0.9726	0.4200	0.9726
	P2	600	0.0725	2.8965	0.9701	0.3811	
	P3	600	0.0664	2.8356	0.9689	0.2147	
G2	P1	600	0.0861	2.7649	0.9601	0.2641	0.9754
	P2	600	0.0845	2.8527	0.9754	0.3979	
	P3	600	0.0689	2.8950	0.9742	0.4215	
G3	P1	600	0.0621	2.6691	0.9597	0.2954	0.9748
	P2	600	0.0859	2.9184	0.9726	0.4021	
	P3	600	0.0746	2.8302	0.9748	0.4593	

Table I shows that the highest percentage in healthy cases is in the first age early stage (uninfected) of plant growth. The highest percentage in cases of the first virus infection (early infection) is in the second age stage. The highest percentage in diseased cases (late infection) is in the third age stage.

The plant in the first age stage is in the sterilization stage and healthy hybridization and not exposed to a large pesticides spraying, taking into account the appropriate weather for cultivation. After that, in later stages of growth, the plant is exposed to larger spraying with pesticides and exposed to different climate factors as well as poor workmanship (farming), its stage will be the highest complete unhealthy rate. The data was divided into 70% training and 30% testing, and determining the number of batches required for the model. The accuracy rate and the loss rate were deduced as shown in the Fig. 5.

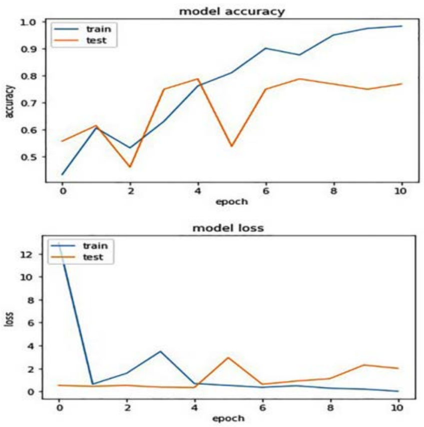


Fig. 5. Model (Accuracy and Loss).

The number of the expected data on the actual data was clear in the following table, and it was found that the second growth stage is the most vulnerable stage to viral infection through the distribution of data by 70% training and 30% testing, as shown in Table II.

['Gen2Phase1', 'Gen2Phase3', 'Gen3Phase3', 'Gen3Phase2', 'Gen1Phase2', 'Gen2Phase2', 'Gen1Phase3', 'Gen3Phase1', 'Gen1Phase1'].

Computed fusion matrix: Heterogeneous data sources can be collectively mined by data fusion. Fusion can focus on a specific target relation and exploit directly associated data together with data on the context or additional constraints [23], as shown in the Fig. 6.

TABLE II. TEST RESULT BY CNN WITH TRAINING AND TESTING

Prediction	G1	G2	G3	All
Actual				
P1	240	210	90	540
P2	120	300	120	540
P3	70	230	240	540
All	430	740	450	1620

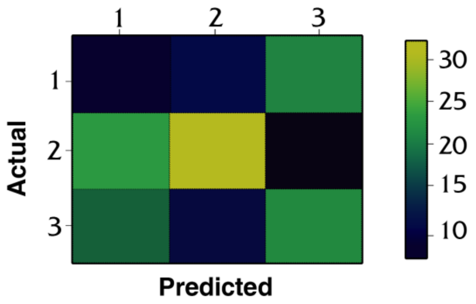


Fig. 6. Computed Fusion Matrix.

3) Generative Adversarial Networks (GANs): After the CNN stage of classification and prediction was completed, the stage of increasing data into by using the (GANs), and during this stage, the experiments were done on the real data and the steps of this stage were as follows:

- Prepare Dataset:
  - import numpy as np
  - import pandas as pd
  - import os
  - print(os.listdir("TomatoDB"))
- Generator and Discriminator for Dataset:
- Image Samples:
  - SAMPLES_TO_SHOW = 8
  - Input: (60, 64, 64, 3), as shown in the Fig. 7.

Fig. 7. Images Sample Complete Description.

- Code implementation steps, as shown in Table III:

TABLE III. RESULT ACCURACY FOR GENERATOR AND DISCRIMINATOR MODEL: "FUNCTIONAL_1"

Layer (type)	Output Shape	Parameter
input_1 (InputLayer)	[(None, 64, 64, 3)]	0
conv1 (Conv2D)	(None, 32, 32, 32)	2432
batch_norm1 (BatchNormalization)	(None, 32, 32, 32)	128
conv1_out (LeakyReLU)	(None, 32, 32, 32)	0
conv2 (Conv2D)	(None, 16, 16, 64)	51264
batch_norm2 (BatchNormalization)	(None, 16, 16, 64)	256
conv2_out (LeakyReLU)	(None, 16, 16, 64)	0
conv3 (Conv2D)	(None, 8, 8, 128)	204928
batch_norm3 (BatchNormalization)	(None, 8, 8, 128)	512
conv3_out (LeakyReLU)	(None, 8, 8, 128)	0

conv4 (Conv2D)	(None, 8, 8, 256)	819456
batch_norm4 (BatchNormalization)	(None, 8, 8, 256)	1024
conv4_out (LeakyReLU)	(None, 8, 8, 256)	0
conv5 (Conv2D)	(None, 4, 4, 512)	3277312
batch_norm5 (BatchNormalization)	(None, 4, 4, 512)	2048
conv5_out (LeakyReLU)	(None, 4, 4, 512)	0
flatten (Flatten)	(None, 8192)	0
ligit (Dense)	(None, 1)	8193
Total params	Trainable params	Non-trainable params
4,367,553	4,365,569	1,984

MODEL: "FUNCTIONAL_3"

Layer (type)	Output Shape	Parameter
input_2 (InputLayer)	[(None, 100)]	0
Dense (Dense)	(None, 8192)	827392
leaky_re_lu (LeakyReLU)	(None, 8192)	0
reshape (Reshape)	(None, 4, 4, 512)	0
trans_conv1 (Conv2DTranspose)	(None, 8, 8, 512)	6554112
batch_trans_conv1 (BatchNormalization)	(None, 8, 8, 512)	2048
trans_conv1_out (LeakyReLU)	(None, 8, 8, 512)	0
trans_conv2 (Conv2DTranspose	(None, 16, 16, 256)	3277056
batch_trans_conv2 (BatchNormalization)	(None, 16, 16, 256)	1024
trans_conv2_out (LeakyReLU)	(None, 16, 16, 256)	0
trans_conv3 (Conv2DTranspose)	(None, 32, 32, 128)	819328
batch_trans_conv3 (BatchNormalization)	(None, 32, 32, 128)	512
trans_conv3_out (LeakyReLU)	(None, 32, 32, 128)	0
trans_conv4 (Conv2DTranspose)	(None, 64, 64, 64)	204864
batch_trans_conv4 (BatchNormalization)	(None, 64, 64, 64)	256
trans_conv4_out (LeakyReLU)	(None, 64, 64, 64)	0
logits (Conv2DTranspose)	(None, 64, 64, 3)	4803
out (Activation)	(None, 64, 64, 3)	0
Total params	Trainable params	Non-trainable params
11,691,395	11,689,475	1,920

- Result plot loss curve (Fig. 8):

The loss result it's almost 2%, and it's the complement the accuracy rate, and also the result rate of generator little some extent and the result rate of discriminator high some extent, this means that the capacity of the model is high, this is what is aimed to be achieved:

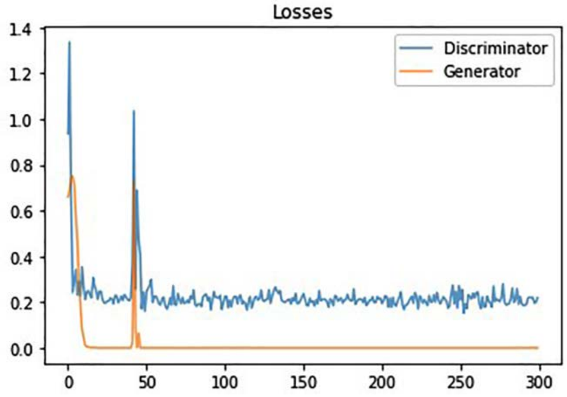


Fig. 8. Losses Result.

- Result Accuracy and Losses (Table IV):

TABLE IV. GANs RESULT

	Losses	Accuracy
GANs:	2.1%	97.9%

V. CONCLUSION AND FUTURE WORK

To summarize, DL was used in early prediction to detect diseases in different plant growth stages using the CNN algorithm for classification and prediction. Here, using the tomato infected with TMV as a model, the accuracy rate of TMV infection was 97%. The GANs used to increase the size of data and prediction accuracy rate by 98% when compared to the original data. For each plant growth phase, it became clear that the most growth stage group is vulnerable to viral infection is the second group. Therefore that determining the growth stages in this paper helped at obtaining results that prove the age group most susceptible to Unhealthy by determining the stages of Unhealthy also (healthy - first infection - Unhealthy), Thus, the study has concluded the previous results by applying to a set of real data that was collected manually from one of the farms in Egypt. Future work will include several DL models for early detection and classification of plant diseases due to using the rapid progress and improvements in DL models, transfer learning techniques, and CNN frameworks. Larger real-time dataset of TMV-infected tomato plants, and other important plant-disease system will be tested for attaining highest prediction accuracy. . Building a robust and accurate digital and computer-based plant pest-infestations and microbial disease-infections early-detection and warning system, will significantly help plant protection in early stages, with increased yield, quality, local marketing, and international exporting competitiveness.

REFERENCES

[1] M. Saleem, J. Potgieter, “Plant Disease Detection and Classification by Deep Learning” MDPI journals, 2019 Oct 31.

[2] P. Sobiczewski, “Bacterial Diseases of Plant: Epidemiology, Diagnostics, and Control” ISSN journals, 1392-3196, No. 3 2008.

[3] D. Singh, S. Teotia, “Fungal Disease Management in Plants” DOI: 10.1007/978-81-322-1620-9_19, October 2014.

[4] I. Ferriol, L. Galipienso, L. Rubio, “Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution” Frontiers in plant science, 17 July 2020.

[5] V. Sharma, A. Verma, “Classification Techniques for Plant Disease Detection” International Journal of Recent Technology and Engineering (IJRTE), March 2020.

[6] S. Kaur, S. Pandey, S.Goel, “Plants Disease Identification and Classification through Leaf Images: A Survey” Journal research gate, 2018.

[7] L. Kawaguchi, B. Saitama, S. Chikusa, N. Nagoya, “How Convolutional Neural Networks Diagnose Plant Disease” 2019.

[8] T. Salimans, I. Goodfellow, A. Radford, “Improve technique for training GANs” SAIL journal, 2016.

[9] J. Amara, B. Bouaziz, A. Algergawy, “A Deep Learning-based Approach for Banana Leaf Diseases Classification” Lecture Notes in Informatics (LNI), 2017.

[10] M. Brahimi, M. Arsenovic, S. Laraba, K. Boukhalfa, “Deep Learning for Plant Diseases: Detection and Saliency Map Visualisation” Part of the Human-Computer Interaction Series book series (HCIS), springer, June 2018.

[11] R. Yamashita, M. Nishio, R. Gian Do, K. Togashi, “Convolutional neural networks: an overview and application in radiology” Springer, 2018.

[12] S. Sakib, N. Ahmed, A. Kabir, H. Ahmed, “An Overview of Convolutional Neural Network: Its Architecture and Applications” Journal research gate, November 2018.

[13] Y. Kawasaki, H. Uga, H. Iyatomi, S. Kagiwada, “Basic Study of Automated Diagnosis of Viral Plant Diseases Using Convolutional Neural Networks” Springer international journal, 18 December 2015.

[14] T. Tran, T. Pham, G. Carneiro, L. Palmer, I. Reid, “A Bayesian Data Augmentation Approach for Learning Deep Models” Research gate, October 2017.

[15] D. Farm, “Image-to-Image Translation with GAN for Synthetic Data Augmentation in Plant Disease Datasets” Journal research gate, July 2019.

[16] A. Odena, C. Olah, J. Shlens, “Conditional Image Synthesis with Auxiliary Classifier GANs” arXiv: 1610.09585v4, 20 Jul 2017.

[17] S. Osindero, M. Montreal, “Conditional Generative Adversarial Nets” ArXiv: 1411.1784v1 cs.LG, 6 Nov 2014.

[18] M. Gong, Y. Xu, C. Li, K. Zhang, K. Batmanghelich, “Twin Auxiliary Classifiers GAN” NIPS journal, 2019.

[19] K. Evtimova, A. Drozdov, “Understanding Mutual Information and its Use in InfoGAN” Semantic scholar, 2016.

[20] A. Khan, A. Sohail, U. Zahoora, A. Saeed, “A Survey of the Recent Architectures of Deep Convolutional Neural Networks” Springer, Published: 21 April 2020.

[21] M. Pashaei, H. Kamangir, M. Starek, P. Tissot, “Review and Evaluation of Deep Learning Architectures for Efficient Land Cover Mapping with UAS Hyper-Spatial Imagery: A Case Study over a Wetland” Multidisciplinary Digital Publishing Institute (MDPI) journal, Published: 16 March 2020.

[22] K. Shridhar, J. Lee, H. Hayashi, P. Mehta, B. Iwana, S. Kang, S. Uchida, S. Ahmed, A. Dengel, “ProbAct: A Probabilistic Activation Function for Deep Neural Networks” arxiv journal, 16 Jun 2020.

[23] M. Zitnik, B. Zupan, “Data Fusion by Matrix Factorization” IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2013.

Mammogram Segmentation Techniques: A Review

Eman Justaniah[1], Dr. Areej Alhothali[2]
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

DR. Ghadah Aldabbagh[3]
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia
MIT Department of Mechanical Engineering
Cambridge, MA, USA

***Abstract*—There is a significant development in computer-aided detection (CADe) and computer-aided diagnostic (CADx) systems in recent years. This development coincides with the evolution of computing power and the growth of data. The CAD systems support detections and diagnosis of significant diseases, including cancer. Breast cancer is one of the most prevalent cancers influencing women and causing death around the world. Early detection of breast cancer has a significant effect on treatment. The typical CAD system goes through various steps, including image segmentation, feature extraction, and image classification. Image segmentation plays an important role in CAD systems and simplifies further processing. This review explores popular mammogram segmentation techniques. A mammogram is medical imaging which uses a low-dose x-ray system to see inner tissues of the breast. There are many segmentation techniques used to segment medical images. These techniques can be categorized into five main categories: region-based methods, boundary-based methods, atlas-based methods, model-based methods, and deep learning. A ground truth image is needed to measure the performance of the segmentation algorithm. Different performance measurements were used to evaluate the segmentation process, including accuracy, precision, recall, F1 score, Hausdorff Distance, Jaccard, and Dice Index. The research in mammogram segmentation has yielded promising results, but there is room for improvements.**

Keywords*—*Mammogram; medical imaging; segmentation; preprocessing; breast cancer

I. Introduction

Over the years, Artificial Intelligence (AI) algorithms have been improving and having impact on every aspect of human life. In recent years, there has been a significant development in machine learning techniques and high-performance computers, along with a massive increase in digital data in various fields. Diagnosing diseases through radiology is an important medical application of AI algorithms. An example of this application is CADe and CADx systems. The CAD systems are used to assess patient's diagnostic images by clinicians and radiologists. Most CAD systems consist of the following steps: image preprocessing, segmentation, feature extraction, and classification. There are many studies conducted in using a CAD system to diagnose and detect breast cancer from medical imaging [1]. This review discuss different aspects related to mammogram segmentation. The rest of the review is divided into the following sections: Section II provides background about medical imaging and mammogram. Section III includes a description of some public mammogram datasets. Section IV explains performance measurements used in mammogram segmentation. Section V discusses different segmentation techniques used in mammogram images. Section VI is a discussion of studies mentioned in the review. The last section concludes the paper.

II. Background

This section provides background about medical image analysis, breast cancer, and mammogram images.

A. Medical Image Analysis

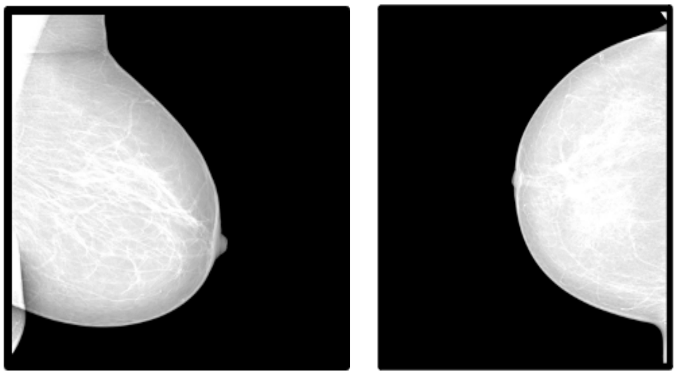
Medical images are different from regular photos; they represent physical features measured from the human body. Therefore, the analysis of medical images must be guided by particular expectations and follow a medical reference. AI has been used in medicine since the 1980s [2]; later on AI medical applications are continuously expanding. Nowadays, medical image analysis has become a branch of artificial intelligence. There are books, academic journals, and conferences for medical image analysis research. There are various types of medical images, including X-ray imaging, magnetic resonance imaging, ultrasound, nuclear imaging, optical microscopy, etc. X-ray Imaging uses electromagnetic waves with a wave-length above the visible spectrum to produce a diagnostically meaningful image. Fluoroscopy and angiography, Computed Tomography (CT), and mammography are kinds of X-ray imaging [1].

B. Breast Cancer

Breast cancer is a disease caused by an abnormal growth of breast cells [3]. It is one of the most prevalent cancers influencing women and causing death around the world [4]. Early detection of this disease increases the recuperating rate significantly [5]. Three main types of examinations are commonly used to detect breast cancer: 1. self-examination performed by the patient herself, 2. a clinical examination conducted by well-trained specialists, 3. a radiology examination conducted by a radiologist using visual evaluation. Studies show that the most accurate radiologic procedure for early detection of breast cancer is the mammogram [4].

C. Mammogram Images

A mammogram is medical imaging that aims to see inside tissues of the breast by using a low-dose x-ray system [3]. There are two imaging modalities of mammograms: digital mammogram and screen-film mammography. The screen-film mammography (SFM) contains conventional analog mammography films. Usually, SFM contains labels and markers in the background, which considered as noise and need to be removed. The digital mammograms are also called Full-Field Digital Mammography (FFDM) images. The FFDM is more recent and does not include labels [2]. Moreover, mammogram images can be found in several formats including LJPG, DI- COM, PGM, and TIFF. In the stander view for each breast, two X-ray images need to be taken on both sides. Therefore, four images of both breasts need to be examined. These four images are called: LEFT CC, LEFT MLO, RIGHT CC, RIGHT MLO [6], [7]. The Craniocaudal (CC) view is obtained from top horizontally compressed breast (head-to- foot picture). The CC view captures the medial portion and the breast's outer lateral region as much as possible. The Medio Lateral-Oblique (MLO)–side view–captures the whole breast and usually contains the lymph nodes with the pectoral muscle. Fig. 1(a) and 1(b) show the example of CC and MLO views. And Fig. 2 illustrates the angle of each view [8].



[(b) MLO view] [(a) CC view]

Fig. 1. FFDM View.

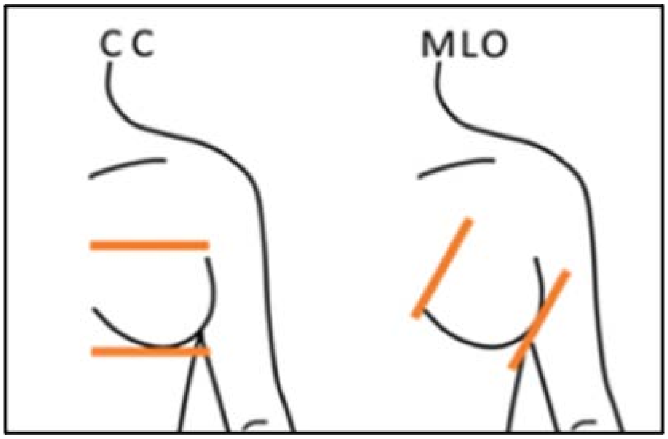


Fig. 2. The Different between CC and MLO Views.

III. Public Mammogram Datasets

There are several mammogram datasets publicly available. Following is a brief description of the most used datasets, which are referenced in studies cited in the review.

A. Mammographic Image Analysis Society (MIAS)

The Mammographic Image Analysis Society (MIAS) is a research group from the UK interested in studying mammograms. This group generated a small mammogram database in 1994 called mini-MIAS or MIAS for short. The mini-MIAS consists of 322 digitized films stored in the PGM image format. Every image has a resolution equal to 1024 × 1024 pixels [9].

B. Digital Database for Screening Mammography (DDSM)

The DDSM project is a collaborative effort between the Massachusetts General Hospital, the University of South Florida, and Sandia National Laboratories. The dataset includes 2620 cases. A case consists of between 6 and 10 files. These are an 'ics' file, an overview "16-bit PGM" file, four image files compressed with lossless JPEG encoding, and zero to four overlay files [10].

C. INbreast

INbreast is a full-field digital mammographic database. The cases were collected from Centro Hospitalar de S. Joa o [CHSJ], Breast Centre in Portugal, in 2011. The database includes 115 cases with a total of 410 images. The resolution of images was 3328 4084 or 2560 3328 pixels and saved in the DICOM format. The region of interest (ROI) was annotated by two specialists and stored in separate .roi and .xml files [11].

D. Breast Cancer Digital Repository (BCDR)

The IMED Project supported the creation of BCDR. The IMED project was supported by FMUP-CHSJ University of Porto, Portugal, INEGI, and CETA-CIEMAT Spain, from March 2009 till March 2013. The BCDR includes 1734 cases with mammography and ultrasound images. Also, it includes clinical history, mammogram lesion segmentation, and selected pre-computed image-based descriptors. The dataset is subdivided into Full Field Digital Mammography-based Repository (BCDR-DM), and Film Mammography-based Repository (BCDR-FM) . Mammogram images were saved in the TIFF format. The BCDR-FM part has a resolution of 720 x 1168 pixels and 8 bits depth. While the BCDR-DM resolution is equal to 3328 x 4084 pixels and 14 bits depth [12].

E. Curated Breast Imaging Subset of DDSM (CBIS-DDSM)

CBIS-DDSM is an updated and standardized version of the Digital Database for Screening Mammography (DDSM) stored in the DICOM file format [13]

Table 1 summarizes the datasets mentioned.

TABLE I. Public Mammogram Datasets

Dataset Name	Size	Format	Type	Published Year
MIAS	322 images	PGM	Digitized	1994
DDSM	2620 cases	PGM, JPEG	Digitized	1998
INbreast	410 images	DIOCM	Digital	2011
BCDR	1734 Cases	TIFF	Hybrid	2012
CBIS-DDSM	2620 cases	DICOM	Digital	2019

IV. Performance Measurements

There are several ways to measure the performance of the segmentation technique. If the ground truth image of the target area is available, then the Dice similarity coefficient or Jaccard Index can be used. Dice Similarity Coefficient (DSC or dice): equivalent to twice the number of elements common on both sets divided by the sum of the number of elements in each set. DSC is usually used for auto-segmentation models and computed by this equation:

$$Dice(A,B) = \frac{2|A \cap B|}{|A|+|B|} \quad (1)$$

Where A represent the segmented image resulted from the algorithm, and B represent the ground truth image. Jaccard Index or Intersection over Union (IoU) is another similarity measurement. IoU computes the similarity ratio of elements in two sets, A and B, as set intersection over the number of elements in the set union:

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Hausdorff Distance is also used to assess the medical image analysis algorithm's performance. This measurement used when outliers need to take it into account. The Hausdorff distance, h (A, B), is given by Equation (3):

$$h(A,B)=\max \left[max_{a\,\in A}\, min_{b\,\in B}\ d(a,b), max_{\,b\,\in B}\, min_{a\,\in A}\ d(a,b) \right] \quad (3)$$

where d (a, b) is the Euclidean distance between the points a and b [14].

Other performance measurements used in CAD systems are accuracy, sensitivity, specificity, precision, recall, and F1 score. Following are equations for these measurements.

$$\text{Accuracy}=\frac{\text{Number of examples identify correctly}}{\text{Total number of example}} \quad (4)$$

To compute precision and recall, the confusion matrix must be created first. A confusion matrix is a table used to describe the performance of a classification model. The table II illustrates the confusion matrix:

Positive here mean a target class, for example in brest cancer detection problem the positive class is a cancer or abnormal masses, and negative class is the mammogram with no cancer detected [15].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1score} = 2\times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (7)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (8)$$

False positive per image compute by following equation:

$$\text{FP/Image}=\frac{\text{Number oF identified}}{\text{Total number of images}}$$

TABLE II. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual class	Positive	True positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

V. Medical Imaging Segmentation

Image segmentation aims to simplify further processing by partitioning the digital image into regions that share similar characteristics. Fig. 3 shows a block diagram of the standard CAD system. There are many segmentation techniques used in segmenting mammograms. These techniques can be categorized into five primary types: region-based methods, boundary-based methods, atlas-based methods, model-based methods [16], and deep learning [17], [5].

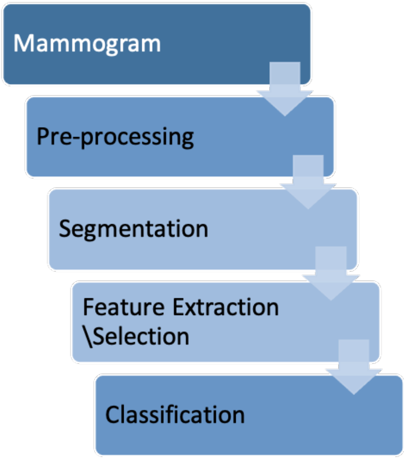


Fig. 3. CAD System Block Diagram.

A. Region based Segmentation

In region-based methods, a segmentation is done based on similarities between regions. Thresholding, Region-growing, watershed, split and merge, and clustering are types of region-based segmentation methods [16].

1) Thresholding: Thresholding is mostly used to separate an image into a background and foreground object. First, a specific value T is selected as a threshold value based on image histogram and local properties. All pixels below T will be considered background, and all pixels equal to or greater than T will be considered foreground. Using multilevel thresholding gives a better result, the authors in [18] proposed a CAD system detecting suspicious mass lesions in the mammogram. The proposed system starts with three pre-processing steps. First, the median filtering with a 3 x 3 window is used to remove noise. Second, morphological operations are applied to remove artifacts and background. At the last preprocessing step, a single-seeded region-growing algorithm is used to remove pectoral muscles. The second phase in the proposed CAD is detecting mass using Dual-stage adaptive thresholding. The performance was measured by sensitivity and false-positive per image (FP/image). The evaluation was done on DDSM and MAIS datasets. The result was sensitivity= 93, FP/image = 0.84. The work [19]. proposed a hybrid approach based on Otsu's multi-thresholding and Watershed Segmentation (WSS) to mine the suspicious sections from mammograms. They used images from the MAIS dataset and measure the performance with many measurements includes Root Mean Square Error (RMSE) and Normalized Absolute Error (NAE). Different thresholding levels were tested; however, th=4 gave the best results, RMSE= 21.7732 and NAE= 0.2429. The authors in [20]. developed a fully automated pectoral muscle segmentation method. This method consists of four steps. First, capturing a small rectangular region in the top-left corner of mammograms and enhancing it using the fractional differential method. Second, segmenting a rough binary boundary of the pectoral muscle in the rectangular region, using an improved iterative threshold method. Third, adapting a rough contour with the least-squares method based on points of the rough boundary. Finally, evaluate the local active contour to acquire the final pectoral muscle segmentation line. The dataset consists of 720 MLO, which are FFDM. The overall performance of this method in the Dice coefficient equal to 0.986±0.005. The authors in [21] , proposed multilevel thresholding based on the electro-magnetism optimization (EMO) technique to segment pectoral muscles. EMO is an evolutionary method that mimics the attraction-repulsion mechanism among charges to evolve the members of a population. The first step is to crop the mammogram image. The second step is extracting a region of interest (ROI) using the stepwise contrast limited adaptive histogram equalization (CLAHE) algorithm. Also, the CLAHE method is used to enhance contrast in mammogram images. The third step is to enhance the image using the histogram equalization technique. In the fourth step, the EMO algorithm with Otsu objective function and Kapur objective function is applied. Finally, the straight-line estimation is used to identify the pectoral muscle. This segmentation was tested on the MAIS dataset and gave an accuracy = 96.58%. The work [22] proposed an adaptive hysteresis thresholding method to detect mammogram masses. This method was applied on MAIS and DDSM datasets and gave sensitivity equal to 96.6%, 96.4%, respectively.

2) Region growing: In a region-growing segmentation, algorithm starts with seed points representing each class of image (e.g., background and foreground classes). Each class grows according to the homogeneity of neighboring pixels; this process continues until reaching homogenous and connected regions [23]. The work [24] ,proposed an automated mammogram segmentation based on region growing and sliding window algorithm (SWA). First, the authors prepared the MIAS dataset by removing artifacts and labels using the opening morphological operator and binary mask. Then remove pectoral muscle using SWA and segment mammogram ROI using Dispersed Region Growing Algorithm (DRGA). The overall accuracy of this approach equals 91.3%. The authors in [25] , proposed a pectoral muscle segmentation and tumor detection approach. This approach starts with the Otsu method to remove artifacts. The region-growing method is used to eliminate the pectoral muscle. Then estimate the number of classes based on the LBP Technique and classify mammogram objects using K-means clustering. Finally, they extract the tumor by a hidden Markov model. The proposed approach was tested on the MAIS dataset, and the overall accuracy = 91.92 %. The work [26] aapplied an adaptive fuzzy region-growing algorithm on two FFDM private datasets to segment suspicious lesions and characterize them. The performance was measured by sensitivity and specificity, and the results were 91.67%, 58.33%, respectively. After detecting suspicious lesions, k-NN and SVM classifiers were used to classify masses as benign masses or malignant tumors. The classification results for k-NN and SVM achieved sensitivity = 84.44% and 85.56%, specificity = 91.11% and 91.67%, FPsI = 0.54 and 0.55 respectively. The authors in [27] , proposed another method to detect the lesion's boundaries in mammogram images based on the region-growing algorithm. The MAIS dataset was used in this work. The performance was measured by accuracy, specificity, sensitivity, and overlap, and the results were 91%, 97 %, 83%, and 79%, respectively. The work [28] , proposed an automatic breast cancer detection approach consisting of four amin processes applied on the MAIS dataset. The first phase is the preprocessing, enhancing images, and removing noise using median filtering. The second phase is mammogram segmentation using region growing. The third phase is feature extraction. Finally, the classification phase uses an optimized fuzzy logic classifier. The performance was measured for the segmentation and classification phase. The segmentation accuracy = 0.98%, and the fuzzy classifier accuracy = 0.91667 %. The authors in [29] proposed a segmentation method based on region-growing techniques. The proposed method included four main steps and was applied on MAIS and DDSM. The first step was extracting the Region of Interest from mammogram images. At the second step, automatic thresholding was applied to binarize the image. The third step was determining the seed points automatically using the density of the pixels' value. Finally, they calculated the threshold value for region creation in seed region growing. The results show that the Dice Similarity Coefficient (DSC) =94.8, 94.6, and Relative Overlap (RO) = 90.2, 89.8 for MAIS and DDSM, respectively.

3) Watershed: The key behind using the watershed transform for segmentation is this: Change the image into another image whose catchment basins are the target objects. Watershed Algorithm is based on simple morphological operations [23]. The work [30] , proposed a two-phase micro-classification segmentation approach. First, detection microcalcifications used morphological operations. Second, the micro-calcification shape was extracted using the watershed. This approach was applied on DDSM, and the overall performance in dice (similarity index) equals 80.5%. The authors in [31] proposed another segmentation approach based on the watershed algorithm. This approach consists of four stages. In the first stage, the ROI images were cropped to 200 x 200 pixels, and the background was removed. In the second stage, the Principal Component Analysis (PCA) method was applied on the cropped image to remove the noise. In the third stage, the Fuzzy C-Means (FCM) was applied to partition the ROI images into the foreground and background clusters. The foreground includes the abnormality region, which will be used in the final stage. In the last step, marker-controlled watershed segmentation was performed with three various structuring elements: disk, diamond, and octagon shapes. The dataset used in this study was obtained from the National Cancer Society of Malaysia. The performance was measured by computing Jaccard Index, Dice Similarity Coefficient (DSC), and Figure of Merit (FOM). The Jaccard index = 0.0452 and DSC = 0.0231 in both disk and octagon structures; in diamond structure, Jaccard index = 0.0405 and DSC = 0.0207. The result of FOM was 0.9594, 0.9700, 0.9842 for disk, diamond, and octagon structures, respectively. The work [32], proposed a semi-automatic segmentation of masses from mammogram images. The proposed approach includes three main stages. First, the median filter was applied to enhance image quality. Second, an initial segmentation was composed based on canny and watershed algorithms. Finally, the boundaries of tumors were extracted using the region- growing algorithm. The MAIS dataset was used and the performance measured by overlap value was equal to 81.3%.

4) Splitting and merge: Split and merge depend on the tree structure, the image splitting successively into quadrants tree based on a homogeneity criterion. Then similar regions are merged to create the segmented result. The work [33] , used a blended approach of region-based method and splitting and merging technique. The proposed approach is applied on MIAS mammogram images. First, the morphological operation is used to remove the noise from images. Then, the splitting step is performed based on the region's growing method (seed points). Finally, at the merge step, the binary values are reconstructed to form a structured mammogram image. The structured image is completed for finding the seed point and grown points. The performance measured by five statistical parameters: mean = 0.0759, variance = 0.0702, entropy = 6.521 standard deviation = 0.2649, and correlation = 0.7869.

5) Clustering: In clustering, pixels are grouped into clusters, in which pixels in the same cluster are more similar to each other than to those in different clusters. The two types of clustering used in image segmentation are K-means clustering, and fuzzy C means clustering [23], [16].

a) K-means clustering: The K-means clustering algorithm start by setting K centroid points (or pixel values). Then assign the remaining pixels to their closest cluster center for each cluster. Based on the resulting cluster, reset a suitable centroid of each cluster. These two steps repeat until the algorithm meets the chosen criteria. In segmentation, the value of k depends on the number of objects want to be extracted.

b) Fuzzy C means clustering: Fuzzy C means (FCM) is a kind of clustering in which a single data point can belong to more than one cluster [34]. The authors in [35] applied FCM and K-mean clustering algorithms on the MAIS dataset to segment mammogram images. The performance was measured by accuracy. The accuracy of FCM was 94.12%, while K-mean accuracy was 91.18%.

B. Boundary-based Segmentation

Unlike region-based segmentation, boundary- or edge-based segmentation depend on differences between regions. There are variety boundary-based segmentation techniques. Roberts, Sobel, Prewitt, Laplacian, and Canny edge detection are examples of boundary-based segmentation techniques [23], [16]. The work [36] proposed a method that segments the breast boundary and pectoral muscle in mediolateral oblique (MLO) views of mammograms automatically. The proposed method consists of three main stages. The first stage removes noise from mammogram images by applying median and anisotropic diffusion filters. The second stage segment the mammogram using Canny edge detection. Finally, the overestimated boundary caused by artifacts was handling by a proposed post-processing stage. Three public datasets were used to evaluate this method including MIAS, INbreast, and BCDR. Experimental results show that dice similarity coefficients on breast boundary and pectoral muscle estimation were equal to 98.8% and 97.8% for MIAS, 98.9% and 89.6% for INbreast, and 99.2% and 91.9% for BCDR respectively.

C. Atlas-based Segmentation

Atlas-based segmentation is an algorithm that aims to extract the relevant anatomy from medical images and to present it in an appropriate view [37]. The atlas-based approach is suitable for segmenting images with unclear associations between regions' and pixels' intensities [38]. The authors in [14] ,proposed an atlas-based algorithm to segment breast area from mammogram images. The preprocessing step includes standardizing mammogram images by flipping the left breast mammograms so that all mammograms had the same orientation. Then make the images square by padding on the left and right, then remove this padding and determine the breast region. The main algorithm consists of two stages. In the first stage, select a set of atlas mammogram images using the K-means clustering algorithm. The number of clusters is determined by applying 2D projection using tributed Stochastic Neighbor Embedding (t-SNE). In the second stage, they used atlas mammogram images with a deformable registration algorithm to segment the images. They tested this algorithm on mini-MIAS and DDSM datasets. The performance measurements used were Hausdorff Distance = 13.34 and Jaccard Index= 0.94.

D. Model-based Segmentation

Model-based segmentation, or energy functions, are based on deformable models. We can define the deformable models as curves that deform due to some external or internal force [16]. This group of segmentation techniques has the ability to integrate high-level knowledge with information from low-level image processing. There are two classes of deformable models, parametric and non-parametric. The parametric deformable model is also called the active counter model. The work [39] , proposed a bimodal level-set formulation-based approach for mammogram segmentation. They used the mini-MIAS dataset and drew ground truths manually using a hand-based polygonal tool. Compared with the Chan-Vese and Zhang models, the pro- posed approach achieved Precision AVG = 0.9448, Recall AVG = 0.975 within only 4-6 iterations, while the other two models required more than 60 iterations to get such results. The authors in [40] , proposed a preprocessing method for the mammogram CAD system, include pectoral muscle segmentation. The proposed method consists of four phases. First, remove noise using median and mean filters. Second, enhance image quality using the CLAHE algorithm. Third, remove radiopaque artifacts and labels present in mammograms by applying thresholding and morphological operations. Finally, using active contours to remove pectoral muscle. This preprocessing approach was tested on two datasets, mini-MIAS and INbreast. The results show that accuracy equals 90%, 98.75% for mini-MIAS and INbreast, respectively. The work [41] , provided an automatic mammogram image segmentation approach based on the Chan–Vese model. The target ROI consists of three classes. The mass class contains pixels pertaining to the mass. Background class includes background pixels which not pertain to the mass class.The remaining pixels, which separate the mass from the background belong to the contour class. The proposed approach consists of the Contour initialization step, fuzzy contours estimating step, and Contour optimization step. In the Contour initialization step, the gamma correction was used to improve the image contrast; then, Otsu thresholding was applied to binarize the mass region. The fuzzy contours estimation step aims to refine the initial contour. The last step is contour optimization using the Chan–Vese model. The accuracy of the proposed method was 93.96%, while precision and recall equal 88.08%, 91.12%, respectively.

E. Deep Learning

Deep learning is an artificial intelligence technique that can learn a pattern from raw data [42]. A typical deep learning algorithm is called artificial neural networks or multi-layer perceptron (MLP). Artificial neural networks (ANNs) is a mathematical model developed to mimic the operations of the human neurophysiological structure [43]. The authors in [44] , proposed a neural network framework to deal with complex shape variations of the pectoral muscle boundary in mammogram segmentation. This framework consists of a convolutional neural network inspired by Holistically Nested Edge Detection network (HND). The main benefit of HND is that it can deal with edge and boundary ambiguities of the object. The performance of this approach was measured by computing Jaccard and Dice metrics. Four public datasets were used in the study, including MIAS, INbreast, BCDR, and CBIS-DDSM. On average the Jaccard equals 94.6%, and dice similarity equals 97.5%. The work [45] proposed automated mass segmentation from mammograms based on a multi-level nested pyramid network (MNP Net). The proposed MNPNet divided into three subsections and employed an Encoder-Decoder framework. In the first section, the atrous spatial pyramid pooling (ASPP) module encoding was used to solve the intra-class inconsistency. In the second section, the multi-level feature pyramid produced by CNN was used to improve inter-class indistinction. In the last section, different ResNet structures were used to perform feature extraction, and the ResNet34 was the best. The proposed segmentation is applied on INbreast and DDSM-BCRP and achieves dice index equal to 91.10% and 91.69%, respectively. The authors in [46] , proposed an approach to segment breast tumors within mammograms' ROI using a conditional Generative Adversarial Network (GAN). They tested the model on two datasets INbreast and a Hospital Sant Joan de Reus private dataset. The cGAN network learns a complex pattern from simple data and has two subnetworks, generative and adversarial networks. The generative network recognizes the tumor area and generates the binary mask that detects it, while the adversarial network distinguishes between real (i.e., true) and synthetic segmentations. The performance of the cGAN model on mammogram segmentation task was measured using dice and intersection over union (IoU) performance metrics. For the INbreast dataset, the dice and IoU on the full mammogram image were 68.69% and 52.31%, respectively. After generating mask images, these images feed to a Convolutional Neural Network (CNN) to classify the tumor into one of four types: irregular, lobular, oval, and round. The The overall accuracy of the CNN classifier was 80%. The work [47] ,proposed an automated segmentation to detect microcalcification from mammograms. This approach consists of five steps; image enhancement, removing skin and air boundary, segmenting pectoral region, selecting suspicious region, and U-net segmentation. First, the Laplacian filter was applied to enhance mammogram images. Second, skin and air boundaries were removed using horizontal line fitting and the image erode method. Third, the breast region is segmented from the pectoral region by K-means pixel-wise clustering. Fourth, suspicious regions were selected using the fuzzy C-means clustering algorithm and were divided into positive and negative patches. Fifth, the U-net was trained on the positive patches of the previous step. Finally, the trained U-net was applied to segment the micro-calcification regions automatically from mammograms. This approach was applied on the DDSM dataset and measured by F-measure = 98.5%, Dice score = 97.8%, and Jaccard index = 97.4%. The authors in [48], used the Dense U-Net algorithm to segment suspicious breast masses from mammograms. The evaluation was done on the DDSM dataset. The performance was measured by F1-score, sensitivity, specificity, and overall accuracy and results equal 82%, 77.89%, 84.69%, and 78.38% respectively. The work [49] also used U-Net to detect mass from mammograms. Moreover, they used different data augmentation techniques, such as image zoom, extracting nine regions of interest, and horizontal reversal. The training and evaluation were done on the DDSM dataset. The overall accuracy was equal to 85.95% and the Dice was equal to 79.39%.

VI. Discussion

The investigations mentioned in this review were selected to meet the following criteria:

1) The date of publication should be after 2016.

2) It was published in the ISI magazine or the ACM or IEEE

3) The work should contain precise quantitative performance.

Several segmentation techniques were discussed in this review. Table III summarizes them. As seen from Table III, there are different targets of segmentation. Some segmentation techniques aim to detect suspicious lesions (mass or tumor). Other methods aim to remove background or pectoral muscles. Moreover, there are segmentation techniques that target microcalcification. Table III categorize the segmentation techniques based on the target area. Table IV lists the pre-processing techniques used with each segmentation. The pre-processing techniques include filtering, applying morphological operations, performing data augmentation, and some other enhancement methods. Median filter, Gaussian filter, Bayesian non-local mean filter, Anisotropic Diffusion filter, and Laplacian filter were used to enhance mammogram images. Also, the morphological operations were used in most mentioned works. The data augmentation was mostly used with deep learning-based segmentation techniques. Moreover, Transformer (Intensity, Gamma) and the CLAHE method were used in some papers [18] [49], [21], [40]. The threshold is mostly used as a segmentation technique, but also could be used as a pre-processing step to remove unwanted regions in mass detection techniques [26], [41]. Also, the principal component analysis PCA was used to denoising images [31]. To simplify the processes, many researchers used to resize or cropping in the pre-processing step.

Some papers combine different segmentation techniques such as watershed and region growing. Among different mass segmentation techniques, improved region growing [28] gave the highest accuracy, while Canny edge detection [36] outperforms other approaches in pectoral muscle segmentation. Although studies in mammogram segmentation have yielded good results, there is room for improvement.

TABLE III. Segmentation Techniques Summary

Technique	**Year**	**Target**	**Enhancing**	**Augmentation**	**Dataset**	**Performance**
A Region-based segmentation *1. Thresholding*						
A dual stage adaptive thresholding [18]	2017	Mass detection	√	–	MIAS and DDSM	Sensitivity=93.0, FP\Image=0.84
Otsu’s Threshold and Watershed [19]	2018	Extract ROI	√	–	MAIS	RMSE= 21.77, NAE= 0.24
Iterative threshold + active contours [20]	2019	Remove pectoral muscle	√	–	720 MLO images	Dice= 0.986±0.005
EMO algorithm [21]	2019	Pectoral muscle segmentation	√	–	MAIS	Accuracy=96.58%
Adaptive hysteresis thresholding [22]	2019	Mass detection	√	√	MAIS DDSM	Sensitivity = 96.6 % Sensitivity = 96.4 %
A. Region-based segmentation *2. Region Growing*						
Region growing and Sliding Window [24]	2017	Remove pectoral muscle	√	–	MIAS	Accuracy= 91
Region Growing [25]	2018	Detecting the lesion’s boundaries	√	–	MIAS	Accuracy = 91%, Specificity= 97 % Sensitivity= 83% Overlap= 79%
Hidden Markov and region growing [26]	2018	Tumor detection	√	–	MIAS	Accuracy= 91.92 %
Adaptive fuzzy region growing [27]	2018	Detect suspicious lesions	–	–	360 FFDM images	Sensitivity=91.67%, Specificity= 58.33%
Improved region growing [28]	2020	Tumor segmentation	√	–	MAIS	Accuracy= 98%
Efficient Seed Region Growing [29]	2020	Tumor segmentation	√	–	MAIS DDSM	Dice= 94.8 Dice =94.6
A. Region-based segmentation *3. Watershed*						
Morphological Approach [30]	2017	Segment microcalcification	√	–	DDSM	Dice=80.5%
Watershed Algorithm [31]	2018	Mass segmentation	√	–	Private SFM data	IoU Disk= 0.045, Dice Disk= 0.023 IoU Diamond= 0.0405 Dice Diamond= 0.0207 IoU Octagon= 0.0452 Dice Octagon= .0231

Watershed and region growing [32]	2018	Tumor segmentation	√	–	MAIS	Overlap =81.3 %
A. Region-based segmentation ***4. Split & merge***						
Split and merge [33]	2016	Mass segmentation	√	–	MAIS	Entropy = 6.521
A. Region-based segmentation ***5. Clustering & merge***						
Clustering K-means and FCM [35]	2019	Mass segmentation	√	–	MAIS	Accuracy=91.18% Accuracy= 94.12%
B. Boundary-based segmentation						
Canny edge detection [36]	2017	Pectoral muscle segmentation	√	–	MIAS BCDR Inbreast	Dice= 98.8% Dice= 98.9% Dice= 99.2%
C. Atlas-based segmentation						
Atlas-based Segmentation [14]	2019	Segmenting breast region	√ –	MAIS, & DDSM	Hausdorff= 13.34, IoU =0.94	Atlas-based Segmentation [44]
D. Model-based segmentation						
Active contours [39]	2017	Mass segmentation	– –	MAIS	Prescision= 94.48% Recall= 97.5%	Active contours [48]
Four phases pre-processing [40]	2017	Pectoral muscle segmentation	√ –	MAIS Inbreast	Accuracy= 90% Accuracy= 98.75%	Four phases pre-processing [6]
Chan–Vese model [41]	2018	Mass segmentation	√ –	MAIS	Accuracy = 93.96%	Chan–Vese model [19]
E. Deep learning						
Convolutional Neural Network [44]	2019	Pectoral muscle segmentation	– –	MAIS BCDR INbreast CBIS-DDSM	IoU =94.6%, Dice=97.5% IoU = 96.9%, Dice=98.8% IoU = 92.6%, Dice=95.6% IoU = 95.1%, Dice=94.8%	Convolutional Neural Network [37]
MNPNet [45]	2019	Mass segmentation	– √	Inbreast CBIS-DDSM	Dice=91.10% Dice=91.69%	MNPNet [51]
cGAN [46]	2019	Segment a breast tumor	√ –	INbreast	Dice=68.69%, IoU=52.3%	cGAN [47]
U-net [47]	2019	Segment microcalcification	√ √	DDSM	Dice=97.8%, IoU=97.4%	U-net [5]
Dense U-Net [48]	2019	Mass segmentation	√ √	DDSM	F1=82 %, Accuracy= 78.38%	Dense U-Net [28]
U-net with Data Augmentation [49]	2020	Mass detection	√ √	DDSM	Accuracy = 85.95% Dice = 79.39%.	U-net with Data Augmentation [54]

TABLE IV. SUMMARY OF PREPROCESSING TECHNIQUES

-	**Data Set**	**Filtering**	**Morphological operations**	**Augmentation**	**Resizing⁄Cropping**	**Other Algorithms**
Mass⁄Suspicious lesions Detection						
A dual-stage adaptive thresholding [18]	MIAS and DDSM	√	√	—	—	CLAHE method
Otsu's Threshold and Watershed [19]	MAIS	—	√	—	—	—
Adaptive hysteresis thresholding [22]	MIAS and DDSM	√	√	√	√	—
Hidden Markov and region growing [26]	MAIS	—	—	—	—	Otsu's method
Adaptive fuzzy region growing [27]	360 FFDM	—	—	—	√	Intensity Transformation
Improved region growing [28]	MAIS	√	—	—	—	—
Efficient Seed Region Growing [29]	MIAS and DDSM	√	√	—	√	—

Watershed Algorithm [31]	Private SFM	—	√	—	√	PCA (to denoise)
Watershed and region growing [32]	MAIS	√	—	—	—	—
Split and marge [33]	MAIS	—	√	—	—	—
Clustering [35]	MAIS	—	√	—	—	—
Chan–Vese model [41]	MAIS	—	√	—	—	Otsu’s method
MNPNet [45]	INbreast CBIS-DDSM	–	–	√	√	–
Dense U-Net [48]	DDSM	—	—	√	√	Gamma transform
U-net with Data Augmentation [49]	DDSM	—	—	√	√	CLAHE
Breast Area ⁄ Pectoral muscle segmentation						
Iterative threshold + active contours [20]	720 MLO images	√	√	—	√	—
EMO algorithm [21]	MAIS	√	—	—	√	CLAHE method
Region growing and Sliding Window [24]	MAIS	—	√	—	—	—
Region Growing [25]	MAIS	√	√	—	—	—
Canny edge detection [36]	MAIS, BCDR, INbreast	√	—	—	√	—
Active contours [39]	MAIS	—	—	—	—	—
Atlas-based Segmentation [14]	MAIS, DDSM	—	—	—	√	Image flipping
Four phases pre-processing [40]	MAIS,INbreast	√	√	—	—	CLAHE method
Convolutional Neural Network [44]	MAIS, BCDR, INbrest CBIS-DDSM	—	—	—	—	—
cGAN [46]	INbrest	√	√	—	√	—
Microcalcification Segmentation						
Morphological Approach [30]	DDSM	√	√	—	√	—
U-net [47]	DDSM	√	√	√	—	—

VII. Conclusion

In this review, an elaborate coverage has been performed in mammogram segmentation techniques. First, we provided an overview of medical image analysis and described the mammogram images. Then we gave a brief description of MIAS, DDSM, INbreast, BCDR, and CBIS-DDSM datasets. We discussed region-based segmentation, boundary-based segmentation, atlas-based segmentation, model-based segmentation, and deep learning approaches for segmentation; we gave an ex- ample from recent papers for each of these segmentation techniques. Then we explained the most-used performance measurement in the segmentation process. Finally, we summarized different mammogram segmentation works in table III, including the preprocessing step, dataset(s), and performance results for each one.

References

[1] K. D. Toennies, Guide to medical image analysis, Springer, 2017.

[2] P. M. P. M. R. a. V. K. Amisha, "Overview of artificial intelligence in medicine," Journal of family medicine and primary care, 2019.

[3] C. E. G. Department, "Cancerous Diseases, Breast Cancer," Ministry of health, August 2018. [Online]. Available: http://tiny.cc/breastDef. [Accessed 2020].

[4] R. M.,. R. H. R. ,. J. C. I. G. C. Alotaibi, "Breast cancer mortality in Saudi Arabia: Modelling observed and unobserved factors," PloS one, vol. 13, no. 10, 2018.

[5] A. J. N. IntisarRizwan I Haque, "Deep learning approaches to biomedical image segmentation," Informatics in Medicine Unlocked, 2020.

[6] A. S. Komen, "Imaging Methods Used to Find Breast Cancer," Susang Komen, 2016. [Online]. Available: https://ww5.komen.org. [Accessed 2020].

[7] T. W. W. Rose, "Digital Database for Screening Mammography," Computer vision and pattern recognition, 2006. [Online]. Available: http://www.eng.usf.edu/cvprg/Mammography/Database.html. [Accessed 2020].

[8] O. C. Y. I. N. H. Y. W. a. M. N. Moghbel Mehrdad, "A review of breast boundary and pectoral muscle segmentation methods in computer-aided detection/diagnosis of breast mammography," Artificial Intelligence Review, pp. 1-46, 2019.

[9] U. r. groups, "Mammographic Image Analysis Society," 2012. [Online]. Available: https://www.mammoimage.org/databases. [Accessed 2020].

[10] D. T. A. W. K. W. a. C. T. Chris Rose, "University of South Florida Digital Mammography Home Page," University of South Florida, 2008. [Online]. Available: http://www.eng.usf.edu/cvprg/Mammography/ Database.html. [Accessed 2020].

[11] I. A. I. D. A. C. M. J. C. a. J. S. C. Inês C.Moreira, "Inbreast: toward a full-field digital mammographic database," Academic radiology, vol. 19, no. 2, pp. 236-248, 2012.

[12] A. M. A. G. López, "More about BCDR," BREAST CANCER DIGITAL REPOSITORY (BCDR, 2012. [Online]. Available: https://bcdr.eu/information/about. [Accessed 2020].

[13] Ksmith, "CBIS-DDSM," The Cancer Imaging Archive (TCIA)The Cancer Imaging Archive (TCIA), April 2019. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM. [Accessed 2020].

[14] J. M. K. V. S. A. a. M. S. Sharma Manish Kumar, "Mammogram segmentation using multi-atlas deformable registration," Computers in biology and medicine, vol. 110, pp. 244-253, 2019.

[15] A. V. M. S. Raschka, Python MAchine learning, BIRMINGHAM: Packt, 2017.

[16] A. H. S. Praylin Selva Blessy, "A comparative study on medical image segmentation methods," Applied Medical Informatics, vol. 43, no. 1, pp. 31-45, 2014.

[17] Y. B. F. P. A. P. N. K. a. D. T. Shervin Minaee, "Applied Medical Informatics Image segmentation using deep learning: A survey," arXiv preprint arXiv:2001.05566, 2020.

[18] J. D. P. a. S. I. A. P. J. Anitha, "A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms," Computer methods and programs in biomedicine, vol. 138, pp. 93-104, 2017.

[19] S. S. R. S. M. ,. M. a. R. aj, "Examination of Digital Mammogram Using Otsu's Function and Watershed Segmentation," in 2018 Fourth International Conference on Biosignals, Images and Instrumentation (ICBSII), IEEE, 2018, pp. 206--212.

[20] K. ,. Y. S. ,. S. C. ,. Z. a. B. Yin, "A robust method for segmenting pectoral muscle in mediolateral oblique (MLO) mammograms," International journal of computer assisted radiology and surgery, vol. 14, no. 2, pp. 237-248, 2019.

[21] B. V. K. A. S. G. K. Avuti Santhos Kumar, "A novel pectoral muscle segmentation from scanned mammograms using EMO algorithm," Biomedical Engineering Letters, vol. 9, no. 4, pp. 481-496, 2019.

[22] N. M. a. M. S. Bushra Mughal, "Adaptive hysteresis thresholding segmentation technique for localizing the breast masses in the curve stitching domain," International journal of medical informatics, vol. 126, 2019.

[23] R. E.-s. I. D. S. T. a. H. M. Ilhame Ait lbachir, "A survey on segmentation techniques of mammogram images," in International Symposium on Ubiquitous Networking, Springer, 2016, pp. 545-556.

[24] A. C. D. K. V. P. S. a. R. S. Ayush Shrivastava, "Automated digital mammogram segmentation using dispersed region growing and sliding window algorithm," in 2017 2nd international conference on image, vision and computing (ICIVC)2017 2nd international conference on image, vision and computing (ICIVC), IEEE, 2017, pp. 366-370.

[25] S. E. I. A. A. T. a. H. El Kaitouni, "A breast tumors segmentation and elimination of pectoral muscle based on hidden markov and region growing," Multimedia Tools and Applications, vol. 77, no. 23, 2018.

[26] "Radiomics based detection and characterization of suspicious lesions on full field digital mammograms," Sapate, Suhas , Mahajan, Abhishek, Talbar, Sanjay, Sable, Nilesh, Desai, Subhash, Thakur, and Meenakshi, vol. 163, pp. 1-20, 2018.

[27] M. a. E. M. A. a. S. M. M. a. B. M. a. A. F. a. o. Rmili, "A New Approach to the Detection of Mammogram Boundary," International Journal of Electrical and Computer Engineering, vol. 8, no. 5, 2018.

[28] A. B. N. atil Rajeshwari S, "Improved region growing segmentation for breast cancer detection: progression of optimized fuzzy classifier," International Journal of Intelligent Computing and Cybernetics, 2020.

[29] A. B. J. Shrivastava Neeraj, "Breast Tumor Detection in Digital Mammogram Based on Efficient Seed Region Growing Segmentation," IETE Journal of Research, pp. 1-13, 2020.

[30] M. Ciecholewski, "Microcalcification segmentation from mammograms: A morphological approach," Journal of digital imaging, vol. 30, no. 2, pp. 172-184, 2017.

[31] A. S. R. A. Rohana Embong, "Structuring Elements in the Watershed Algorithm for the Segmentation of Mammography Images," in TENCON 2018-2018 IEEE Region 10 Conference, IEEE, 2018, pp. 2144--2147.

[32] E. M. A. a. R. M. Saleck Moustapha Mohamed, "Semi-automatic segmentation of breast masses in mammogram images," in Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence, 2018.

[33] S. P. a. R. A. Jothilakshmi GR, "Mammogram segmentation using region based method with split and merge technique," Indian Journal of Science and Technology, vol. 9, no. 40, 2016.

[34] A. A. D. b. I. C. Chowdhary Chiranji Lal, "Segmentation of mammograms using a novel intuitionistic possibilistic fuzzy c-mean clustering algorithm," in Nature Inspired Computing, Springer, 2018, pp. 75--82.

[35] M. Y. S. a. A. M. Kamil, "Mammography Images Segmentation via Fuzzy C-mean and K-mean," International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, pp. 22-29, 2019.

[36] M. P. J. S. B. W. a. W. J. Rampun Andrik, "Fully automated breast boundary and pectoral muscle segmentation in mammogramsFully automated breast boundary and pectoral muscle segmentation in mammograms," Artificial intelligence in medicine, vol. 79, pp. 28-41, 2017.

[37] A. A. O. Akinyemi, "Atlas-based segmentation of medical images," University of Glasgow, 2011. [Online]. Available: http://theses.gla.ac.uk/2623/. [Accessed 2020].

[38] "Atlas-based image segmentation: A Survey," Pregled bibliografske jedinice broj, pp. 1-7, 2008.

[39] A. C. K. N. Soomro Shafiullah, "Robust active contours for mammogram image segmentation," 2017 IEEE International Conference on Image Processing (ICIP), pp. 2149--2153, 2017.

[40] E.-S. R. D. I. a. T. S. Lbachir Ilhame Ait, "A new mammogram preprocessing method for Computer-Aided Diagnosis systems," EEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), p. 2017, 2017.

[41] H. K. S. B. a. B. S. Hmida Marwa, "Mammographic mass segmentation using fuzzy contours," Computer methods and programs in biomedicine, vol. 164, pp. 131-142, 2018.

[42] A. J. N. IntisarRizwan I Haque, "Deep learning approaches to biomedical image segmentation," Informatics in Medicine Unlocked, vol. 18, 2020.

[43] P. Y. S. V. K. T. S. a. M. W. Puri Munish, "Artificial neural network for drug design, delivery and disposition," Academic Press, 2015.

[44] L.-L. K. M. P. J. S. B. W. W. H. O. I. G. M. G. Z. R. B. M. A. G. a. M. I. Rampun Andrik, "Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network," Medical image analysis, vol. 57, pp. 1-17, 2019.

[45] M. Y. S. W. G. Y. W. W. Q. Y. a. G. X. Wang Runze, "Multi-level nested pyramid network for mass segmentation in mammograms," Neurocomputing, vol. 363, pp. 313-320, 2019.

[46] R. H. A. R. S. A. F. P. N. S. M. M. K. S. A. A. M. A. M. P. D. a. J.-B. Singh Vivek Kumar, "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network," Expert Systems with Applications, vol. 139, 2020.

[47] M. S. Hossain, "Microc alcification Segmentation Using Modified U-net Segmentation Network from Mammogram Images," ournal of King Saud University-Computer and Information Sciences, 2019.

[48] D. M. D. G. a. M. X. Li Shuyi, "Attention dense-u-net for automatic breast mass segmentation in digital mammogram," IEEE Access, vol. 7, pp. 37-47, 2019.

[49] C. C. A. Z. T. ,. M. N. M. R. A. V. M. M. a. R. R. R. Zeiser Felipe Andre, "Segmentation of Masses on Mammograms Using Data Augmentation and Deep Learning," Journal of Digital Imaging, pp. 1-11, 2020.

Sensed-Lexicon based Approach for Identification of Similarity among Punjabi Documents

Jasleen Kaur[1]
School of Engineering
P P Savani University
Kosamba, Gujarat, India

Jatinderkumar R Saini[2]*
Symbiosis Institute of Computer Studies And Research
Symbiosis International (Deemed University)
Pune, India

Abstract—**Textual similarity among documents often leads to copyright issues. Manual measurement of similarity among documents is a time consuming infeasible activity. In this paper, we proposed a technique for measuring similarity at sensed-lexicon level for documents written in Punjabi language using Gurumukhi script. 50 Punjabi document pairs were manually collected with the help of Punjabi native writers. The proposed technique consisted of major 4 levels. Level 0 consists of data collection phase. Level 1 consists of noise removal and stop word removal sub levels. Extracted tokens were stemmed, lemmatized and synonyms were replaced based on part of speech tagging in level 2. Vector space representation corresponding to each document leads to n-gram generation of documents in level 2. Extracted n-grams were weighted based on term frequency. In level 3, string based token level similarity indexes such as Jaccard Similarity Index (JSI), Cosine Similarity Index (CSI) and Levenshtien Distance Index (LDI) were experimented with weighed tokens. In this work, Human Intelligence Task (HIT) based rating has been utilized for measuring the similarity among documents between 0-100. Results obtained from HIT based rating are compared with results obtained from the proposed technique with various combinations of pre-processing levels. Results revealed that on the basis of majority voting, combination of stop word removal with stemming and 'noun' based synonym replacement leads to the best combination with bi-gram tokens. Statistical analysis indicates strong correlation between CSI and HIT based rating.**

Keywords—Cosine Similarity Index (CSI); Jaccard Similarity Index (JSI); Levenshtien Distance Index (LDI); n-gram; Punjabi; similarity checker

I. Introduction

"It is better to fail in originality than to succeed in imitation" Herman M.

Measuring similarity between words/ terms, sentences, paragraph and document plays an important role in computational linguistics. Similarity measurement is significant component for text classification, search engine, topic modelling, text summarization, legal documents, question answer generation, information retrieval, plagiarism detection and other language related research. Similarity is associated with finding the overlapping index among two documents. This overlapping can be present at sentence level or document level. Similarity among documents can be identified at lexical level and semantic level. In lexical level, words and/or phrases are compared to identify the similarity whereas in semantic level, contextual information associated with words or phrases is extracted and used for comparison.

In general, an automatic document similarity analyzer takes two documents and generates similarity index for them. In this paper, document level similarity is identified at sensed-lexicon level. These documents are written in Punjabi language using Gurumukhi script which adds one more layer of complexity to this task. This work has potential application in plagiarism detection in Punjabi documents. India is the land of languages. Numerous languages and its dialect are being used in spoken as well as written form. Punjabi is one of them. Punjabi falls in Indo –Aryan language category. It is indicated as first language for about 130 million people and is the 10th most spoken language in the world [1-2].

A lot of research has been carried out in area of measuring similarity among documents written in foreign languages, especially English. But this area still needs to be explored in Indian languages. No work has been reported for Punjabi language.

II. Related Work

This section presents different works carried out in area of detecting similarity among documents. Indexes for finding documents similarity are broadly categorized into string based, corpus based and knowledge based measure [5]. String based algorithms perform character level or token level comparison. Corpus based methods detect similarity based on semantic information extracted from large corpus and Knowledge based methods extract semantic similarity based on information extracted from semantic network.

A. Similarity Checking Work in Foreign Languages

Researchers [6] proposed a technique for handling semantically similar words/ paraphrases in Arabic language. Open Source Arabic Corpora (OSAC) was utilized for identifying suspected documents and Word2vec was used for experimentation. Various methods such as Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), word2vec, Global Vector Representation (GloVe), and Convolutional Neural Network (CNN) were experimented for paraphrase detection. Another group of researchers [10] used a deep learning based method to detect Arabic paraphrasing. This method consists of pre-processing phase, and word2vec phase. Convolutional Neural Network was used to generate sentence vector. Authors [12] proposed two layer plagiarism

*Corresponding Author

detection method for Arabic documents. This method consists of two layer: Fingerprinting and Word embedding. Documents were weighted using different techniques such as word alignment, POS tags, and inverse document frequency. With recall of 88% and precision of 86%, this method outperformed Plagdet. Different word embedding models were experimented for capturing semantic similarity among sentences. In this work, authors proposed a model (M-MaxLSTM-CNN) for employing multiple sets of word embedding for evaluating sentence similarity. Multi-level comparisons among sentence embedding, generated by multiple word embedding, leads to sentence similarity information. Proposed technique experimented with STS Benchmark dataset and SICK dataset from SemEval and outperform all other existing methods [7].

Saptono et al. experimented with Vector Space Model (VSM) for detecting plagiarism. In this work, cosine similarity method was used to generate the rank of textual paragraphs from query as well as collection vector. Conditional probability concept was utilized to extract number of words from a paragraph. Results revealed that 54.28% average precision and 100% average recall is achieved with threshold value of 0.3 for the conditional probability and 0.2 for cosine similarity [8]. Authors introduced the project ParaPhraser.ru for collecting of Russian paraphrase corpus and organizing a Paraphrase Detection Shared Task. Different techniques were experimented for finding paraphrases among Russian language. Result revealed that traditional classifiers with linguistics features outperformed other methods [13].

B. Similarity Checking work in Indian Languages

Automatic plagiarism software Maulik was developed to check the plagiarism among Hindi documents. Approach used for detecting plagiarism is based on n-grams and comparison with repository and online documents. Input text was pre-processing using stop word removal and stemming. Different values of n were compared with cosine similarity index to find the best value of n. Accuracy reported was 96.3 which is better as compared to existing techniques [3]. Authors proposed Document Synset Matrix for Marathi (DSMM) technique for measuring among Marathi documents. In this work, proses and verses were used for experimentation. Dataset consists of 1206 proses and verses. Different problems such as sense identification of words, polysemy were handled using proposed technique. Accuracy reported was 80 which was better than existing techniques [4]. In this paper, authors presented fuzzy semantic based and Naïve Bayes model for identifying obfuscated plagiarism in English as well as Marathi Language. Semantic relatedness information was analysed based on part of speech tags and WordNet measures. Results revealed that Naïve Bayes Model performed better as compared to fuzzy method [9]. Authors proposed technique for detecting plagiarism in Urdu documents. Reordering of sentences, and inter-textual similarity among Urdu documents was handled in this work. Proposed technique was evaluated using Support vector machine (SVM) and Naïve Bayes (NB). Performance of this proposed method was better as compared to existing techniques [11]. Author proposed Deep learning based methods for handling paraphrase detection task in Indian languages. Convolutional neural network with word embedding, WordNet score and LSTM based methods were experimented [14].

III. Methodology

This section provides the detailed architecture of system for finding similarity among Punjabi documents. Fig. 1 presents the architecture of Punjabi document similarity analyzer. Punjabi document similarity analyzer consists of mainly 4 different levels. Each level (except level 0) takes input from previous level and provides some output to the next level. Working and detail about each level is as follows:

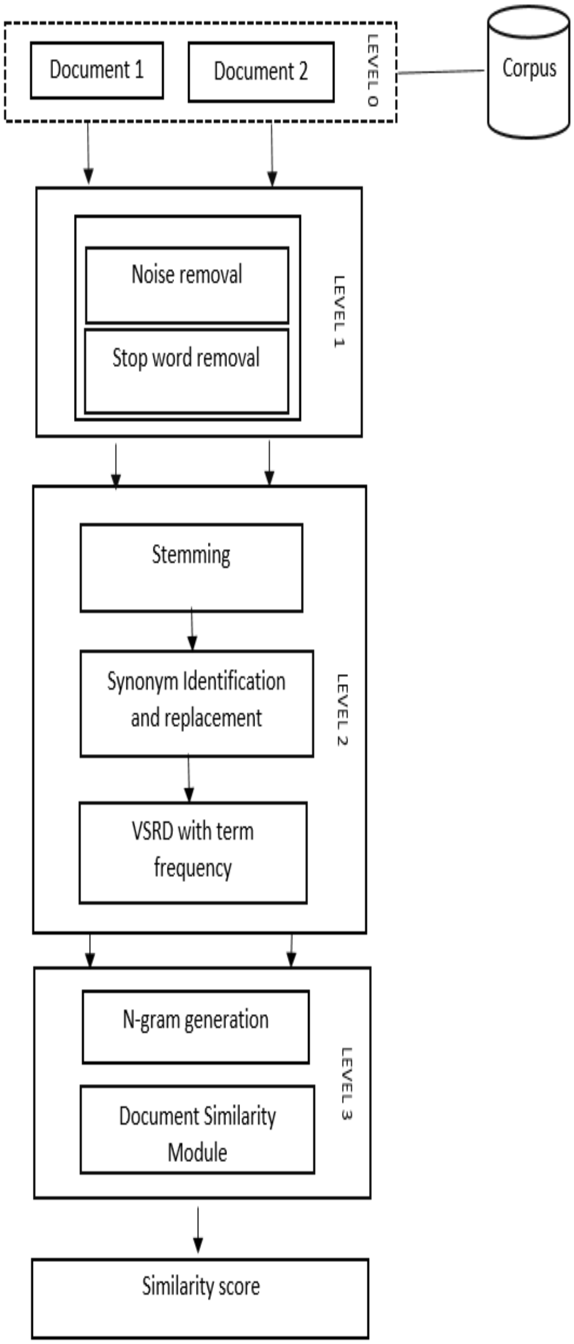


Fig. 1. Architecture of Punjabi Document Similarity Analyzer.

A. Level 0

First step for any kind of analysis is corpus. Due to unavailability of textual similarity corpus in Punjabi, similar document pairs were created. For the creation of these documents, two techniques were followed. In first one, two human annotator (Punjabi native users (writers and speakers)) were requested to write one page on a given topic. As this process was very time consuming, so internet was used as second source for generating similar document pair. Topic selection is versatile from latest topic such as corona virus to festivals of Punjab, from motivational write up to a small story, from real heroes such as Bhagat Singh to real world problems such as pollution, religious gurus to motivational thoughts. Total 50 document pairs (100 documents) written in Punjabi language using Gurumukhi script were collected for further experimentation. Out of 50 document pairs, 26 document pairs were annotated by Punjabi native user whereas, remaining 24 document pairs were generated through internet. Each document pair consists of two documents. So, these two documents D1, and D2 were passed through following phases/levels. Table I provides the statistical details about the dataset.

TABLE I. STATISTICAL DETAILS ABOUT DATASET

Sr. No.	Description	Count
1	Document pair count	50
2	Total documents	100
3	Total token count	12342
4	Unique token count	8976
5	Stop words removed	1679

B. Level 1

D1 and D2 are passed through various pre-processing sublevels. Existing similarity checker for various language (such as English) consists of comparison based on phrases and terms only. Whereas, in this proposed technique, comparison was not just based on exact phrases and but contextual information association with phrases and terms was also checked using IndoWordNet [19]. The purpose of this sub level is to reduce the noise in the input data. So various punctuation marks, symbols were removed from documents (D1 and D2). Stop words were also removed from D1 and D2 [16] [22].

C. Level 2

As mentioned earlier, in this work, document similarity is identified at sensed-lexicon level. Lexical level comprises of lexicons that are being used in both the documents. Lexical features are proven to be effective in Punjabi poetry classification work [17] [21]. Correct sense of these lexicons leads to sensed-lexicon. In next sublevel, remaining words were normalized into their root form. For word normalization, Punjabi stemming rules were used [18]. ND1 and ND 2 (normalized words from document 1 and document 2) were passed to next sublevel. Another important aspect of near copy similarity is synonym replacement. To identify the synonyms replacement among the documents, an algorithm is devised. Detailed steps are presented in algorithm 1. Effect of synonym replacement with stemming is presented separately in the results section. IndoWordNet was utilized for synonym replacement based on Part of Speech (POS) tags [19-20]. In this word, two part of speech tags ('noun' and 'verb') were experimented for identifying the synonym information from document. These normalized words (ND1, ND2) from D1 and D2 represent Vector Space Representation of both documents (VSRD1, VSRD2) [24]. With an intention to give more preference to higher occurring word in document, term frequency (TF) was used to weight the words in D1 and D2. Formula for term frequency is as follows:

$$tf(ND_i, VSRD_i) = count\ of\ ND_i\ appearing\ in\ VSRD_i$$

D. Level 3

In this level, weighted ND1 and ND2 tokens from VSRD1 and VSRD2 were divided into n-grams. Results are presented for n is equal to 1 to 5. Generated n-grams were passed to next sublevel: document similarity level. Lexical similarity between documents was identified through following techniques. Similarity of documents was generated on the basis of scale from 0-100. 0 means no overlapping between the documents and 100 means completely copied document.

Jaccard Similarity Index (JSI): This index was used to measure the similarity between two sets using the formula as given below [24]

$$J(nw - ND_1 \text{ and } \mathrm{nw} - ND_2\,) = \frac{|nw{-}ND_1\ \cap \mathrm{nw}{-}ND_2\,|}{|nw{-}ND_1\ \cup \mathrm{nw}{-}ND_2\,|} \tag{1}$$

Where $nw - ND_1$ and $\mathrm{nw} - ND_2$ represents the n-gram representation of weighted ND1 and ND2.

Cosine Similarity Index (CSI): This index was used to measure the similarity based on angle between two vectors [25] where document were represented as vectors.

$$\cos \Theta = \frac{\overrightarrow{nw{-}ND_1}.\overrightarrow{\mathrm{nw}{-}ND_2}}{\|\overrightarrow{nw{-}ND_1}\|\|\overrightarrow{\mathrm{nw}{-}ND_2}\|} \tag{2}$$

Where $nw - ND_1$ and $\mathrm{nw} - ND_2$ represents the n-gram representation of weighted ND1 and ND2

c. Levenshtien Distance index (LDI): This is edit based similarity index. Number of edits in form of insertion, deletion and substitution is calculated. Overall bounded similarity index is generated between 0 and 1 [26].

$$lev_{(nw-ND_1 \text{ and } \mathrm{nw}-ND_2\,)}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ \min \begin{cases} lev_{(nw-ND_1 \text{ and } \mathrm{nw}-ND_2\,)}(i-1,j)+1 \\ lev_{(nw-ND_1 \text{ and } \mathrm{nw}-ND_2\,)}(i,j-1)+1 \\ lev_{(nw-ND_1 \text{ and } \mathrm{nw}-ND_2\,)}(i-1,j-1)+1 \end{cases} & otherwise \end{cases} \tag{3}$$

It measures first i characters and j characters of $nw - ND_1$ and $\mathrm{nw} - ND_2$, respectively.

Implementation of this entire work was done in Python 3.7 [15]. Different packages such as nltk, inltk, sklearn were used in this work.

Algorithm I: Algorithm for finding synonyms of tokens based on part of speech associated

```
Input: Document1 (D1) and Document2 (D2)
Output: All synonyms replaced in Document2
Step 1: Both documents were tagged based Part of Speech with the help of part of speech tagger.
Step 2: Divide the document D2 into tokens a ($t_1$ …$t_n$) and form Bag of Word2 ($BOW_2$).
Step 3: Extract 'noun'/ 'verb' from document D1 and form Bag of Word1 ($BOW_1$) with tokens ($t_1$---$t_n$).
Step 4: for each token ($t_1$---$t_n$) in $BOW_1$
    If token is present in $BOW_2$
            Continue with the next token in Bag of Word1 ($BOW_1$),
    Else
            a) Find the synonyms of token using IndoWordNet and search the presence of each synonym in $BOW_2$
            b) If match found in BOW2, replace synonym matched with the original token in $BOW_2$
            c) Goto step3
Step 5: End
```

IV. Results and Analysis

The purpose of this research work was to find the most suitable similarity index for Punjabi documents. Similarity between the documents can be identified either at Lexical level or at Semantic level. In this work, similarity between Punjabi documents has been measured at lexical level (indicated with 'A' in this work) with different combination of pre-processing techniques. For finding the similarity index, document vectors of TF weighted n-grams have been used. For evaluating the system, results are presented in two sections. Section 1 consists of results by the algorithm and section 2 consists of evaluation results by human linguistic expert through HIT.

A. Results based on Algorithm

In order to find similarity index at lexical level (A), different measures (as specified in previous section) were experimented with different combinations of pre-processing techniques. These combinations have been labelled with characters a to e. Details of these measures with code are presented in Table II. It is notable that these codes have been coined by us for simplicity.

TABLE II. Coined Codes for different Combinations of Pre-Processing and Normalization Techniques

Sr. No.	Coined Codes	Details of different combinations of pre-processing and normalization techniques
1	A.a	Without any pre-processing
2	A.b	Stop words removed from documents
3	A.c	Stop words removed and tokens are stemmed
4	A.d	Stop words are removed, words are stemmed and 'noun' synonym are replaced using IndoWordNet
5	A.e	Stop words are removed, words are stemmed and 'verb' synonym are replaced using IndoWordNet

Each document pair has been evaluated using 5 combinations of pre-processing techniques (as indicated in Table II) in addition with n-gram values from 1 to 5. For a single document pair, 5x5x3 combination have been tested where 5 were the combinations, 5 n-gram values and 3 similarity indexes. In total, 50x5x5x3 combination of experiments have been performed to analyze the result where number of document pairs are 50. For each document pair, each combination from A.a to A.e was tested with value of n - gram used was 1 to 5. Result of each combination (considering only non-zero results for n-grams have been averaged. Results were analyzed based on two valid findings:

1) *Finding 1:* For more than 38 document pairs, similarity index values have been reported to be 0 for n-gram having value 4 and 5. So, these values were excluded while calculating average.

2) *Finding 2:* By averaging the n-gram results (as per finding1) obtained in each combination, best combination was selected. Although, combination A.a comes out to be the best combination in all of them. But, A.a results were ignored considering the presence of stop words and so is the maximum overlapping. Detail results are presented in the next subsection.

B. Results based on Human Intelligence Task (HIT)

For this work, each document pair was shared among 10 Punjabi language native speakers. Users selected for this research are from technical background and have sound knowledge about plagiarism and similarity. They were requested to rate the similarity between two documents on the scale of 0-100. Rating value equal to 0 or 100 was ignored considering it as outlier, and such values were not considered while calculating Average Human Intelligence Task (AHIT) rating.

C. Analysis of Similarity Indexes

For each document pair, the best combination is selected on the basis of Average Jaccard Similarity Index (AJSI), Average Cosine Similarity Index (ACSI), and Average Levenshtien Distance Index (ALDI). Table III provides the results obtained with algorithm and index value obtained with AHIT score. Values in column AHIT were averaged and rounded off to 2 decimal points.

TABLE III. RESULTS OBTAINED WITH ALGORITHM AND HIT SCORING

Sr. No.	Document Pair	Best Code	AJSI	ACSI	ALDI	AHIT
1	DP-1	A.e	0.068	0.192	0.07	0.17 0.17
		A.d	0.001	0.189	0.021	
2	DP-2	A.d	0.121	0.314	0.132	0.34
3	DP-3	A.e	0.078	0.321	0.046	0.35
4	DP-4	A.d	0.068	0.412	0.063	0.45
5	DP-5	A.d	0.023	0.342	0.021	0.33
6	DP-6	A.e	0.064	0.286	0.053	0.21
7	DP-7	A.d	0.053	0.332	0.083	0.28
8	DP-8	A.c	0.058	0.409	0.049	0.3
9	DP-9	A.c	0.055	0.234	0.038	0.19
10	DP-10	A.c	0.055	0.291	0.07	0.24
11	DP-11	A.d	0.084	0.324	0.113	0.31
		A.c	0.068	0.356	0.07	0.28
12	DP-12	A.c	0.043	0.215	0.04	0.17
		A.d	0.041	0.231	0.042	0.31
13	DP-13	A.d	0.52	0.142	0.034	0.29
14	DP-14	A.d	0.064	0.231	0.062	0.25
15	DP-15	A.c	0.014	0.45	0.023	0.27
		A.b	0.012	0.213	0.14	0.19
16	DP-16	A.d	0.06	0.256	0.071	0.17
17	DP-17	A.d	0.031	0.134	0.012	0.19
18	DP-18	A.d	0.075	0.309	0.053	0.18
19	DP-19	A.d	0.05	0.154	0.038	0.34
20	DP-20	A.d	0.054	0.254	0.041	0.27
21	DP-21	A.c	0.1	0.578	0.149	0.53
22	DP-22	A.e	0.046	0.287	0.08	0.26
23	DP-23	A.d	0.042	0.456	0.04	0.46
24	DP-24	A.d	0.09	0.422	0.078	0.44
25	DP-25	A.d	0.09	0.422	0.078	0.43
26	DP-26	A.c	0.046	0.142	0.034	0.18
27	DP-27	A.d	0.082	0.409	0.079	0.43
28	DP-28	A.e	0.119	0.219	0.123	0.17
29	DP-29	A.c	0.134	0.234	0.098	0.23
30	DP-30	A.d	0.054	0.209	0.041	0.18
31	DP-31	A.d	0.123	0.381	0.14	0.39
32	DP-32	A.d	0.057	0.19	0.069	0.19
33	DP-33	A.c	0.123	0.667	0.149	0.21
34	DP-34	A.d	0.041	0.212	0.056	0.23
35	DP-35	A.b	0.139	0.183	0.045	0.19
36	DP-36	A.c	0.062	0.267	0.068	0.17
37	DP-37	A.d	0.087	0.414	0.078	0.42
38	DP-38	A.b	0.084	0.398	0.113	0.34
39	DP-39	A.d	0.023	0.234	0.012	0.24
40	DP-40	A.e	0.058	0.177	0.049	0.21
41	DP-41	A.d	0.045	0.167	0.038	0.29
42	DP-42	A.d	0.021	0.335	0.067	0.39
43	DP-43	A.d	0.075	0.341	0.054	0.37
44	DP-44	A.e	0.074	0.47	0.0123	0.12
45	DP-45	A.d	0.021	0.127	0.049	0.21
46	DP-46	A.d	0.038	0.177	0.043	0.16
47	DP-47	A.d	0.021	0.532	0.099	0.52
48	DP-48	A.d	0.023	0.452	0.113	0.47
49	DP-49	A.c	0.033	0.145	0.043	0.12
50	DP-50	A.d	0.083	0.318	0.128	0.39

From Table III, Table IV is derived based on the frequency count of each combination. From Table IV, it can be observed that combination A.c is proven to be the best combination so-far on the basis of majority voting mechanism. Result of combination A.a is ignored as stated in finding 1. *Total value reflected in Table IV is 54 because in DP-1, DP-11, DP-12 and DP-15, two combinations comes out to be the best instead of one.

In second phase of experimentation, all the results for combination A.c were compared for checking the existence of correlation with AHIT obtained. For finding the correlation among these values, distribution of data was identified.

Distribution details were presented in Fig. 2. As it can be observed from Fig. 2, data is not normally distributed, so spearman correlation coefficient method was used for finding the correlation between values obtained by algorithm and human score [23]. Correlation strength values lies between -1 and 1. Table V presents the different strength values.

TABLE IV. FREQUENCY DISTRIBUTION FOR COMBINATIONS

Sr. No.	Combination Code	Frequency Count
1.	A.b	3
2.	A.c	12
3.	A.d	32
4.	A.e	7
Total		54*

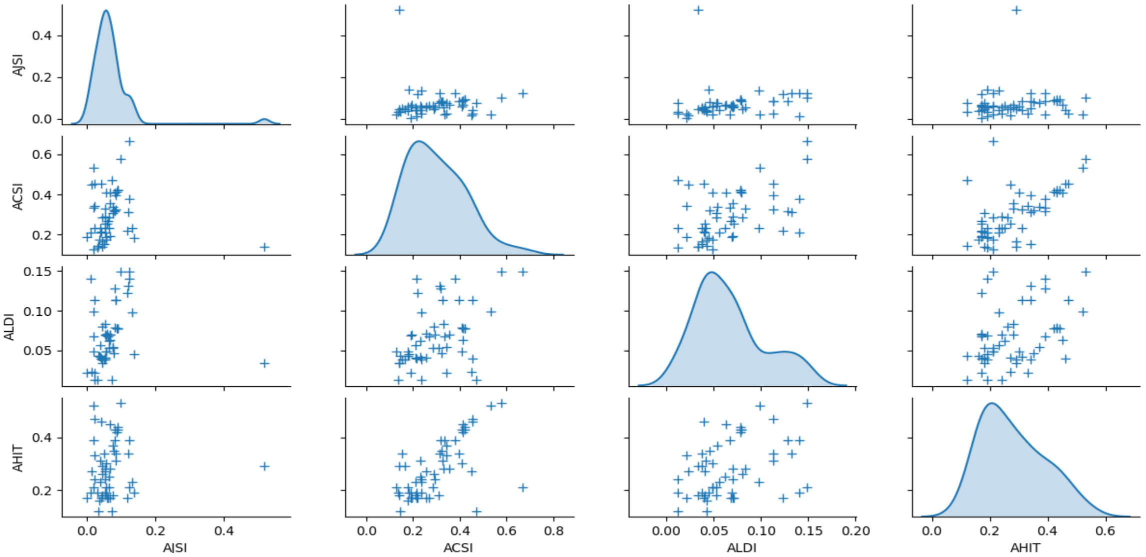


Fig. 2. Distribution of AJSI, ACSI, ALDI and AHIT.

TABLE V. STRENGTH VALUES FOR CORRELATION

Sr. No.	Coefficient Value	Interpretation
1.	0.00-0.19	Very Weak
2.	0.20-0.39	Weak
3.	0.40-0.59	Moderate
4.	0.60-0.79	Strong
5.	0.80-1.00	Very Strong

TABLE VI. COEFFICIENT SCORE

Sr. No.	Correlation Between	Coefficient Value
1.	AHIT and AJSI	0.184
2.	AHIT and ACSI	0.621
3.	AHIT and ALDI	0.351

Spearman correlation coefficient was obtained between 3 similarity index values and average HIT score. Table VI presents the coefficient values. From Fig. 3, it can be observed that highest coefficient value is 0.621 with p-value >0.05. So, AHIT score is more correlated with average cosine similarity index value. So, ACSI values obtained with algorithm has strong association with AHIT (as indicated from Table V values).

D. Analysis of n-gram

In this section, n-gram effect on similarity task is studied. For this work, value of n is taken from 1 to 5. As per assumption specified in result section, results are taken into consideration for n equal to 4 and 5. Analysis is carried out on unigram (n=1), bigram (n=2) and trigram (n=3). Table VII presents the results obtained for 50 document pairs for these n-grams.

For n-gram analysis, n-gram wise result for each combination (A.a to A.e) are averaged. Value for trigrams in document pair 4 and 10 are ignored and are not considered while calculating column average. It can be observed from the Table VII and Fig. 4 that bigram (n=2) gives the best result whereas as n is increased to 3, index values have been reduced.

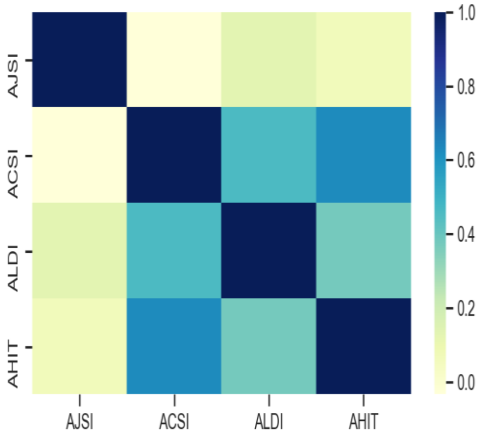


Fig. 3. Correlation Coefficient between Similarity Index Values and Average HIT.

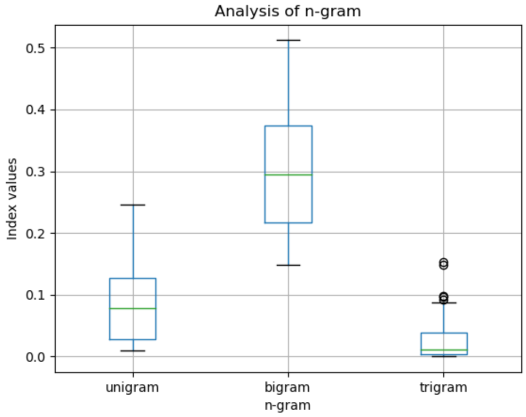


Fig. 4. Analysis of n-gram Index Values.

TABLE VII. ANALYSIS OF N-GRAM VALUES FOR 50 DOCUMENT PAIRS

Sr. No.	Document pair	Unigram (n=1)	Bigram (n=2)	Trigram (n=3)
1.	DP-1	0.127	0.343	0.01275
2.	DP-2	0.123	0.343	0.015
3.	DP-3	0.087	0.334	0.043
4.	DP-4	0.016	0.221	0
5.	DP-5	0.065	0.328	0.008
6.	DP-6	0.197	0.383	0.098
7.	DP-7	0.040	0.316	0.010
8.	DP-8	0.040	0.278	0.002
9.	DP-9	0.015	0.185	0.001
10.	DP-10	0.009	0.166	0
11.	DP-11	0.060	0.313	0.012
12.	DP-12	0.022	0.213	0.014
13.	DP-13	0.013	0.189	0.001
14.	DP-14	0.021	0.181	0.003
15.	DP-15	0.090	0.273	0.008
16.	DP-16	0.026	0.154	0.005
17.	DP-17	0.031	0.240	0.003
18.	DP-18	0.075	0.279	0.003
19.	DP-19	0.130	0.429	0.038
20.	DP-20	0.079	0.386	0.019
21.	DP-21	0.080	0.398	0.012
22.	DP-22	0.246	0.450	0.153
23.	DP-23	0.177	0.442	0.065
24.	DP-24	0.201	0.434	0.046
25.	DP-25	0.202	0.444	0.034
26.	DP-26	0.011	0.176	0.002
27.	DP-27	0.080	0.271	0.007
28.	DP-28	0.024	0.149	0.001
29.	DP-29	0.035	0.242	0.005
30.	DP-30	0.081	0.273	0.002
31.	DP-31	0.145	0.299	0.038
32.	DP-32	0.078	0.375	0.018
33.	DP-33	0.078	0.478	0.010
34.	DP-34	0.232	0.512	0.148
35.	DP-35	0.185	0.374	0.097
36.	DP-36	0.026	0.154	0.005
37.	DP-37	0.011	0.176	0.002
38.	DP-38	0.128	0.328	0.016
39.	DP-39	0.085	0.312	0.042
40.	DP-40	0.145	0.299	0.038
41.	DP-41	0.078	0.375	0.018
42.	DP-42	0.172	0.283	0.092
43.	DP-43	0.017	0.178	0.005
44.	DP-44	0.145	0.299	0.038
45.	DP-45	0.078	0.375	0.018
46.	DP-46	0.118	0.289	0.088
47.	DP-47	0.080	0.271	0.006
48.	DP-48	0.022	0.148	0.001
49.	DP-49	0.029	0.243	0.004
50	DP-50	0.029	0.165	0.006
Average		0.085	0.295	0.027

V. CONCLUSION

As the Punjabi textual content is increasing day by day on web, there is a need to check many of such documents for similarity. Manually detecting the similarity is a tedious task. So, the main objective of this work was to automate the similarity detection process. As there was unavailability of similarity textual corpus, it was created manually through human annotators. 50 document pairs were collected for further experimentation. Each document pair consists of information about the same topic. These document pairs were passed through various pre-processing techniques such as stop word removal, stemming, part of speech based synonym replacement with the help of IndoWordNet. Different combinations of these techniques were tested with n-gram with value of n from 1 to 5. JSI, CSI, LDI and HIT based rating have been used for evaluation. Results indicated that combination of pre-processing technique (stop word removal with root word conversion using stemming and synonym replacement with 'noun' based part of speech tag) proven to be the best combination so-far for detecting similarity among Punjabi documents. Out of the 3 indexes used for experimentation, values obtained for CSI are highly correlated with HIT based rating.

REFERENCES

[1] Punjabi language accessed from https://simple.wikipedia.org/wiki/Punjabi_language in Jan 2020.

[2] S. Jatinderkumar, and K. Jasleen, "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa'". Proc. Comp. Sci., vol. 167, pp. 1220-1229, March 2020.

[3] G. Urvashi, and G. Vishal, "Maulik: A Plagiarism Detection Tool for Hindi Documents" Ind. J. of Sci. and Tech., vol. 9, no.12, pp. 1-6, March 2016.

[4] B. Prafulla, and R. S. Jatinderkumar, "Marathi Document: Similarity Measurement using Semantics-based Dimension Reduction Technique" Int. J. of Adv. Comp. Sci. and App., vol. 11, no. 4, pp. 138-143, 2020.

[5] H. G. Wael, and A. F. Aly, "A Survey of Text Similarity Approaches" Int. J. of Comp. Appl., vol. 68, no. 13, pp. 13-18, 2013.

[6] A. Mahmoud, and M. Zrigui, "Similar Meaning Analysis for Original Documents Identification in Arabic Language" In International Conference on Computational Collective Intelligence, pp. 193-206. Springer, Cham, 2019.

[7] N. H. Tien, M. N. Le, Y. Tomohiro, and I. Tatsuya, "Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity" Inf. Process. & Manage., vol.56 no. 6, pp. 1-10, 2019.

[8] R. Saptono, H. Prasetyo, and A. Irawan, "Combination of Cosine similarity method and conditional probability for plagiarism detection in the thesis documents vector space model" J. of Telecomm., Elect. and Comp. Engi., vol. 10 no. (2–4), pp. 139–143, 2018.

[9] N. Shenoy, and M. A. Potey, "Semantic similarity search model for obfuscated plagiarism detection in Marathi language using Fuzzy and Naïve Bayes approaches" IOSR J. of Comp. Engi., vol.18 no.3, pp. 83–88, 2016.

[10] A. Mahmou, A. Zrigui, and M. Zrigui, "A text semantic similarity approach for Arabic paraphrase detection" In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2017. Lecture Notes in Computer Science, 10762. Springer, Cham. pp. 338-349, 2018.

[11] W. Ali, T. Ahmed, Z. Rehman, A. Rehman, and M. Slaman, "Detection of Plagiarism in Urdu Text Documents" 14th International Conference on Emerging Technologies (ICET), Islamabad, pp. 1-6, 2018, doi: 10.1109/ICET.2018.8603616.

[12] B. Nagoudi, A. Khorsi, H. Cherroun, and D. Schwab, "A two-level plagiarism detection system for Arabic document" Cyber. and Info. Tech., vol. 18 no.1, pp. 1–18, 2018.

[13] L. Pivovarova, E. Pronoza, E. Yagunova, and A. Pronoza, "ParaPhraser: Russian paraphrase Corpus and shared task" In: Filchenkov A., Pivovarova L., Žižka J. (eds) Artificial Intelligence and Natural Language. Communications in Computer and Information Science, Springer, 789, pp. 211–225, 2018.

[14] B. Rupal, S. Gargi, and S. Yashvardhan, "Deep Paraphrase Detection in Indian Languages" In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp.1152–1159, 2017.

[15] S. Bird, L. Edward and K. Ewan, Natural Language Processing with Python. O'Reilly Media Inc, 2009.

[16] K. Jasleen, and S. Jatinderkumar. Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle. In Proceeding of ACM Symposium WIR'16, 32-37. DOI: 10.1145/2909067.2909073. 2016.

[17] K. Jasleen, and S. Jatinderkumar, "Designing Punjabi Poetry classifiers using machine learning and different textual features" Int. Arab J. of Info. Tech., vol.17, no. 1, pp. 38-44, 2020.

[18] V. Gupta, "Automatic Stemming of Words for Punjabi Language," Advances in Signal Processing and Intelligent Recognition Systems, vol. 264, pp. 73-84, 2014.

[19] B. Pushpak, IndoWordNet, Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.

[20] Punjabi Part of Speech Tagger available at http://punjabipos.learnpunjabi.org/.

[21] K. Jasleen, and S. Jatinderkumar, "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms", International Conference on machine learning and computing, Singapore, February 24-26, 2017, pp. 1–5, https://doi.org/10.1145/3055635.3056589.

[22] K. Jasleen and S Jatinderkumar, "Automatic Punjabi Poetry Classification Using Machine Learning Algorithms with Reduced Feature Set" Int J. of Art. Int. and Soft comp. Inderscience Publishers. vol 5, no 4, pp 311-319. DOI: 10.1504/IJAISC.2016.10002239.

[23] M. Jerome, and W. Arnold, Research Design and Statistical Analysis (2nd ed.). Lawrence Erlbaum. pp. 508. ISBN 978-0-8058-4037-7, 2003.

[24] M. Melucci, "Vector-Space Model". In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009.

[25] L. Michael, and D. Winter, "Distance between sets" Nature, vol. 234 no.5, pp. 34–35, 1971.

[26] A. Singhal, "Modern Information Retrieval: A Brief Overview" Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24 no.4, pp. 35–43, 2001.

A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System

Majid Torabi[1]*, Nur Izura Udzir [2]*, Mohd Taufik Abdullah[3], Razali Yaakob[4]
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 Serdang
Selangor, Malaysia

***Abstract*—Intrusion detection has drawn considerable interest as researchers endeavor to produce efficient models that offer high detection accuracy. Nevertheless, the challenge remains in developing reliable and efficient Intrusion Detection System (IDS) that is capable of handling large amounts of data, with trends evolving in real-time circumstances. The design of such a system relies on the detection methods used, particularly the feature selection techniques and machine learning algorithms used. Thus motivated, this paper presents a review on feature selection and ensemble techniques used in anomaly-based IDS research. Dimensionality reduction methods are reviewed, followed by the categorization of feature selection techniques to illustrate their effectiveness on training phase and detection. Selection of the most relevant features in data has been proven to increase the efficiency of detection in terms of accuracy and computational efficiency, hence its important role in the design of an anomaly-based IDS. We then analyze and discuss a variety of IDS-based machine learning techniques with various detection models (single classifier-based or ensemble-based), to illustrate their significance and success in the intrusion detection area. Besides supervised and unsupervised learning methods in machine learning, ensemble methods combine several base models to produce one optimal predictive model and improve accuracy performance of IDS. The review consequently focuses on ensemble techniques employed in anomaly-based IDS models and illustrates how their use improves the performance of the anomaly-based IDS models. Finally, the paper laments on open issues in the area and offers research trends to be considered by researchers in designing efficient anomaly-based IDSs.**

***Keywords*—Intrusion detection system (IDS); anomaly-based IDS; feature selection (FS); ensemble**

I. Introduction

Intrusion detection system (IDS) is one of the widely used security mechanisms intended to protect computers, programs, networks, and information against intrusion, illegitimate access, alteration, or demolition. At a minimum, security systems for computers (host) and networks requires firewalls, antivirus applications and IDSs. Intrusion detection aims to detect acts performed against information systems by intruders, which attempt to gain illegitimate access to a computer asset (data, information, and network). Intruders may involve local or remote intruders: local intruders are network users with some level of legal access which attempts to elevate their levels of access through abuse of unauthorized privileges; while remote intruders refer to users who attempt to obtain illegal access to device data outside the target network [1], [2].

One of the techniques utilized to construct the intrusion detection system in order to track and deter attacks are machine learning (ML) algorithms. These techniques analyze and distinguish between from normal and abnormal packets, are attempting to avoid system harm from the attack.

A. Challenges

Among the main challenges in IDS research are related to the selection of relevant data to be investigated. For training expert machine in IDS datasets, there are some attributes which are irrelevant or may not influence the final results and also increases the execution time. The strain on expert machine is minimized by eliminating these attributes by utilizing dimensionality reduction techniques [3], [4].

The other challenge is to build an appropriate feature subset selection to be used in the process of intrusion detection, which will not only reduce the detection time but also increase the accuracy of detection. In addition, generating proper feature subset helps expert machine by avoiding over fitting issue and enhances predictive performance [5], [6].

Utilizing suitable machine learning algorithms in order to detect the intrusion is another challenge in IDS. Threats and security landscape are become increasingly complex, and strategies based on low-level machine learning are ineffective in coping with rising security issues.

There are various machine learning algorithms and methods but the main challenge is to select the one that yields optimal performance for the IDS model [7], [8].

B. Motivation

The above challenges have inspired the discussion in this review paper, which focuses on application of machine learning algorithms for feature selection and ensemble-based detection in anomaly-based IDSs. The papers are classified base on mentioned issues and addressed their objectives and attributes. This study discusses and compares the models and examines the particularities of each model in order to promote more studies in this area.

C. Previous Study

Numerous existing review articles on IDS concentrated on feature selection or detection mechanism without taking into account the future trends and open topics. Anomaly-based intrusion detection has already been studied in several review articles [9], [10]. They described various elements of IDS but did not discuss articles in-depth, the strengths and weaknesses

*Corresponding Author

in different methods of feature selection and ensemble detection for anomaly-based intrusion detection system. Furthermore, future trends and open issues are also addressed.

D. Contributions

The contributions of this review are summarized below:

- Classification of detection methods in IDS. The methods, feature selection technique used, classification type, evaluation tool and dataset are all mentioned.
- Classification of machine learning techniques used in anomaly-based IDS.
- Identification of feature selection for anomaly-based IDS, as summarized in Tables II and III.
- Identification of ensemble classification for anomaly-based IDS.
- Presentation of future directions on the state-of-the-art anomaly-based feature selection and ensemble classification.

To achieve the mentioned contributions, some research questions are ready for this analysis and the responses are given in the following sections:

RQ1.What are the detection methods utilized for IDS?

RQ2.Which evaluation tools are utilized to assess the effectiveness of the IDS?

RQ3.What are the datasets reported in the review to be used in anomaly-based IDS?

RQ4.Which feature selection methods are used for anomaly-based IDS?

RQ5.What are the machine learning algorithms used for detecting intrusions in anomaly-based detection?

RQ6.What of the ensemble techniques included in the review are reported to be used in anomaly-based IDS?

The remainder of this paper is organized as follows: Section 2 provides an overview of detection methods in IDS. Then, the taxonomy of machine learning algorithms and their methods which are employed in IDS follows in Section 3, while Section 4 reviews and compares different techniques of feature selection. Next, ensemble classification algorithms and their methods which are employed in anomaly-based IDS follows in Section 5. In Section 6, discussions on the open issues and future trends for IDSs are provided, and finally Section 7 concludes this survey.

II. Detection Methods

Intrusion detection methods are classified into four groups based on the detection method used in the system: signature-based, anomaly-based, specification-based, and hybrid. In signature-based detection the IDS identifies threats when the system or network operation matches the threat pattern (called signature) stored in the IDS local databases, and an alert will be activated. Signature-based IDSs are effective and efficient in identifying existing attacks, and their task is simple to comprehend. However, this technique has not been effective in identifying zero day attacks and new variants of previously identified attacks which are still elusive as the associated signature for these attacks [11]. Signature-based schemes offer very strong outcomes for popular, well-known threats. However, they are unable to identify new, unseen attacks, even though they are designed as minimal variants of attacks previously identified. Examples of signature-based IDSs are Artificial Immune System (AIS) [12], Collaborative Block Chained Signature-Based IDS (CBSigIDS) [13], IPFIX-based IDS (FIXIDS) [14].

Anomaly-based detection aims to predict the system's "ordinary" pattern to be covered and produce an anomaly warning whenever the difference between an immediate occurrence and normal pattern reaches a predetermined threshold. The key benefit of anomaly-based detection method is their ability to recognize previously undiscovered attack incidents. Nevertheless, in anomaly-based systems, the rate of false positives (FP), or wrongly defined as attacks is typically higher than that of signature-based method, considering the possible inaccuracy in formal signature specifications. Examples of anomaly-based IDSs are Hybridized Feature Selection Approach (HFSA) [15], Hybrid Anomaly Detection Model (HADM) [16], Unsupervised Heterogeneous Anomaly Based IDS [17].

For specification-based detection method, a human expert manually constructs the desired template which consists of a series of rules (specifications) that aim to evaluate valid behavior of a device. If the parameters are sufficiently accurate, the template may identify unlawful patterns of behavior. In addition, the false positive rate is decreased, primarily because benign behaviors that were not previously observed are not flagged as intrusions in this type of system. Specifications could also be created using some formal tool, for example, with a sequence of states and their transitions, the Finite State Machine (FSM) methodology appears suitable for modelling network protocols [18] [19]. Standard languages of representation such as LOTOS, UML and N-grammars can be considered for this reason.

Hybrid detection aims to benefit from the strengths of each intrusion detection method, minimized their weaknesses and build strong schema to detect the intrusion. A notable aspect in hybrid detection is common uses of a key signature-based detection system in conjunction with an additional anomaly-based model. This integration of the two forms of detection strategies in a "Hybrid NIDS" [20] aims to increase the final accuracy of signature-based models for intrusion detection while eliminating the usual high level of false positives of network-based IDS (NIDS) approaches, hence a hybrid approach is embraced by most existing platforms. Other examples of hybrid are Signature-Based Anomaly Detection Scheme (SADS) [21]' Artificial Bee Colony and Artificial Fish Swarm (ABC-AFS) [22], Hybrid Intrusion Detection Approach In Cloud Computing (HIDCC) [23].

Table I shows the type of detection methods utilized by researchers in IDS. RQ1, RQ2, and RQ3 are all answered in detail in the table. It specifies the detection method, the evaluation tool, dataset used in the articles, and so on. From the table it is apparent that signature-based and specification-based detection methods did not utilize feature selection and ensemble classifier to detect intrusions, in contrast to the anomaly-based detection which utilized both of them. For evaluation tools and dataset, signature-based and specification-based models were deployed and validated using simulation and real data while anomaly-based approaches were evaluated by experiments and standard IDS datasets. The NSL-KDD dataset is the most utilized dataset based on the articles in this review.

TABLE I. COMPARISON OF DIFFERENT DETECTION METHODS FOR IDS

Author	**Signature-based**	**Anomaly-based**	**Specification-based**	**Hybrid**	**Feature selection**	**Single classifier**	**Multi classifier/ Ensemble**	**Evaluation Tool**	**Dataset**
2011 [30]	×	×	√	×	×	×	×	Simulation	Real Data
2012 [31]	×	×	√	×	×	×	×	Simulation	Real Data
2013 [32]	×	√	×	×	√	√	×	Experiment	NSL-KDD
2014 [33]	√	×	×	×	×	√	×	Experiment	DARPA 1999 Real Data
2015 [15]	×	√	×	×	√	×	√	Experiment	NSL-KDD
2015 [21]	×	×	×	√	×	×	√	Experiment	DARPA 1999 ISCX 2012
2016 [34]	×	√	×	×	√	√	×	Experiment	KDD Cup 99 NSL-KDD
2017 [35]	×	×	×	√	√	×	×	Simulation	Real Data
2017 [36]	×	×	√	×	×	×	×	Simulation	Real Data
2018 [37]	√	×	×	×	×	×	×	Simulation	Real Data
2018 [22]	×	×	×	√	√	×	√	Experiment	NSL-KDD UNSW-NB15
2018 [14]	√	×	×	×	×	×	×	Simulation	Real Data
2019 [13]	√	×	×	×	×	×	×	Simulation	Real Data
2019 [38]	×	√	×	×	√	×	√	Experiment	NSL-KDD
2020 [39]	×	√	×	×	×	×	√	Experiment	CICIDS-2017 CSIC-2010v2 UNSW-NB15 NSL-KDD
2020 [40]	×	√	×	×	√	×	√	Experiment	ISCX 2012 NSL-KDD CIC-IDS2017
2020 [41]	×	×	×	√	×	×	√	Experiment	Real Data
2020 [6]	×	√	×	×	√	√	×	Experiment	KDD Cup99 NSLKDD UNSW-NB15
2020 [42]	×	√	×	×	×	×	√	Experiment	ADFA NSL-KDD
2021 [43]	×	√	×	×	√	√	×	Experiment	UNSW-NB15

III. Machine Learning in Anomaly-based IDS

Machine learning (ML) algorithms are classified into unsupervised learning and supervised learning, depending on the availability of training dataset and the successful outcome of learning algorithms. Fig. 1 illustrates the taxonomy of machine learning algorithms in anomaly-based IDS. Regarding RQ5, it has been noted that most studies focus on the following algorithms for IDS.

In supervised learning, the training function is provided with input and target output pairs, and an expert model is trained to predict the output of functions at minimal expense. Supervised learnings are classified based on learning algorithms, frameworks and objective functions. Support vector machine (SVM), decision trees, and artificial neural network (ANN) are common categorizations.

For unsupervised learning there is no tag or label in the sample dataset. Unsupervised learning algorithms are proposed to simplify the data's key features and shape clusters of natural input patterns due to a particular cost function. Hierarchical clustering, K-means clustering, and self-organization map are the most common unsupervised learning methods. One of the challenges of unsupervised training is that it is hard to evaluate because it does not have a specific educator and therefore does not have labelled test data.

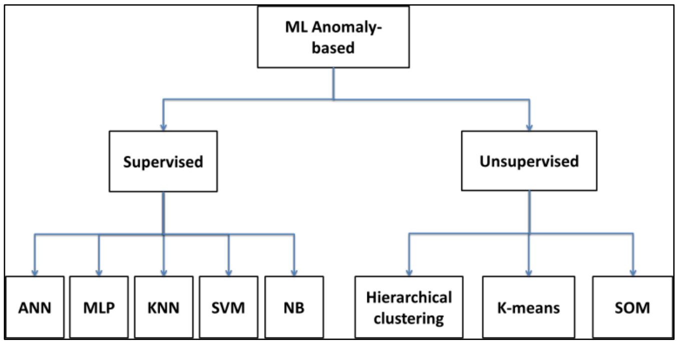


Fig. 1. Taxonomy of Machine Learning Algorithms in Anomaly-based IDS.

A. Supervised learning

1) Artificial Neural Network (ANN): ANN is one of the major algorithms of machine learning which is widely utilized as a detector operator in IDSs in many studies. ANN is used to solve a variety of issues faced by other existing intrusion detection approaches and has been suggested as a substitute for the statistical analysis aspect of detection of anomaly schemes. Initially, the ANN acquires its expertise by training the machine to properly detect pre-selected examples of problems. The neural network result will be checked and the machine configuration will be optimized until the training data neural network response reaches a sufficient level. Besides the initial training phase, the neural network often gains expertise over time as it performs review of the problem-related data [24], [25] . A hypervisor layer anomaly detection system called Hypervisor Detector that employs a combination algorithm that is a hybrid of Fuzzy C-Means clustering algorithm and Artificial Neural Network (FCM-ANN) was introduced to enhance the detection system accuracy [26]. The KDD Cup 99 sample dataset is used to test the design system to test for the reliability of five attack forms. The model was good at finding normal and probe attacks, but did not yield good results for DOS (99.96–5.33), U2R (96.78–3.22) and R2L (93.73–6.27) attacks, even for accuracy and false alarm rate. A reasonable solution using ANN in hierarchical anomaly-based IDS can be pointed to [27], which used neural Self Organization Map (SOM) networks to identify and distinguish normal packet from the attack traffic. The proposed machine was used to configure, train and evaluate the SOM Neural Network for intrusion detection. Detection output was performed to evaluate the SOM efficiency in detecting anomaly intrusion and the findings show that SOM with the KDD Cup 99 dataset can distinguish attack packet from normal one at 92.37%, while with NSL-KDD the detection rate is 75.49%.

The work in [28] tackles detection problems by presenting a simple ANN-based IDS system, utilizing back propagation and feed forward algorithms together with different other optimization methods to minimize the total computing overhead while maintaining a high level of performance. Results of the experiment on the NSL-KDD benchmark dataset showed that the quality of the proposed ANN (accuracy and detection speed) was 98.86% for accuracy and 95.77% for detection rate. An effective method to identify brute force attack in the Secure Shell (SSH) was proposed by [29]. A brute force attack is performed by the implementation into the private cloud of a client-server SSH model and the server captures traffic related to attack and normal. Next, ANN's Multi-Layer Perceptron model extracts indicative traffic characteristics and uses them to distinguish the attack and normal packets. Results obtained from this approach indicate that the suggested framework is able to detect the attack successfully with great accuracy and minimal false alarm.

2) Multi-layer Perceptron Neural Network (MLP): MLP is a supervised learning classifier which utilizes back propagation algorithm in the learning phase to train the model. It can learn a non-linear approximate function for both regression and classification task, by providing a group of features and a target in which one or more non-linear layers called hidden layers between the inputs and outputs are distinguished from logistic regressions [44].

The MLP neurons are positioned in layers with always-flowing outputs toward the output layer, either one layer (called a perceptron) or a multilayer perceptron, if multiple layers exist [45], where every neuron in a single layer has direct connections to the subsequent layer's neurons. The units of those networks apply a sigmoid function as activation function in many applications.

A wrapper-based feature selection is designed by utilizing the Discernibility Function as algorithm for search to construct subsets of feature and the MLP classifier is used to determine the subsets of features. Thus, the C4.5 decision tree and the MLP classifier, which are commonly utilized in the IDS, are used to illustrate better classification rates. With this hybrid method, the findings for the KDD Cup 99 shows improved accuracy of approximately 12% for U2R, 2% for Probe, and 1% for DOS classes [46].

To build effective IDS, a hybrid multi-layer perceptron (MLP) and Artificial Bee Colony (ABC) algorithm were designed. The MLP classifier was used to distinguish among the attack and normal traffic in network. Training and testing have been conducted using the NSL-KDD dataset. Results of the experiments show that the suggested solution gives a high detection rate of about 87.54 % and error rate of 0.124% [47].

3) K-nearest neighbor (KNN): KNN algorithm is a nonparametric technique for classification and is a simple and straightforward machine learning algorithm. It is vast used based on many experiments reported on intrusion detection, pattern recognition, text categorization and countless others [48].

A combination of the Learning Vector Quantization ANN and KNN method for intrusion detection was suggested by [49]. The analysis was performed on the NSL-KDD dataset and the proposed model has a detection rate of 97.2% (five classes) with a false alarm rate of approximately 1%.

4) Support Vector Machines (SVM): SVM is one of the algorithms in machine learning that used labeled instance (packet) to train the model and differentiate the packet to different classes by generating templates that could determine which class a new instance belongs into [50], [51]. SVM's main objective is to discover a linear optimized hyper plane that maximizes the isolation boundary between groups. The SVM then trains the model across sections or portions of the data[52].

A hybrid intrusion detection KPCASVM with GAs design was proposed [53], where KPCA is implemented in the N-KPCA-GA-SVM system to obtain the key data features of intrusion detection, and a multi-layer SVM classifier is used to determine normal or attack behavior. The test was conducted on the KDD Cup 99 dataset and the detection rate was 96%. BIRCH hierarchical clustering SVM-based network intrusion detection framework [54] was proposed for pre-processing of data. Instead of the original large data set, the BIRCH hierarchical clustering could provide the SVM learning with highly qualified, abstracted and reduced data sets. The proposed solution could achieve a 95.72% accuracy with a false positive rate of 0.7% overall, but was not satisfactory with the division accuracy for each attack type (Prob= 97.55%, U2R=19.73% and R2L=28.81%).

A new Combining Support Vectors with Ant Colony (CSVAC) algorithm was proposed to produce cluster classifiers in intrusion detection [55] using two existing machine learning techniques (SVM and CSOACN) to improve overall detection rates and speed. The method is applied and tested using the standard KDD Cup 99 dataset benchmark, and yields a classification rate of 94.86% with a false negative ratio of about 1% and the false-positive ratio of 6.01%.

5) Naive Bayes Network (NB): Naive Bayes (NB) is a simple method of creating classifiers that allocate labels of class to problematic cases identified as values of feature vectors, where class tags are drawn from a restricted set. There is no single algorithm for learning such classifiers but a set of algorithms based on a common concept. A Directed Acyclic Graph (DAG) usually describes the structure of an NB, that each node represents a process variables and each reference encodes one node's control over another [56]. By comparing the decision tree and Bayesian techniques, the decision tree's accuracy is much higher but the processing time of the Bayesian network is low [57]. Therefore, it will be effective to use NB models when the dataset is very large.

A Naive Bayes-based IDS which obtained better findings than neural network IDS while tested on the KDD Cup 99 was proposed by [58]. The average accuracy obtained by utilizing Naive Bayes was 91.52%. While being basic in design, it can produce accurate results. A hybrid intrusion detection system based on Naive Bayes and decision tree was proposed by [59]. The model has been compared and tested using benchmark KDD Cup 99 dataset, the detection rate was 99.63%. A Fuzzy Intrusion Recognition Engine (FIRE) Intrusion Detection System Simple data mining approaches used to process network stimulus data packets and reveal essential anomaly detection indicators was developed by [60]. Such indicators were accessed for each observed value and used afterwards to classify network attacks. An intrusion detection model with information gain for feature selection and SimpleCart algorithm to detect the intrusion was suggested by [61]. First, the features were reduced to 33 and then the SimpleCart algorithm used for detection. The model was applied on NSL-KDD dataset and the detection accuracy was 82.32% and error rate was 17.67%. A hybrid strategy to learning is suggested by integrating Naive Bayes and K-Means clustering classifier. The suggested solution has been compared and tested using the benchmark dataset KDD Cup 99. These combinations learning methodology achieved rather low error rates with an average of less than 0.5% while retaining accuracy and detection rates above 99%. The method is capable of accurately classifying all data except the U2R and R2L attacks. to overcome this limitation, it was recommended to consider the Integrated Intrusion Detection Program which is ideal for identifying R2L and U2R threats [62], [63]. In SSH traffic, a combination of Bayesian Network and Genetic Algorithm was introduced to improve identification of brute force attacks [64]. The proposed method implements brute force attack data obtained in a client-server model. Their findings show that the most effective features were chosen and the final result was better than the benchmark.

B. Unsupervised learning

Unsupervised detection of anomalies (often recognized as outlier detection) employs clustering approaches to classify potentially malicious incidents without previous knowledge in a dataset. Clustering aims to divide a limited unlabeled data into a discrete and finite collection of "natural" unseen structures of data instead of providing a precise non-observed characterization incidents produced within the same distribution probability [65]. In another aspect, the goal of unsupervised algorithms is to divide the data into categories (clusters) that reach great similar internal and external dissimilarities without previous knowledge.

All clustering approaches are based on the following hypotheses for this reason. First, the volume of normal instances in a database surpasses the volume of anomalies. Next, the anomaly packet themselves vary from normal instances qualitatively [66]. Scores are allocated to the installed clusters after the cluster formation. If a cluster's score reaches the threshold pre-defined or automatically determined, a potential anomaly is considered. When clustering is utilized to identify attacks on the network, respectively, one believes that malicious traffic is less than the normal packet and normal packet is distinguished from the malicious one in some way. In other words, the features that characterize the attacks well enough to be defined must be selected concerning to the process of detection. The aim of clustering is to categorize network packets or flows without prior knowledge, but based solely on their relationships. As a result, large normal packet clusters would be formed when attack packets produce small clusters and cases not belonging to other groups. A static or dynamic threshold may be utilized to determine that clusters are deemed to be attack based on the testing and algorithm adjustment used. The main benefit of clustering models is their capability to identify unseen threats without previous information, thereby eliminating the need for labeled traffic. The main disadvantage is their high false-positive rate.

The extraction or selection of features is among the most critical stages of unsupervised detection. The use of clustering techniques to identify a range of attacks by checking alarm records from heterogeneous database was proposed [17], instead of utilizing the attributes of abnormalities that carry specific actions to suit instances or the standard approach of testing and training currently utilized in abnormal detection. Even though it required less time for the three clustering algorithms tested in the system to forecast and build clusters, the clusters' accuracy produced by one algorithm was not consistent across various logs and subsets. The obtained result indicates the way or route to develop abnormal detectors that could use pure activity logs obtained from heterogeneous databases on the tracked network and compare instances through alarm records to identify intrusion.

IV. Feature Selection Techniques

Feature Selection (FS) is a method for removing unnecessary and redundant features and choosing the most suitable feature subset that will result in a better classification of patterns which belong to various classes of attack. So, from researchers' view there are reasons why feature selection needs to be performed:

1) A single selection strategy is not adequate to obtain consistency across multiple datasets, as network traffic activity is changing [67]–[69].

2) An appropriate subset for each attack types should be identified, since one general subset of features is insufficient to properly represent all the various attacks[69]–[71].

3) FS can significantly improve not only the accuracy of detection but also the computational efficiency, where:

a) features which are irrelevant or redundant can result in poor detection rate and overfitting, therefore, reducing them can increase the detection accuracy; and

b) more features for each data point would cause higher computational costs and complexity—reducing irrelevant features will increase the computational efficiency [67], [69]–[74].

4) Ultimately, R2L (Remote-to-Local) and U2R (User-to-Root) attack groups are known to become the most challenging to identify since they are too isolated and could be mislabeled as normal packet. Studies and experiments have shown that FS can solve this issue by defining a feature subset adapted to the behavior of each attack type classes [70], [71], [74].

Methods of FS are generally classified into filter, wrapper and optimization-based FS methods for selecting features. Table II illustrates the advantages and disadvantages of the mentioned features selection methods and Table III summarized the reviewed feature selection for anomaly-based IDS. RQ3 and RQ4 are all answered in detail in the table. It specifies the feature selection methods, the algorithm's origin, subset size, strength, weakness, dataset used in the articles, and so on.

TABLE II. Comparison of different Feature Selection Methods

Method	**Advantages**	**Disadvantages**	**Examples**
filter	• Faster than wrapper • Not dependent on classifier • Less computational complexity than wrapper • Less over-fitting issues • Use statistical methods for evaluation of the attributes	• Lack of interaction to classifiers • Lack of dependency among attributes • Less detection rate compared to wrapper	Euclidean distance, information gain, correlation-based, etc.
Wrapper	• Interact with classifier • Consider attributes dependency • Better detection rate • Use cross validation for evaluation attributes	• Longer execution time • More risky for over-fitting issues	Sequential forward selection, Sequential backward selection, Hill climbing, Stepwise selection, etc.
Optimization-based	• Interact with classifier • Less over-fitting issues • Better detection of global optima • Better attribute selection • Simple to implement	• Difficult to be adjusted to a new situation • Complexity to adjusted different parameters	Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Cuckoo Search Algorithm (CSA), Genetic Algorithm (GA), etc.

TABLE III. SUMMARY OF THE REVIEWED FEATURE SELECTION FOR ANOMALY-BASED INTRUSION DETECTION SYSTEM

Author	**Method**	**Algorithm**	**Type of feature subset**	**Subset size**	**Strengths**	**Weaknesses**	**Dataset**
2011 [79]	Filter	Mutual information-based feature selection	Class-based Subset	15 out of 41	Improve relevancy and reduce redundancy for feature selection.	Performance metric is low.	KDD Cup 99
2013 [32]	Wrapper	Wrapper based on Bayesian Network classifier	Single Subset	11 out of 41	Improve the performance metric.	Accuracy of U2R attack was not satisfactory.	NSL-KDD
2015 [102]	Wrapper	Layered wrapper feature selection approach (LAWRA)	Single Subset	16 out of 41	LAWRA utilized external cluster validity indices, F-measure, and Fowlkes–Mallows index, for feature selection.	Overall accuracy (around 83%) and false parameter was not satisfactory. Lack of class-based feature selection.	NSL-KDD
2016 [34]	Optimization-based	Ant Colony Optimization (ACO) and K-Nearest Neighbor (KNN)	Class-based Subset	4 to 8 out of 41	Needs no prior knowledge of features.	False positive was not satisfactory (2.59).	KDD Cup 99 NSL-KDD
2016 [103]	Filter	Flexible Mutual Information Feature Selection (FMIFS) and Least Square SVM	Single Subset exempt Class-based Subset for KDD Cup 99	KDD Cup 99 (12 to 23 features) NSL-KDD (18 features) Kyoto2006+ (-)	Proved the generality of their model by utilizing different datasets.	Accuracy of U2R attack was not satisfactory. Lack of class-based feature selection two dataset.	KDD Cup 99 NSL-KDD Kyoto2006+
2017 [92]	Optimization-based	Hypergraph based Genetic Algorithm (HG-GA)	Single Subset	35 out of 41	HG-GA utilized a weighted objective function to improve performance metrics.	Overall accuracy 97.14%. Lack of class-based feature selection.	NSL-KDD
2019 [104]	Filter/ Optimization-based	Linear correlation coefficient algorithm and cuttlefish algorithm (CFA) and Decision tree	Single Subset	10 out of 41	Integration of filter method with cuttlefish algorithm optimization helps the model to detect the attack with less false alarm.	Overall accuracy 95.03%. Lack of class-based feature selection.	KDD Cup 99
2019 [38]	Filter/Wrapper	CART Algorithm	Single Subset	17 out of 41	Utilizing Gini and CART algorithm.	Overall accuracy 79.7%. Lack of class-based feature selection.	NSL-KDD
2020 [105]	Wrapper / Optimization-based	Genetic Algorithm with Multi-parent Crossover and Multi-Parent Mutation (MGA)	Single Subset	4 out of 41	Propose a new operator, called multi-parent-crossover-mutation to enhance the GA performance.	Lack of class-based feature selection.	NSL-KDD
2020 [94]	Optimization-based	Mutation Cuckoo Fuzzy (MCF) and ANN classifier	Single Subset	22 out of 41	Integrates mutation operator with cuckoo search and Fuzzy C Means (FCM).	Lack of class-based feature selection.	NSL-KDD
2020 [106]	Optimization-based	Pigeon Inspired Optimizer	Single Subset	KDDCUPP99 (7 features), NSL-KDD (5 features), and UNSW-NB15 (5 features)	Proved the generality of their model by utilizing different datasets.	Performance metric was not satisfactory. Lack of class-based feature selection.	KDD CUPP 99 NSL-KDD UNSW-NB15
2020 [40]	Filter/ Optimization-based	Ensemble of (mRMR, JMI CMIM) and Chaotic Adaptive Grasshopper Optimization Algorithm (CAGOA)	Single Subset	ISCX 2012 (20 features), NSL-KDD (19 features) and CIC-IDS2017 (12 features)	This feature selection combination gives good accuracy and less false alarm.	Lack of class-based feature selection.	ISCX 2012 NSL-KDD CIC-IDS2017
2020 [107]	Optimization-based	Multi-objective method (NSGAII) and ANN	Single Subset	NSL-KDD (24 features) UNSW-NB15 (19 features)	Proved the generality of their model by utilizing different datasets.	Performance metric was not satisfactory. Lack of class-based feature selection.	NSL-KDD UNSW-NB15

2020 [6]	Wrapper / Optimization-based	Hybrid of Fruit Fly Algorithm (FFA) and Ant Lion Optimizer (ALO) Algorithm.	Single Subset	KDD Cup99 (12 features), NSLKDD (16 features), UNSW-NB15 (15 features)	This hybrid algorithms increase the diversity of populations, which yields better detection.	Lack of class-based feature selection.	KDD Cup99 NSLKDD UNSW-NB15
2020 [108]	Optimization-based	Many Objective Evolutionary Algorithm and Artificial Bee Colony (MaOEA-ABC)	Single Subset	11 out of 41	Propose an adaptive selection probability approach that will adjust the selection probability and enhance the algorithm's ability to find the best solution.	Lack of class-based feature selection.	NSLKDD
2020 [43]	Wrapper / Optimization-based	Tabu Search and Random Forest (TS-RF)	Single Subset	16 features out of 41	Reducing feature vector by more than 60%. This reduces computational complexity of the proposed solution.	Lack of solving class imbalance problem present in UNSW-NB15. Performance metric was not satisfactory.	UNSW-NB15

According to the reviewed articles in Table III and the result from Fig. 2, it shows that optimization-based methods were mostly utilized for feature selection in the recent years. This method has undergone a significant improvement in terms of feature numbers. Based on the review, NSL-KDD dataset was mostly used by researchers to prove their models. In addition, some research utilized different datasets to highlight the generality of their solutions, like Kyoto2006+, ISCX 2012, UNSW-NB15, and CIC-IDS2017.

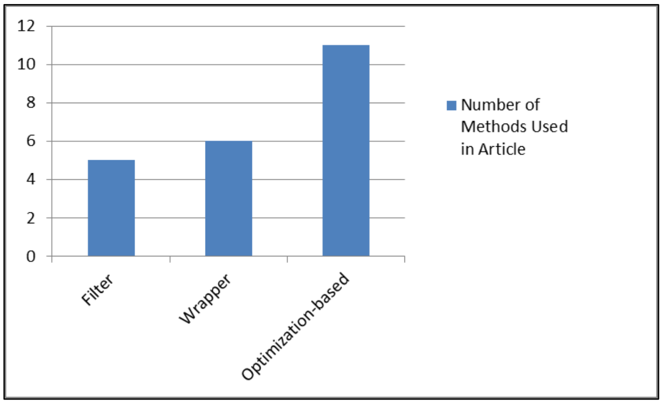


Fig. 2. Number of Studied Feature Selection Methods.

5) Filter: Filter methods use different information theory and mathematical formula for feature selection. Due to their simplicity, ranking methods are used and had good performance for practical applications. The rating of variables is based on an accepted ranking criterion and the threshold is being utilized to eliminate variables just below the value of threshold. The methods of ranking are filtering approaches as less relevant variables are extracted before the classification. A fundamental characteristic of a distinctive feature is the provision of useful information on the various classes of data. This characteristic could be described as a feature relevance[75] which defines a measure of the efficacy of the feature in order to distinguish among various classes. There are different ways to calculate a feature's relevance to the data point or outcome. Different publications [75]–[77] proposed different understandings and measurements for a variable's importance and relevancy. One description that can be listed that would be valuable is ‘‘A feature can be regarded as irrelevant if it is conditionally independent of the class labels.’’ [78]. This clearly stipulates that the data could be distinct but not separate from labels of the class, if the feature is to be relevant. The feature that does not impact class labels can be omitted. As noted above, for assessing specific features, correlation of features plays a key role. The underlying distribution of practical uses is unclear: it is calculated by the accuracy of the classifier. Because of this, an ideal subset of features may not be special because using different feature sets it may be possible to reach the similar accuracy of classifier. An improved feature selection algorithm has been proposed [79] to efficiently classify the attacks behaviors by measuring mutual information. Correlation can also be extended to evaluate of the efficiency of a feature subset, where a subset of features is perfect if the correlation among the classification and the feature subset is significant, but the correlation among the specific feature and the other features within the subset of features is poor. In addition, distance calculation for the selection of features can also be utilized [80]. Widely utilized distance calculation includes Euclidean distance, Martensitic distance and standardized Euclidean distance.

6) Wrapper: Wrapper feature selection use machine learning as a fitness function and determine the best feature subset across all subsets of features. This problem formulation allows generic optimization techniques to be used with the machine learning to rank subsets of feature based on their prediction. Therefore, in the aspect of a machine learning final predictive accuracy, the wrapper method typically surpasses the filter approach. The wrapper technique was widely popularized by [75], and provides an easy but efficient way to tackle the issue of selection of features. However, the wrapper method incurs more computation cost and need more execution time compared to filtering methods. A feature selection method using machine learning algorithms was proposed [81] for efficient intrusion detection, which blends the characteristics of distributed denial of service (DDoS) characteristic-based features (DCF) and consistency set evaluation (CSE). To identify the most relevant features, the NSL-KDD dataset is utilized as an attack dataset and is built on a few selections of feature methods, along with consistency-based evaluation of subsets and DDoS characteristic-based features (DCF). The experimental result shows that their proposed system has greater accuracy and efficiency compared to other approaches.

7) Optimization-based methods: Classic wrapper and filter strategies are independently evaluated and subset chosen. However, some features are not independent, but they are really successful when they work together. Therefore, the classic strategies in this respect are not very successful. Metaheuristic-based methods were already used to select and classify the selected features as a result of its vast improvement capability of in detection [82], [83]. Examples of optimization-based methods are Particle Swarm Optimization (PSO) [84]–[86] entropy of network features [87], Genetic Algorithm [88], [89], ant colony optimization [34], [90] and Kernel Principal Component Analysis (KPCA) [91]. With the increase in the dataset dimension, the space of the problem of selection of feature rises significantly. This leads to a large solution space with additional features. Furthermore, in a wide solution space, a huge proportion of duplicate or uncorrelated features generate several local optima.

A new anomaly based detection model of Hypergraph based Genetic Algorithm (HG - GA) was proposed by [92]. The Hypergraph's attribute was used to generate initial population in order to speed up the quest for the optimum solution and avoid trapping at local minima. HG-GA utilized a weighted objective function to achieve the balance among maximum detection rate and reducing false positive, as well as reducing features number. HG-GA SVM performance was assessed by NSL-KDD dataset.

An Ant Colony Optimization (ACO) for selection of feature method was proposed [34] using K-Nearest Neighbor (KNN) for the classification process and the accuracy was utilized as the assessment function for the model. The studies were performed using the KDD Cup 99 dataset, giving 98.9 % for accuracy and 2.59% for false positive rate.

A learning model for fast learning network (FLN) based on PSO was proposed by [93]. The PSO-based optimized FLN was trained using particle swarm optimization to pick weights. For evaluation, the research utilized KDD Cup 99 dataset to explore the effects of PSO-FLN model. The findings indicated that the model had good impact on intrusion detection.

An enhancement of Cuckoo Search Algorithm (CSA), named Mutation Cuckoo Fuzzy (MCF) was proposed by [94] for feature selection method and multiverse optimization ANN for classification at anomaly-based IDS. For feature selection phase, MCF that integrates mutation operator with cuckoo search and Fuzzy C Means (FCM) clustering was utilized. Through this method, the cuckoo Search efficiency to detect the optimal features was increased. The proposed feature selection choses 22 out of 41 features and for evaluation part well known dataset, called NSL-KDD was used to illustrate the effectiveness of their anomaly-based IDS.

A. *Limitations of the Related Works*

After analyzing the data collected from the literature related to feature selection, some limitations and shortcomings of the works are identified:

1) The optimal detection methods or strategies for various datasets have yet to be established.

2) There is a lack of proper feature subset to train faster with minimal computation and optimal performance in detecting intrusion with high accuracy and less false alarms.

V. Ensemble

The idea of merging results from a collection of learners into one is known as ensemble [95]. To obtain reliable and more accurate predictions, an ensemble can integrate multiple learners. It is possible to use a variety of techniques to generate and incorporate learners. Various datasets could be utilized to train the same training frameworks or the similar dataset could be utilized to learn various frameworks [96]. The biggest issue on the learning of the ensemble is to choose the algorithms that construct the ensemble and the function of decision or fusion that incorporates these algorithms' results. Of course, it is easy to use more algorithms to enhance the fusion results, but bearing in mind the computing cost of adding a new algorithm, it needs careful consideration. Dietterich [95] offered three key explanations for the use of an ensemble-based system. First, the empirical justification is related to the absence of sufficient knowledge to accurately classify the quest space's best hypothesis. Second, the computational description is to resolve the issue that most machine learning methods might be trapped in the local optima when looking for the perfect solution. Finally, the rationale for representation is to resolve the problem of the failure of several machine learning methods to accurately depict the border of the searched decision. Creating an ensemble takes two main parts: creating and combining [97].

The creation process has to construct a collection of base classifiers. The decision on how to integrate the results of the base classifiers into one is taken in the combining process. Many of the well-known modern ML algorithms were constructed around the idea of the ensemble. The three widely used ensemble model are bagging, boosting, and stacking [98]. Such techniques combine various models of learning into a single model so that bagging (variance), boosting (bias) or stacking (predictions) can be minimized. Fig. 3 demonstrates the general design methods of the ensemble.

A. *Bagging*

Among the first ensemble algorithms, one of the simplest and easiest way to accomplish a better efficacy was bagging [99]. When bootstrapped copies were used, varieties of results are generated in bagging, which is to say, various data subsets are randomly selected from the complete dataset of training. A different same type of classifier is designed by utilizing the learning data portion. Using a majority vote on their lists, the fusion of different classifiers is accomplished. Therefore, the decision of the ensemble is the category chosen by the largest number of classifiers for any instance data.

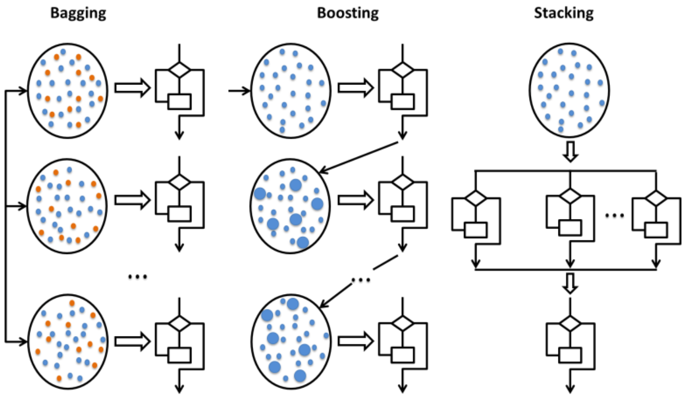


Fig. 3. Three General Ensemble Designs.

Random Forests is a method which is produced from bagging [100]. Training of multiple decision trees and randomly changing parameters relevant to training is a way to create this sort of classifier. As in bagging, copies of the training data could be bootstrapped from those parameters; but, unlike bagging, they can also be unique subsets of features, which is the case in the random subspace process.

From bagging, another method was generated, named "pasting small votes." It was a technique designed to run on huge datasets, unlike bagging [101]. Large size datasets are divided to the small size portions called "bites," used for learning various classifiers.

Small votes have resulted in the design of two combinations: first, named as Rvotes, randomly produces the subsets of data; second, named as Ivotes, creates consecutive datasets, taking into account the importance of the instances. Ivotes has been shown to deliver better results similar to the approach in boosting methods where the classifier advises the most suitable instances for the ensemble component used [109].

New method of ensemble classification [110] are proposed using bagging classifiers and their performance is evaluated with accuracy in mind. A classifier ensemble is built as a base classifier using the Support Vector Machine (SVM) and Radial Basis Function (RBF). The effectiveness and advantages of the approaches proposed are demonstrated through NSL-KDD datasets. The accuracy for bagged RBF was 86.40% and bagged SVM was 93.92%.

B. Boosting

In 1990, Schapire [111] demonstrated a weak learner (algorithm) which produces classifiers that can moderately surpass random guessing, can be converted into a powerful learner that can properly classify all instances except an extremely small fraction. The boosting created a group of classifiers by resampling the data and integrating results by majority voting. Re-sampling in boosting is designed to provide the most detailed training data for successive classifiers. In general three classifiers are created by boosting: a randomized subset of available training data is utilized to construct the first one. For training of the second classifier, knowledgeable subset of data provided to the first classifier is utilized where the knowledgeable data portion includes instances of training dataset, so the first classifier correctly identified half of them and the other half was misidentified. Ultimately, learning information for the third classifier is made up of cases where there was a conflict between the first and second classifiers. The results of the three classifiers would be combined with a majority vote.

A simplified edition of the initial boosting algorithm called "adaptive boosting" or "AdaBoost" was proposed in 1997 by Freund and Schapire [112]. Two algorithms of this group, AdaBoostM1 and AdaBoostR are the most commonly used variants, as they are perfect to cope for problems of regression and multiclass. AdaBoost generates some assumptions and the same assumptions apply to aggregate decisions by weighted majority voting of the groups decided. By extracting instances from a successively updated distribution of training data, a weak classifier is trained to build the assumptions. Updating the distribution ensures that the following classifier examples that were incorrectly identified by the prior classifier are return back to dataset to train other classifiers. Therefore, training data from various classifiers continue to move into instances that are becoming increasingly difficult to classify.

C. Stacking

Many cases are very likely to be miscategorized because they may happen to be in the near neighboring of the decision line and thus are typically located on the incorrect side of the line identified by the machine learning classifier. On the other hand, since it is on the right side and far from the boundaries of the appropriate decision, there may be instances that are likely to be well defined. If a group of classifiers performs with a dataset from an undefined source, could we create a relationship among the classifiers' results and correctly detect groups? The concept motivating generalization of Wolpert's is that the results of a classifying ensemble serve as sources to the next meta-classifier at second level with the goal of learning the manner in which the ensemble's findings are related to the correct label instances [113].

Stacking is the term used for Stacked Generalization [113], which is to find the ideal composition of a base learner set. Stacking is an algorithm class that requires training a "meta-learner" second level to find the combination. Stacking aims to combine solid, different sets of learners, unlike bagging and boosting. Besides, ensemble methods such as boosting and bagging are often utilized to construct alike ensembles, while stacking could be utilized to create diverse ensembles.

D. Other Work

New ensemble methods [114] proposed are Net-GR based ANN-Bayesian approach that implies ensemble of Bayesian Net with Gain Ratio (GR) feature selection approach and ANN. They have applied a variety of single classification methods and their proposed ensemble on NSL-KDD and KDD Cup 99 datasets to evaluate for model's robustness. With 29 features which were selected, a 97.78% and 99.38% accuracy detection were achieved when the model was applied to the NSL KDD and KDD Cup 99 datasets to detect intrusions.

A hybrid approach that combines the synthetic minority oversampling technique (SMOTE) and cluster center and nearest neighbor (CANN) was proposed [115]. Significant features were selected by utilizing the leave one out (LOO) approach. In addition, the research utilized the NSL-KDD dataset and the results illustrate that the proposed approach increases the accuracy of the R2L and U2R attacks as opposed to the benchmark paper by 50% and 94%, respectively.

A Hybrid RBF-SVM ensemble classification was proposed by [110] utilizing Support Vector Machine (SVM) and Radial Basis Function (RBF) as base classification. The efficacy and advantages of the proposed model are presented using NSL-KDD datasets, and their finding illustrates that the proposed ensemble RBF-SVM is superior to single-method approaches in terms of accuracy as it achieved 98.46%.

An ensemble-based IDS model was designed using integrated feature selection approach and an ensemble of ML classifiers comprising Bayesian Network, J48, and Naive Bayes [15]. In this model, features are reduced from 41 to 12, and majority vote is used for combing the findings. The true positive rate (TP) of the proposed model is 98.0% with a false-positive rate (FP) of 0.021%.

A hybrid classification approach was proposed to detect and forecast DDoS threats. Using the KDD Cup 99 dataset as attack data, related features were chosen based on information gain. The experimental result revealed that each step of the threat case is well divided, and they can identify DDoS threat precursors as well as the threat itself [116].

A model for Adaptive Ensemble Learning was proposed by [38] by changing the learning data ratio and constructing a MultiTree algorithm which deploys multiple decision trees. To increase detection efficiency, a number of base classifiers are chosen, including Random Forest, decision tree, deep neural network (DNN), KNN, and an adaptive voting algorithm were developed. For the validation part, the NSL-KDD dataset was used, and the MultiTree algorithm accuracy was 84.2%, while the final adaptive voting ensemble accuracy was 85.2%.

A model called SCDNN combines spectral clustering (SC) and DNN algorithms was proposed by [117]. In this model, k subsets were created from the dataset based on the similarity of the sample utilizing cluster centers as in SC. Then, the distance between data points in the training set and the test set was calculated on the basis of features similarity and was applied into the DNN algorithm to detect intrusion. NSL-KDD dataset was used for evaluation benchmark and the overall accuracy was 92.1%.

A framework with feature selection and ensemble method [118] integrates correlation-based feature selection with Bat algorithm (CFS-BA), and an ensemble of Random Forest (RF), C4.5 and Forest by Penalizing Attributes (Forest PA) is developed for the detection model. The evaluation experiments used the CIC-IDS2017, AWID, and NSL-KDD datasets. The results show that this framework has better accuracy than other research work.

A hierarchical ensemble classifier and knowledge-based method was proposed by [119]. In order to determine the specific attack class, it used a weighted voting fusion technique for specific classes to obtain a more accurate classification. The KDD Cup 99 dataset was used to prove the model. This IDS model has more complexity during the learning phase and it consumes more time in contrast to other work.

A Hybrid IDS of One Class Support Vector Machine (OC-SVM) and C5 decision tree classifier [42] was proposed to detect unknown and known intrusion. To the model was evaluated using the ADFA and NSL-KDD datasets. Their finding demonstrated that the hybrid schema has better performance than other models.

An IDS ensemble model of convolutional neural network, Random Forest, and gated recurrent unit (GRU) was proposed by [120]. NSL-KDD dataset was utilized to prove the performance of the model. The detection accuracy was 76.61% with reduced learning time and resource usage than other schema.

An IDS model with combination of ensemble (Random Forest, J48, and Reptree) and CFS algorithm suggested by [121]. Experimented on the KDD Cup 99 and NSLKDD datasets, their finding illustrates that the proposed ensemble has 99.90% for the KDD99 dataset, and 98.60% detection rate for NSLKDD. However, this model could not handle imbalance data.

A stacked ensemble classifier with a combination of gradient boosting machine, XGBoost, and Random Forest [39] was proposed and experimented on CICIDS-2017, CSIC-2010v2, UNSW-NB15, and NSL-KDD. The result shows that the proposed ensemble model has good impact on detection of attack in a Web application.

Table IV introduces a comparative analysis of different ensemble algorithms used in the literature to handle anomaly-based IDS. The table presents a comprehensive review of several ensemble classifications, showing their methods, strength, weakness and the dataset utilized for evaluation. RQ3 and RQ6 are addressed in table.

According to the reviewed articles presented in the table, different combination of classifiers and algorithms were utilized for ensemble detection. An ensemble with diversity of classifier types had significant improvements in detection accuracy and reduces the false alarm for anomaly-based IDS.

Based on the review, NSL-KDD dataset was mostly used to show the efficacy and advantages of the proposed ensemble models. Furthermore, some articles utilized different datasets to highlight their generality of their solutions, like AWID, ISCX 2012, UNSW-NB15, CIC-IDS2017 and CSIC-2010v2.

Based on the analysis of the studied articles in the review, Fig. 4 illustrates that NSL-KDD dataset was mostly utilized to highlight the effectiveness of their anomaly-based IDS models. The KDD Cup 99 dataset came in second as to be used to evaluate their solutions.

TABLE IV. SUMMARY OF THE REVIEWED INTRUSION DETECTION SYSTEM

Author	Method	Strength	Weakness	Dataset
2013 [27]	Unsupervised Artificial Neural Network	Hierarchical Anomaly-based Intrusion Detection System	Overall detection accuracy of 75.49%. No class-based detection.	KDD Cup 99 NSL-KDD
2014 [110]	Support Vector Machine (SVM) And Radial Basis Function (RBF)	They develop a bagging classifier	Overall accuracy was not reasonable. No class-based detection.	NSL-KDD
2015 [47]	Hybrid Artificial Bee Colony Algorithm and Multi-Layer Perceptron	The proposed model has reasonable detection time and Error rate (0.124%)	Overall accuracy was 87.54 % No class-based detection.	NSL-KDD
2015 [15]	Ensemble of Bayesian Network, J48, and Naive Bayes	The overall accuracy and error rate was reasonable	No class-based detection.	NSL-KDD
2016 [115]	Hybrid cluster center and nearest neighbor (CANN) and synthetic minority oversampling technique (SMOTE)	Reasonable detecting R2L and U2R attacks (50% and 94%,)	Not detect the rest of attacks. Not mentioned about time.	NSL-KDD
2016 [26]	Hybrid of Fuzzy C-Means clustering algorithm and Artificial Neural Network	The model was good at finding normal and probe attacks	The results for other attack types did not yield good results even for accuracy and false alarm rate (DOS (99.96–5.33), U2R (96.78–3.22) and R2L (93.73–6.27)	KDD Cup 99
2016 [46]	Hybrid of C4.5 decision tree and the MLP classifier	Use feature selection by utilizing the Discernibility Function and MLP to provide feature subset	Detection rate R2L and U2R attacks were not satisfactory.	KDD Cup 99 ISCX dataset
2016 [28]	ANN-based IDS with back propagation and feed forward algorithms	The overall accuracy was reasonable (98.86%)	No class-based detection.	NSL-KDD
2019 [38]	Ensemble of Decision Tree, Random Forest, KNN, DNN and MultiTree	Detection based on attack class	Class based accuracy was not satisfactory. Performance metric was low.	NSL-KDD
2020 [118]	Feature Selection CFS-BA and an Ensemble of Random Forest (RF), C4.5 and Forest by Penalizing Attributes (Forest PA)	They train and test their framework with different dataset	Detection rate R2L and U2R attacks were not satisfactory.	NSL-KDD AWID CIC-IDS2017
2020 [119]	Hierarchical ensemble classifier and knowledge base method	The class-based accuracy and error rate was reasonable	The IDS model has more complexity during learning phase and it consume more time in contrast to other work.	KDD Cup 99
2020 [42]	Hybrid of One Class Support Vector Machine (OC-SVM) and C5 decision tree classifier	The overall accuracy and error rate was reasonable	No class-based detection.	ADFA NSL-KDD
2020 [120]	Ensemble of Convolutional Neural Network, Random Forest, and Gated Recurrent Unit (GRU)	The model has improvement on reduction of learning time and resource usage	Overall accuracy was not reasonable. No class-based detection.	NSL-KDD
2020 [121]	Combination of Ensemble (Random Forest, J48, and Reptree) and CFS algorithm	The overall accuracy and error rate was reasonable	The model could not handle imbalance data issue.	KDD Cup 99 NSL-KDD
2020 [39]	Ensemble of Gradient Boosting Machine, XGBoost, and Random Forest	The overall accuracy and error rate was reasonable	No class-based detection.	CICIDS-2017 CSIC-2010v2 UNSW-NB15 NSL-KDD

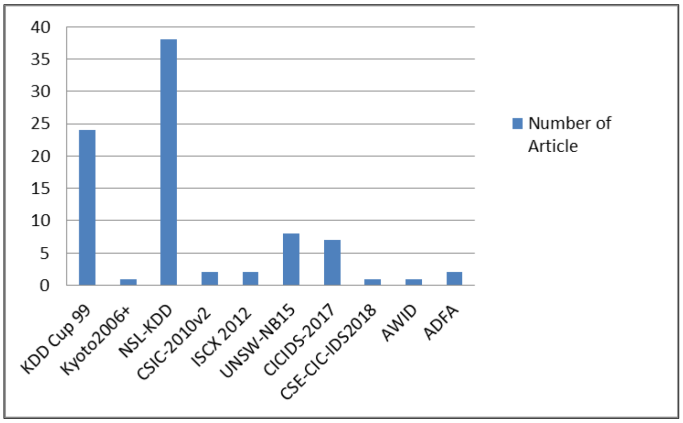


Fig. 4. Number of Datasets in the Reviewed Articles.

1) Limitations of the ensemble classification: After analyzing the data collected from the literature related to ensemble, some limitations and shortcomings of the works are identified and in order to reach maximum diversity with various boundaries of decision, the identified limitation should be considered:

a) Multiple datasets have to be utilized to prove the generality of the ensemble model.

b) In order to handle imbalance data issues in anomaly-based IDS, different types of classifiers have to be deploy in ensemble machine. Therefore, selection of various classifiers and the fusion of their outcomes empower the final result.

VI. Discussion

Upon studying and reviewing the different IDS models, we found challenges that motivate research in utilizing machine learning for feature selection and ensemble techniques in IDS. In this paper, we discuss future trends in anomaly-based IDS, in particular feature selection and ensemble techniques. Some of the critical topics in the existing research with view of future trends are described below:

1) Anomaly-based IDS datasets have a crucial impact on the proposed approaches in terms of performance assessment. To be current, it is necessary to utilized updated datasets to illustrate that the proposed solution works well with new attack types. Although KDD Cup 99 is an old dataset used by most of researchers as benchmark comparisons, the attack packets and even the features are dated 20 years ago. In addition, researchers can deploy their model on different anomaly-based IDS datasets to prove the generality of their model to detect different attacks.

2) Finding the appropriate feature selection schema plays an important role in anomaly-based IDS. Proper selection of feature subset helps expert machine in the learning phase to detect attacks in the testing phase. Optimization-based feature selection aims to acquire an optimal subset of features among all features in different domains. The role of new optimization-based feature selection methods in the success of anomaly-based IDS must be considered.

3) Ensemble-based modern anomaly-based IDS techniques allow multiple combinations of models or algorithms to identify new unseen cases. In the implementation, after a variety of classification models are typically constructed utilizing some portion of datasets, the various classifiers results are merged to form the final conclusion. Various schemes may be suggested for the generation of classifiers and for the combination of the ensembles.

The future trends mentioned above and open issues discussed in anomaly-based intrusion detection system should be considered by researchers in the field of anomaly-based IDS.

VII. Conclusions

Intrusion detection system is a prominent security mechanism designed to prevent intrusion, illicit entry, modification or demolition by intruders. For efficient intrusion detection process vital components like feature selection and detection mechanism have to be considered when designing the model. The article reviews the studies on feature selection and ensemble approaches utilized for anomaly-based intrusion detection systems. We discussed the main challenges in IDS, namely the dimensionality reduction in anomaly-based IDS that reduces irrelevant attributes from dataset; and how to build an appropriate feature subset selection, in order to better detect intrusion by increasing the performance metrics. Consequently, the study categorizes and discusses feature selection methods and presents their performance in detection accuracy. Another important challenge in anomaly-based IDS lies in utilizing suitable machine learning algorithms in the detection process. To illustrate their effectiveness in improving the IDS performance, this paper reviews and categorizes various machine learning schema and discussed their utilization in IDS, giving emphasize on ensemble methods as an emerging trend in anomaly-based IDS. Based on our study on anomaly-based IDS and the assessment and comparison of feature selection and detection module, we can summarize two points about how to boost the performance of anomaly-based IDS as follows:

1) Optimization-based feature selection with excellent combination and well tune up parameters will select the proper feature subset for IDSs. Through this study, it is clear that optimization-based have significant performance to design the optimal feature set. Furthermore, if their parameters are adjusted well, feature selection could be significantly enhanced.

2) Ensemble detection with different types of classification can empower the detection phase and reduce the false alarm rate. If the diversity occurred, a fusion of the outcome has better chance to detect properly.

Finally, we present some open issues and offered research trends, including the datasets used, the role of optimization-based algorithm-ms and ensemble methods, in the area of anomaly-based IDS. We expect that this review paper will furnish scientists with innovative ideas and serve as a springboard for them to undertake better studies. We acknowledge that this article has some limitations due to the scope of the review:

1) This review focused on the feature selection and ensemble detection for anomaly-based IDS.

2) This review does not focus on performance parameter which is utilized at IDS.

3) This article does not study IDS datasets in-depth, like their features, attack types, etc.

Having listed the limitations of the paper, a deep analysis on the following issues can be considered as future work:

1) Other detection methods for anomaly-based IDS, apart from the feature selection and ensemble detection methods that are discussed here, could be studied too, in order to acquire a more holistic understanding of the research area.

2) Extra studies could be performed on performance parameters which are utilized in IDS, and how we can obtain the optimal set of parameters for better detection performance.

3) An in-depth study on IDS datasets could be carried out, such as their features, attack types, etc. to understand the pattern in their attributes that may affect the detection performance.

References

[1] Patel, Q. Qassim, and C. Wills, “A survey of intrusion detection and prevention systems,” Inf. Manag. Comput. Secur., vol. 18, no. 4, pp. 277–290, 2010.

[2] J. R. Vacca, Computer and Information Security Handbook, vol. 82, no. 90001. 2013.

[3] H. I. Alsaadi, R. M. Almuttairi, O. Bayat, and O. N. Ucani, “Computational intelligence algorithms to handle dimensionality reduction for enhancing intrusion detection system,” J. Inf. Sci. Eng., vol. 36, no. 2, pp. 293–308, 2020.

[4] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” IEEE Access, vol. 8, pp. 54776–54788, 2020.

[5] O. Almomani, “A feature selection model for network intrusion detection system based on pso, gwo, ffa and ga algorithms,” Symmetry (Basel)., vol. 12, no. 6, pp. 1–20, 2020.

[6] M. Samadi Bonab, A. Ghaffari, F. Soleimanian Gharehchopogh, and P. Alemi, “A wrapper-based feature selection for improving performance of intrusion detection systems,” Int. J. Commun. Syst., vol. 33, no. 12, pp. 1–25, 2020.

[7] A. Aldweesh, A. Derhab, and A. Z. Emam, “Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues,” Knowledge-Based Syst., vol. 189, p. 105124, 2020.

[8] W. Fang, X. Tan, and D. Wilbur, “Application of intrusion detection technology in network safety based on machine learning,” Saf. Sci., vol. 124, no. December 2019, p. 104604, 2020.

[9] N. Moustafa, J. Hu, and J. Slay, “A holistic review of Network Anomaly Detection Systems: A comprehensive survey,” J. Netw. Comput. Appl., vol. 128, pp. 33–55, 2019.

[10] S. Jose, D. Malathi, B. Reddy, and D. Jayaseeli, “A Survey on Anomaly Based Host Intrusion Detection System,” J. Phys. Conf. Ser., vol. 1000, no. 1, 2018.

[11] H. J. Liao, C. H. Richard Lin, Y. C. Lin, and K. Y. Tung, “Intrusion detection system: A comprehensive review,” J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16–24, 2013.

[12] C. Liu, J. Yang, R. Chen, Y. Zhang, and J. Zeng, “Research on immunity-based intrusion detection technology for the Internet of Things,” in Proceedings - 2011 7th International Conference on Natural Computation, ICNC 2011, 2011, vol. 1, pp. 212–216.

[13] W. Li, S. Tug, W. Meng, and Y. Wang, “Designing collaborative blockchained signature-based intrusion detection in IoT environments,” Futur. Gener. Comput. Syst., vol. 96, pp. 481–489, 2019.

[14] F. Erlacher and F. Dressler, “FIXIDS: A high-speed signature-based flow intrusion detection system,” in NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, 2018, pp. 1–8.

[15] N. F. Haq, A. R. Onik, and F. M. Shah, “An ensemble framework of anomaly detection using hybridized feature selection approach (HFSA),” IntelliSys 2015 - Proc. 2015 SAI Intell. Syst. Conf., pp. 989–995, 2015.

[16] M. Monshizadeh, V. Khatri, B. G. Atli, R. Kantola, and Z. Yan, “Performance Evaluation of a Combined Anomaly Detection Platform,” IEEE Access, vol. 7, pp. 100964–100978, 2019.

[17] A. I. Hajamydeen and N. I. Udzir, “A detailed description on unsupervised heterogeneous anomaly based intrusion detection framework,” Scalable Comput., vol. 20, no. 1, pp. 113–160, 2019.

[18] J. M. Estevez-Tapiador, P. Garcia-Teodoro, and J. E. Diaz-Verdejo, “Stochastic protocol modeling for anomaly based network intrusion detection,” in Proceedings - 1st IEEE International Workshop on Information Assurance, IWIA 2003, 2008, vol. 02798, pp. 3–12.

[19] R. Sekar et al., “Specification-based anomaly detection: A new approach for detecting network intrusions,” Proc. ACM Conf. Comput. Commun. Secur., pp. 265–274, 2002.

[20] PMG, “Maximizing the value of network intrusion detection,” in A corporate white paper from the product management group of intrusion.com, 2001.

[21] W. Yassin, N. I. Udzir, A. Abdullah, M. T. Abdullah, H. Zulzalil, and Z. Muda, “Signature-Based Anomaly intrusion detection using Integrated data mining classifiers,” Proc. - 2014 Int. Symp. Biometrics Secur. Technol. ISBAST 2014, pp. 232–237, 2015.

[22] V. Hajisalem and S. Babaie, “A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection,” Comput. Networks, vol. 136, pp. 37–50, 2018.

[23] M. A. Hatef, V. Shaker, M. R. Jabbarpour, J. Jung, and H. Zarrabi, “HIDCC: A hybrid intrusion detection approach in cloud computing,” Concurr. Comput. , vol. 30, no. 3, 2018.

[24] I. Lorenzo-Fonseca, F. Maciá-Pérez, F. J. Mora-Gimeno, R. Lau-Fernández, J. A. Gil-Martínez-Abarca, and D. Marcos-Jorquera, “Intrusion detection method using neural networks based on the reduction of characteristics,” in International Work-Conference on Artificial Neural Networks, 2009, pp. 1296–1303.

[25] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, “Modeling intrusion detection system using hybrid intelligent systems,” J. Netw. Comput. Appl., vol. 30, no. 1, pp. 114–132, 2007.

[26] N. Pandeeswari and G. Kumar, “Anomaly Detection System in Cloud Environment Using Fuzzy Clustering Based ANN,” Mob. Networks Appl., vol. 21, no. 3, pp. 494–505, 2016.

[27] L. M. Ibrahim, D. T. Basheer, and M. S. Mahmod, “A comparison study for intrusion database (KDD99, NSL-KDD) based on self organization map (SOM) artificial neural network,” J. Eng. Sci. Technol., vol. 8, no. 1, pp. 107–119, 2013.

[28] B. Subba, S. Biswas, and S. Karmakar, “A Neural Network Based System for Intrusion Detection and Attack Classification,” 2016 22nd Natl. Conf. Commun. NCC 2016, pp. 1–6, 2016.

[29] M. Barati, A. Abdullah, N. I. Udzir, M. Behzadi, R. Mahmod, and N. Mustapha, “Intrusion detection system in secure shell traffic in cloud environment,” J. Comput. Sci., vol. 10, no. 10, pp. 2029–2036, 2014.

[30] P. Jokar and V. C. M. Leung, “Intrusion Detection and Prevention for ZigBee-Based Home Area Networks in Smart Grids,” IEEE Trans. Smart Grid, vol. 9, no. 3, pp. 1800–1811, 2011.

[31] H. C. Lin, M. K. Sun, H. W. Huang, C. Y. H. Tseng, and H. T. Lin, “A specification-based intrusion detection model for wireless ad hoc networks,” Proc. - 3rd Int. Conf. Innov. Bio-Inspired Comput. Appl. IBICA 2012, pp. 252–257, 2012.

[32] F. Zhang and D. Wang, “An effective feature selection approach for network intrusion detection,” Proc. - 2013 IEEE 8th Int. Conf. Networking, Archit. Storage, NAS 2013, pp. 307–311, 2013.

[33] W. Meng, W. Li, and L. F. Kwok, “EFM: Enhancing the performance of signature-based network intrusion detection systems using enhanced filter mechanism,” Comput. Secur., vol. 43, pp. 189–204, 2014.

[34] M. H. Aghdam and P. Kabiri, “Feature selection for intrusion detection system using ant colony optimization,” Int. J. Netw. Secur., vol. 18, no. 3, pp. 420–432, 2016.

[35] H. Bostani and M. Sheikhan, “Hybrid of anomaly-based and specification-based IDS for Internet of Things using unsupervised OPF based on MapReduce approach,” Comput. Commun., vol. 98, pp. 52–71, 2017.

[36] A. Althubaity, H. Ji, T. Gong, M. Nixon, R. Ammar, and S. Han, “ARM: A hybrid specification-based intrusion detection system for rank attacks in 6TiSCH networks,” IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA, pp. 1–8, 2017.

[37] Y. Wang, W. Meng, W. Li, J. Li, W. X. Liu, and Y. Xiang, “A fog-based privacy-preserving approach for distributed signature-based intrusion detection,” J. Parallel Distrib. Comput., vol. 122, pp. 26–35, 2018.

[38] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, “An Adaptive Ensemble Machine Learning Model for Intrusion Detection,” IEEE Access, vol. 7, pp. 82512–82521, 2019.

[39] B. A. Tama, L. Nkenyereye, S. M. R. Islam, and K. S. Kwak, “An enhanced anomaly detection in web traffic using a stack of classifier ensemble,” IEEE Access, vol. 8, pp. 24120–24134, 2020.

[40] S. Dwivedi, M. Vardhan, and S. Tripathi, “An effect of chaos grasshopper optimization algorithm for protection of network infrastructure,” Comput. Networks, vol. 176, no. March, 2020.

[41] A. R. Gupta and J. Agrawal, “The multi-demeanor fusion based robust intrusion detection system for anomaly and misuse detection in computer networks,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 1, pp. 303–319, 2020.

[42] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, and A. Alazab, “Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vector machine,” Electron., vol. 9, no. 1, 2020.

[43] A. Nazir and R. A. Khan, “A novel combinatorial optimization based feature selection method for network intrusion detection,” Comput. Secur., vol. 102, p. 102164, 2020.

[44] “MLP structure,” https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

[45] Margaret H. Dunham, “Data mining – introductory and advanced topics,” Pearson Educ., pp. 106–114, 2003.

[46] A. Akyol, M. Hacibeyoglu, and B. Karlik, “Design of multilevel hybrid classifier with variant feature sets for intrusion detection system,” IEICE Trans. Inf. Syst., vol. E99D, no. 7, pp. 1810–1821, 2016.

[47] M. S. Mahmod, Z. A. H. Alnaish, and I. A. A. Al-hadi, “Hybrid Intrusion Detection System Using Artificial Bee Colony Algorithm and Multi-Layer Perceptron,” vol. 13, no. 2, pp. 1–7, 2015.

[48] Liao and V. R. Vemuri, “Use of k-nearest neighbor classifier for intrusion detection,” Comput. Secur., vol. 21, no. 5, pp. 439–448, 2002.

[49] R. S. Naoum and Z. N. Al-Sultani, “Learning Vector Quantization (LVQ) and k-Nearest Neighbor for Intrusion Classification,” World Comput. Sci. Inf. Technol. J., vol. 2, no. 3, pp. 105–109, 2012.

[50] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Min. Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1998.

[51] H. Eid, “Computational Intelligence in Intrusion Detection System,” 2013.

[52] A. Chalak, “Data Mining Techniques for Intrusion Detection and Prevention System,” 2011, vol. 11, no. 8, pp. 200–203.

[53] F. Kuang, W. Xu, and S. Zhang, “A novel hybrid KPCA and SVM with GA model for intrusion detection,” Appl. Soft Comput. J., vol. 18, pp. 178–184, 2014.

[54] S. J. Horng et al., “A novel intrusion detection system based on hierarchical clustering and support vector machines,” Expert Syst. Appl., vol. 38, no. 1, pp. 306–313, 2011.

[55] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, “Mining network data for intrusion detection through combining SVMs with ant colony networks,” Futur. Gener. Comput. Syst., vol. 37, pp. 127–140, 2014.

[56] C. F. Tsai, Y. F. Hsu, C. Y. Lin, and W. Y. Lin, “Intrusion detection by machine learning: A review,” Expert Syst. Appl., vol. 36, no. 10, pp. 11994–12000, 2009.

[57] S. Agrawal and J. Agrawal, “Survey on anomaly detection using data mining techniques,” Procedia Comput. Sci., vol. 60, no. 1, pp. 708–713, 2015.

[58] N. Ben Amor, S. Benferhat, and Z. Elouedi, “Naive bayesian networks in intrusion detection systems,” in 14th European Conference On Machine Learning 17th European Conference On Principles And Practice Of Knowledge Discovery In Databases, 2003.

[59] D. M. Singh, N. Harbi, and M. Zahidur Rahman, “Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection,” Int. J. Netw. Secur. Its Appl., vol. 2, no. 2, pp. 12–25, 2010.

[60] J. E. Dickerson and J. A. Dickerson, “Fuzzy network profiling for intrusion detection,” in PeachFuzz 2000. 19th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 00TH8500), 2000, pp. 301–306.

[61] K. Bajaj and A. Arora, “Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach.,” … J. Comput. Sci. Issues (IJCSI …, vol. 10, no. 4, pp. 324–329, 2013.

[62] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, “Intrusion detection based on K-Means clustering and Naïve Bayes classification,” 2011 7th Int. Conf. Inf. Technol. Asia Emerg. Converg. Singul. Forms - Proc. CITA’11, pp. 1–6, 2011.

[63] W. Yassin, N. I. Udzir, and Z. Muda, “Anomaly-Based Intrusion Detection Through K- Means Clustering and Naives Bayes Classification,” Proc. 4th Int. Conf. Comput. Informatics, ICOCI 2013, no. 049, pp. 298–303, 2013.

[64] M. Barati, A. Abdullah, R. Mahmod, N. Mustapha, and N. I. Udzir, “Features Selection for Ids in Encrypted Traffic Using Genetic Algorithm,” Proc. 4th Int. Conf. Comput. Informatics, ICOCI 2013, no. 038, pp. 279–285, 2013.

[65] R. Xu and D. C. Wunsch, “Survey of clustering algorithms,” 2005.

[66] E. Vasilomanolakis, S. Karuppayah, M. Muhlhauser, and M. Fischer, “Taxonomy and survey of collaborative intrusion detection,” ACM Comput. Surv., vol. 47, no. 4, pp. 1–33, 2015.

[67] A. Fahad, Z. Tari, I. Khalil, A. Almalawi, and A. Y. Zomaya, “An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion,” Futur. Gener. Comput. Syst., vol. 36, pp. 156–169, 2014.

[68] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, “Toward an efficient and scalable feature selection approach for internet traffic classification,” Comput. Networks, vol. 57, no. 9, pp. 2040–2057, 2013.

[69] Z. Liu, R. Wang, M. Tao, and X. Cai, “A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion,” Neurocomputing, vol. 168, pp. 365–381, 2015.

[70] E. De La Hoz, E. De La Hoz, A. Ortiz, J. Ortega, and A. Martínez-Álvarez, “Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps,” Knowledge-Based Syst., vol. 71, pp. 322–338, 2014.

[71] Y. Li, J. L. Wang, Z. H. Tian, T. B. Lu, and C. Young, “Building lightweight intrusion detection system using wrapper-based feature selection mechanisms,” Comput. Secur., vol. 28, no. 6, pp. 466–475, 2009.

[72] K. Deb, A. Member, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multi-objective genetic algorithm:NSGAII,” vol. 6, no. 2, pp. 182–197, 2002.

[73] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, “Feature selection for optimizing traffic classification,” Comput. Commun., vol. 35, no. 12, pp. 1457–1471, 2012.

[74] Y. Zhu, J. Liang, J. Chen, and Z. Ming, “An improved NSGA-III algorithm for feature selection used in intrusion detection,” Knowledge-Based Syst., vol. 116, pp. 74–85, 2017.

[75] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997.

[76] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, no. Mar, pp. 1157–1182, 2003.

[77] G. H. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem,” in Machine Learning Proceedings 1994, Elsevier, 1994, pp. 121–129.

[78] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 9, pp. 1154–1166, 2004.

[79] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, “Mutual information-based feature selection for intrusion detection systems,” J. Netw. Comput. Appl., vol. 34, no. 4, pp. 1184–1199, 2011.

[80] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” Proceedings, Twent. Int. Conf. Mach. Learn., vol. 2, pp. 856–863, 2003.

[81] A. R. A. Yusof, N. I. Udzir, A. Selamat, H. Hamdan, and M. T. Abdullah, “Adaptive feature selection for denial of services (DoS) attack,” 2017 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2017, vol. 2018-Janua, pp. 1–4, 2018.

[82] H. Chen, R. Cheng, J. Wen, H. Li, and J. Weng, “Solving large-scale many-objective optimization problems by covariance matrix adaptation evolution strategy with scalable small subpopulations,” Inf. Sci. (Ny)., 2018.

[83] Y. Xue, B. Zhao, T. Ma, and A. X. Liu, “An evolutionary classification method based on fireworks algorithm.,” IJBIC, vol. 11, no. 3, pp. 149–158, 2018.

[84] K. Chen, F.-Y. Zhou, and X.-F. Yuan, “Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection,” Expert Syst. Appl., vol. 128, pp. 140–156, 2019.

[85] R. Vanaja and S. Mukherjee, “Novel Wrapper-Based Feature Selection for Efficient Clinical Decision Support System,” in International Conference on Intelligent Information Technologies, 2018, pp. 113–129.

[86] Y. Zhang, D. Gong, and J. Cheng, “Multi-objective particle swarm optimization approach for cost-based feature selection in classification,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 1, pp. 64–75, 2017.

[87] B. Agarwal and N. Mittal, “Hybrid approach for detection of anomaly network traffic using data mining techniques,” Procedia Technol., vol. 6, pp. 996–1003, 2012.

[88] B. M. Aslahi-Shahri et al., “A hybrid method consisting of GA and SVM for intrusion detection system,” Neural Comput. Appl., vol. 27, no. 6, pp. 1669–1676, 2016.

[89] B. Ma and Y. Xia, “A tribe competition-based genetic algorithm for feature selection in pattern classification,” Appl. Soft Comput., vol. 58, pp. 328–338, 2017.

[90] T. Mehmod and H. B. M. Rais, “Ant colony optimization and feature selection for intrusion detection,” Lect. Notes Electr. Eng., vol. 387, pp. 305–312, 2016.

[91] F. Kuang, S. Zhang, Z. Jin, and W. Xu, “A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection,” Soft Comput., vol. 19, no. 5, pp. 1187–1199, 2015.

[92] M. R. Gauthama Raman, N. Somu, K. Kirthivasan, R. Liscano, and V. S. Shankar Sriram, “An efficient intrusion detection system based on hypergraph - Genetic algorithm for parameter optimization and feature selection in support vector machine,” Knowledge-Based Syst., vol. 134, pp. 1–12, 2017.

[93] M. H. Ali, B. A. D. Al Mohammed, A. Ismail, and M. F. Zolkipli, “A New Intrusion Detection System Based on Fast Learning Network and Particle Swarm Optimization,” IEEE Access, vol. 6, pp. 20255–20261, 2018.

[94] S. Sarvari, N. F. Mohd Sani, Z. Mohd Hanapi, and M. T. Abdullah, “An Efficient Anomaly Intrusion Detection Method with Feature Selection and Evolutionary Neural Network,” IEEE Access, vol. 8, pp. 70651–70663, 2020.

[95] T. G. Dietterich, “Ensemble methods in machine learning,” in International workshop on multiple classifier systems, 2000, pp. 1–15.

[96] G. Folino and F. S. Pisani, “Combining ensemble of classifiers by using genetic programming for cyber security applications,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9028, no. February, 2015.

[97] M. P. Sesmero, A. I. Ledezma, and A. Sanchis, “Generating ensembles of heterogeneous classifiers using stacked generalization,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 5, no. 1, pp. 21–34, 2015.

[98] A. A. Aburomman and M. B. I. Reaz, “A survey of intrusion detection systems based on ensemble and hybrid classifiers,” Comput. Secur., vol. 65, pp. 135–152, 2017.

[99] L. Breiman, “Bagging predictors,” Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[100] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[101] L. Breiman, “Pasting small votes for classification in large databases and on-line,” Mach. Learn., vol. 36, no. 1–2, pp. 85–103, 1999.

[102] Sangeeta Bhattacharya and S. Selvakumar, “LAWRA: a layered wrapper feature selection approach for network attack detection,” Secur. Commun. NETWORKS, vol. 2, pp. 71–81, 2015.

[103] M. Ambusaidi, X. He, P. Nanda, and Z. Tan, “Building an intrusion detection system using a filter-based feature selection algorithm,” IEEE Trans. Comput., vol. PP, no. 99, p. 1, 2016.

[104] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, “Cyber intrusion detection by combined feature selection algorithm,” J. Inf. Secur. Appl., vol. 44, pp. 80–88, 2019.

[105] S. Hosseini and B. M. H. Zade, “New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN,” Comput. Networks, vol. 173, no. March, p. 107168, 2020.

[106] H. Alazzam, A. Sharieh, and K. E. Sabri, “A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer,” Expert Syst. Appl., vol. 148, 2020.

[107] A. Golrang, A. M. Golrang, S. Y. Yayilgan, and O. Elezaj, “A novel hybrid ids based on modified NSGAII-ANN and random forest,” Electron., vol. 9, no. 4, pp. 1–19, 2020.

[108] Z. Zhang, J. Wen, J. Zhang, X. Cai, and L. Xie, “A Many Objective-Based Feature Selection Model for Anomaly Detection in Cloud Environment,” IEEE Access, vol. 8, pp. 60218–60231, 2020.

[109] N. V Chawla, L. O. Hall, K. W. Bowyer, T. E. Moore, and W. P. Kegelmeyer, “Distributed pasting of small votes,” in International Workshop on Multiple Classifier Systems, 2002, pp. 52–61.

[110] M. Govindarajan, “Hybrid Intrusion Detection Using Ensemble of Classification Methods,” Int. J. Comput. Netw. Inf. Secur., vol. 6, no. 2, pp. 45–53, 2014.

[111] R. E. Schapire, “The strength of weak learnability,” Mach. Learn., vol. 5, no. 2, pp. 197–227, 1990.

[112] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.

[113] D. H. Wolpert, “Stacked generalization,” Neural networks, vol. 5, no. 2, pp. 241–259, 1992.

[114] A. KumarShrivas and A. Kumar Dewangan, “An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set,” Int. J. Comput. Appl., vol. 99, no. 15, pp. 8–13, 2014.

[115] M. R. Parsaei, S. M. Rostami, and R. Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset,” vol. 7, no. 6, pp. 20–25, 2016.

[116] A. R. Yusof, N. I. Udzir, and A. Selamat, “An Evaluation on KNN-SVM Algorithm for Detection and Prediction of DDoS Attack,” Springer Int. Publ. Switz., vol. 9799, no. 61272374, pp. 841–852, 2016.

[117] T. Ma, F. Wang, J. Cheng, Y. Yu, and X. Chen, “A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks,” Sensors (Switzerland), vol. 16, no. 10, 2016.

[118] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an efficient intrusion detection system based on feature selection and ensemble classifier,” Comput. Networks, vol. 174, no. March, 2020.

[119] M. Sarnovsky and J. Paralic, “Hierarchical Intrusion Detection Using Machine Learning and Knowledge Model,” Symmetry (Basel)., vol. 12, no. 203, pp. 1–14, 2020.

[120] A. Andalib and V. Tabataba Vakili, “An Autonomous Intrusion Detection System Using an Ensemble of Advanced Learners,” 2020 28th Iran. Conf. Electr. Eng., 2020.

[121] C. Iwendi, S. Khan, J. H. Anajemba, M. Mittal, M. Alenezi, and M. Alazab, “The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems,” Sensors (Switzerland), vol. 20, no. 9, pp. 1–37, 2020.

Design, Aggregation and Analysis of Power Consumption Data using the Jump Process

Yazid Hambally Yacouba[1]
Amadou Diabagaté[2], Adama Coulibaly[4]
Training and Research Unit for Mathematics and Computer Science, Félix Houphouet Boigny University
UPHB, Abidjan, Côte d'Ivoire

Michel Babri[3]
Training and Research Unit for Mathematics and Computer Science, INPHP, Yamoussoukro
Côte d'Ivoire

***Abstract*—This work aims to seek a pragmatic approach to assess electricity consumption at the level of households, buildings and neighborhoods. The main concern consists in proposing aggregation methods based on jump process according to a customer environment that is intrinsically linked to the implementation of a centralized system. The aim of the approach is to present data aggregations that derive their basis from a data model in order to facilitate the processing of electricity data at different scales of analysis. Such a smart meter data management process merits the design of an aggregated database that can store data for a house, a building and a neighborhood. The advantage of this system lies in the facilitation of data interpretation and the ability to guide decision-makers in the management of electricity consumption. An analysis of the behavior of electricity consumption is also proposed based on the monitoring of the electricity consumption of the various devices connected to a smart meter.**

Keywords*—Design; aggregation; analysis; jump process; electricity consumption; smart meter

I. Introduction

The systems for managing data from smart meters remain closed in the way they process data. However, these systems have the merit of presenting detailed output states. This work consists of:

- Analyzing the existing framework in favor of measuring methods for the smart meter in particular.
- Studying and adapting this research work to the documents and output reports made available by data managers of these smart meters.
- Initializing a data model for the management of smart metering data.
- Explaining the treatments of the central system including the various techniques applied and the presentation of data structures for facilitating data aggregation.
- Presenting the aggregation methods implemented using the jump process.

This work follows a previous work [1] related to the design of a multi-agent system for the management of data from smart meters. The advantage of this approach is to deepen the design of the system by providing it with data aggregation methods that would lead to results for the benefit of decision-makers.

In [1], presents a formal framework for storing smart meter data and applying methods for analyzing electricity consumption data. For this, a database model for evaluating the electricity consumption data of a house, a building and a neighborhood is proposed. This work highlights a structured set of measurement data from smart meters. There are many elements to consider for the implementation of a data model [2, 3, 4]. Jaime Lloret and al. propose an integrated architecture based on the Internet of Things (IoT) for the deployment of smart meter networks [2]. This article [3] offers a large number of datasets as well as a web portal for visualizing this data. The work in this publication presents key technological solutions for better data analysis and energy efficiency research. In [4], reviews technologies oriented and adapted to the development of applications for the management of smart meter data. Yi Wang and al. provide insight into the future challenges of descriptive analytics including load analysis, predictive analytics, forecasting and prescriptive analytics for load management [4].

The processing of the data collected is carried out by means of aggregation methods. It is also about presenting data structures that will evolve progressively as needed. This work presents existing studies highlighting the importance of forecasts for a good management of electricity consumption, mainly at the level of end customers [5, 6]. In [5], highlights the need for detailed technical information on the devices used by households in order to measure their impacts on household behavior. Ya Wu and Li Zhang study the factors influencing energy saving in tiered electricity pricing and presents empirical results based on variables such as personal characteristics and living conditions [6].

A survey of studies similar to the approach conducted in this paper help to identify a few cases with different interests from each other. Xiaochen Zhang and al. propose a time-variant load model based on the exploration of a historical database of smart meter data [7]. A relational database management system is set up to leverage information that consumers and utilities can get from smart meters data [8]. The achievement of new energy efficiency services is highlighted by the use of smart meters with smartphones through an infrastructure and a set of algorithms [9]. Ming Dong and al. propose a method to help monitoring energy consumption and identifying individual consumption from large household appliances [10]. Xiufeng Liu and al. [11] propose an innovative ICT solution based on a hybrid architecture, which proves the importance of smart meters in socio-economic surveys and proposes a rational approach to dealing with the complexity of their data. This solution concerns more areas of interest such as geographic locations, weather conditions and user information. Peng Xu and al. propose an ICT solution that aims to highlight the efficiency of smart meters and remote displays through real-time monitoring functions of energy consumption to achieve objectives in environmental and supply security [12]. Juan I. Guerrero and al. provide a method for integrating smart meter data from heterogeneous data sources and modeling the integrated information through an automatic data mining framework [13]. Jinsong Liu and al. [14] present the requirements of the fundamental models for the management of smart metering data and presents a concept of three-mode models as well as the design of the service architecture for accessing to generic models.

The review of work around the evaluation of electricity consumption and its opportunities has made it possible to identify other areas of interest, including studies relating to aggregations, forecasts, evaluation of electrical data and their impacts. There are many approaches of data aggregation that can be temporal, zone and by use. The large volume of these data has also led to the exploration of aggregation methods allowing this data to be represented in other forms while retaining the initial information contained in the raw data [15, 16, 17, 18, 19, 20, 21, 22]. Yi Wang, Qixin Chen and Chongqing Kang provide an overview of research related to the challenges of smart meter data analysis [15]. This work [16] presents a method based on Blockchain and Homomorphic Encryption (HE) technologies to ensure the confidentiality of smart meter data during the data aggregation phase. Mohamed Saleem Haja Nazmudeen and al. aim to adopt the fog computing architecture to reduce the amount of data collected in order to improve the performance of the process applicable to data within the central system [17]. In [18], highlights the need to implement a sub-metering of individual devices in a house to give a better interpretation to the data thanks to a proposed model called the Explicit-Duration Hidden Markov Model (EDHMM-diff). This paper [19] presents a model of data aggregation with fault tolerance, preventing data leakage and demonstrating robustness with negligible cost. Toshichika Shiobara and al. propose a method of reducing the cost of storing and using smart meter data through an aggregation method that reduces the total size of the data and the processing cost [20]. Accurate estimation of the energy [21] requirement forecast based on the amount of load growth and geographic space is carried out based on knowledge of small zone consumption, historical loads and weather conditions. The vector machine algorithm is used for data prediction, classification and analysis using a deep learning approach. Anastasia Ushakova and al. use a set of Gaussian models to aggregate time-series data from smart meters and to understand energy consumption based on the aggregation of data from the target population, particularly by inducing energy-efficient behavior [22].

There are also studies on both long-term and short-term forecasts of electricity demand [23, 24, 25]. In [23], highlights a new method of probabilistic forecasting of electricity demand through a hierarchy with different levels of aggregation such as substations, cities and regions. This study [24] proposes the evaluation of seven algorithms from measurement data collected every fifteen minutes from sensors to predict the electricity consumption of residential buildings in the next hour. This work confirms the best results already obtained from commercial buildings data with methods based on the neural network, unlike the results obtained from residential data. However, the conclusion of this work also shows that the results with the data of residential buildings are better with the Least Squares Support Vector Machines. This article [25] shows the important issue of forecasting electricity consumption by proposing prediction methods based on Short-Term Load Forecast (STLF), the Self-Recurrent Wavelet Neural Network (SRWNN) and the Levenberg–Marquardt (LM) learning algorithm.

Upstream of the aggregation and forecasting of electricity consumption, there is first the need to formalize the data from smart meters in a database model and this is why a set of studies on the evaluation of electricity consumption data from smart meters is presented. The evaluation of electricity data is indeed a major concern and a prerequisite for understanding consumer behavior in order to guarantee energy efficiency at the level of producers, suppliers and customers alike [26, 27, 28, 29]. [26] explores the behavior of electricity customers on the use of heating and cooling. This work also shows the importance of using smart a meter data for understanding a consumer behavior with respect to thermal comfort, particularly in regions where automatic HVAC systems are virtually absent. Ilze Laicane and al. prove the importance of having up-to-date electricity data to determine the profile of households and reduce electricity consumption [27]. This work shows that the energy performance of households depends on the energy efficiency of equipment due to technological progress while highlighting an important part of the change in user behavior in the results observed [27]. In [28], is based on the PROBE and CarbonBuzz initiatives which illustrate that the concept of energy performance for more energy efficiency advocated by the construction industry does not live up to expectations and is biased by taking into account unrealistic factors such as occupancy behavior and facility management relating to the energy models used within these buildings. The study is based on the evaluation of data after occupancy of buildings to establish more realistic energy performance models. In [29], shows the impact of airflow on electricity consumption in computer data centers based on a comparative study of four different cities. Analysis of the cooling periods and energy saving periods yielded results in showing the importance of climatic conditions, energy prices and cooling technologies on cooling efficiency and costs of cooling exploitation.

Likewise, the impacts of the evaluation of smart meter data are numerous [30, 31, 32] and fully participate in encouraging consumers to save energy by identifying different criteria for reducing electricity consumption. This study [30] presents the mistaken motivations of a sample of US grid electricity customers for choosing smart meters and associated technologies in residential homes and discusses the policy implications and risks perceived by customers. Jacopo Torriti shows that the demands of electricity customers using smart meters are 5.2 times lower than those of users of conventional meters using a comparative study carried out on three different floors of the same building in Italy [31]. Gordon Rausser, Wadim Strielkowski and Dalia Streimikien highlight relevant findings for the attention of stakeholders and policymakers on issues to consider in encouraging positive electricity consumption with smart meters [32].

The benefits of analyzing electricity consumption data from smart meters [5, 33, 34, 35, 36, 37] for consumers and for the environment in particular deserve to be highlighted. Iana Vassileva, Fredrik Wallin and Erik Dahlquist focuse on identifying appropriate power saving measures from important data collected and analyzed monthly over a long period from identical buildings [13]. This work consists in defining different behavioral consumption profiles from technical data on electricity consumption, tenant characteristics, energy consumption behaviors as well as type and use of electrical devices [13]. In [33], is a review of the literature that assesses the effectiveness of smart metering in reducing energy consumption while explaining factors related to users behavior. This article [34] is an in-depth analysis of the environmental impacts resulting from the use of smart meters by considering all stages of the life cycle of these meters. In [35], illustrates the importance of electricity consumption data in studying users' consumption patterns in order to classify households according to predefined criteria. This study shows that a priori knowledge of certain criteria such as floor level or number of occupants can improve the accuracy of household classification. In [36], aims to provide industrial consumers with a methodological approach allowing the evaluation and choice of offers from electricity suppliers. The method used is based on characterization of electricity consumption, analysis of tariff offers and forecasts based on energy factors. The results of the study were drawn from the evaluation of fourteen different contracts of electricity suppliers. In [37,] emphasizes the importance of long-term forecasting of electricity demand based on forecasts obtained from multivariate regression analysis of electricity consumption data.

Some works using the jump process among which [38, 39, 40, 41, 42] have been identified. The paper [38] determines the integral transforms of the joint distribution of the first-exit time from an interval and the value of a jump of a process over the boundary at exit time and the joint distribution of the supremum, infimum, and value of the process. Clifford A. Ball and Walter N. Torous show a simplified jump process for common stock returns using the jump process models information arrivals and, as such, stock price jumps [39]. The article [40] explores the use of the Markov jump process to model vehicular mobility at the macroscopic level. Lydia Chabane and al. study the fluctuations of systems modeled by Markov jump processes with periodic generators using large deviation theory; canonical biasing and generalized Doob transform [41]. They show that the asymptotic fluctuations process, called driven process, is the minimum under constraint of the large deviation function for occupation and jumps [41]. Alexander Sikorski and al. present how augmenting the spatial information of the embedded Markov chain by the temporal information of the associated jump times [42]. The approach presented in this work is a very different application of the jump process in a context of electricity consumption management.

II. Principle for the Aggregation Methods of the Smart Metering System

First, it should be noted that the data collection mechanisms both from smart meters and from the customer base make it possible to automatically provide data without duplication and by identifying proven or probable anomalies in the collection process in case of missing data or outliers through business analytics [1].

Data quality assessment should be subject to automated treatments. In addition, the storage of processed data must be possible over periods of time in accordance with the regulations in force in each country.

After data collection, it is necessary to aggregate the raw data as shown in Fig. 1.

- The aggregation of measurement data can be done according to a time step (hourly, daily, monthly, etc.) regardless of any consideration of the zone and use to which the measurements are linked. This temporal aggregation process can be done in the data collection transaction from the smart meters or in a separate transaction from the data collection process. It is sufficient to rely on the arrival date "ARRIVAL_DATE" of each data in the central system for the need of ordering the data of a source [1]. The collection of data from each smart meter requires the storage of the date "LAST_COLLECTION_DATE" of the last measurement data recorded in the source [1]. This mechanism makes it possible to ensure the collection of all data of a smart meter within the source to which it is attached. Several smart meters could also be associated with a single source. However, it is essential for the aggregation mechanism to save in a transaction, the "LAST_AGGREGATION_DATE" parameter for each source data. The "SOURCE" table has therefore been modified by adding the "LAST_AGGREGATION_DATE" field. "LAST_ AGGREGATION_DATE" corresponds to the arrival date in the central system of the last aggregated data for each source. Aggregation is therefore done in order of arrival of data from one or all the smart meters associated with a source as represented in Fig. 2.
- The aggregations are then carried out on the sub-zones of the lowest level towards the sub-zones of the highest level. The aggregation of the data of a bounding zone is conditioned by the aggregation of all its sub-zones.
- Data aggregations are also possible by use. Indeed, the aggregation of an use can be applied to all the measurement data of this use for a given zone.

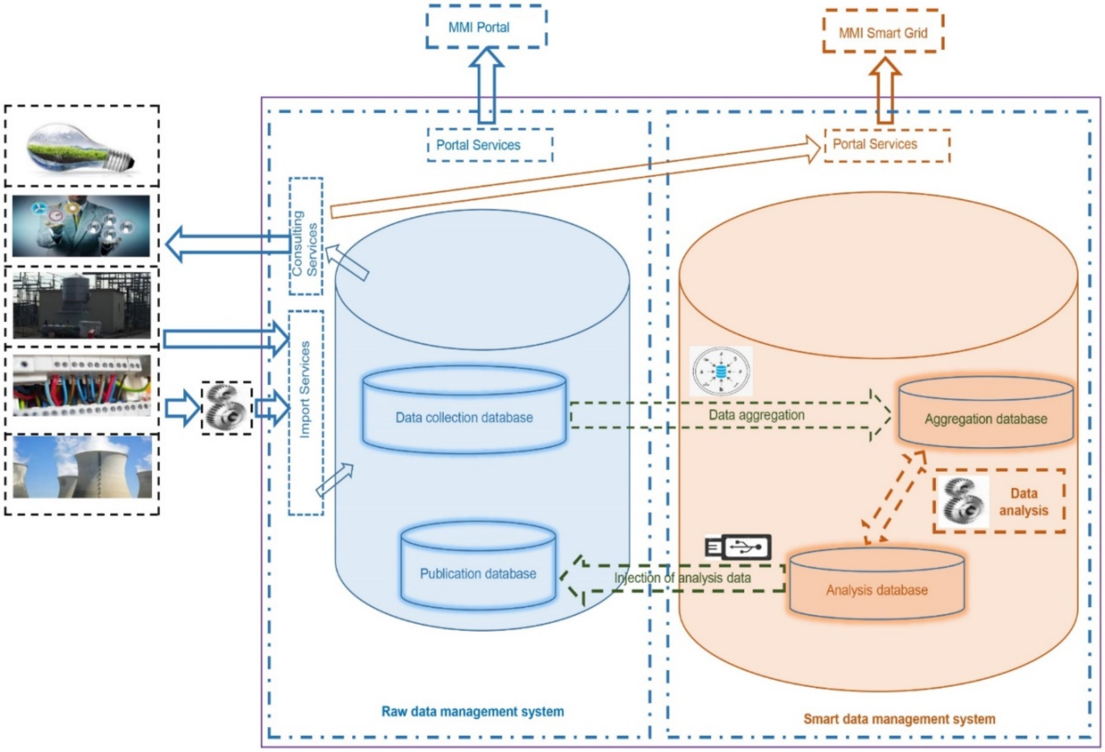
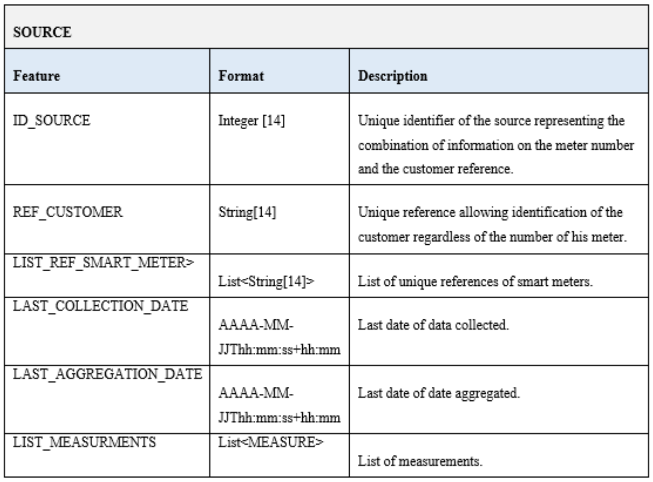


Fig. 1. Functional Architecture of the Electricity System Including Aggregations and Interactions with other Actors in the Electricity Value Chain.

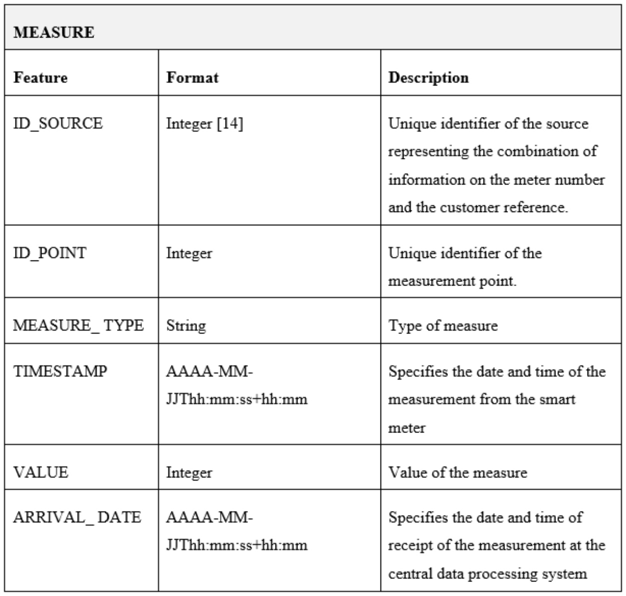


SOURCE		
Feature	**Format**	**Description**
ID_SOURCE	Integer [14]	Unique identifier of the source representing the combination of information on the meter number and the customer reference.
REF_CUSTOMER	String[14]	Unique reference allowing identification of the customer regardless of the number of his meter.
LIST_REF_SMART_METER>	List<String[14]>	List of unique references of smart meters.
LAST_COLLECTION_DATE	AAAA-MM-JJThh:mm:ss+hh:mm	Last date of data collected.
LAST_AGGREGATION_DATE	AAAA-MM-JJThh:mm:ss+hh:mm	Last date of date aggregated.
LIST_MEASURMENTS	List<MEASURE>	List of measurements.

Fig. 2. Representation of a Source.

- It will no longer be possible to integrate data for a validated (procedurally completed) or non-initiated billing period in the aggregated database. This condition helps to limit and avoid the overlap in the collection and aggregation of electricity consumption data over time. This will also make it possible to identify functional and technical irregularities, in particular cases of attempted fraud.

Let's present a view of the measurement data table regardless of its relationship to other entities in the master data model as indicated below in Fig. 3.



MEASURE		
Feature	**Format**	**Description**
ID_SOURCE	Integer [14]	Unique identifier of the source representing the combination of information on the meter number and the customer reference.
ID_POINT	Integer	Unique identifier of the measurement point.
MEASURE_ TYPE	String	Type of measure
TIMESTAMP	AAAA-MM-JJThh:mm:ss+hh:mm	Specifies the date and time of the measurement from the smart meter
VALUE	Integer	Value of the measure
ARRIVAL_ DATE	AAAA-MM-JJThh:mm:ss+hh:mm	Specifies the date and time of receipt of the measurement at the central data processing system

Fig. 3. Representation of the Electricity Consumption of a Measurement.

The analysis of this table highlights the correlation between the volume of data and the time step for collecting the data values for each measurement. The metering data is voluminous and therefore difficult to use.

Reducing complexity involves moving to a more understandable data presentation scale by aggregating data over hourly, daily and monthly ranges. This aggregation consists of the accumulation of consumption data collected over the various intervals for each measurement.

At first glance, collecting the measurements of a house can generate a significant volume of data. The amount of data from a building, or even a neighborhood might be gigantic. Temporal aggregation of data is necessary not only to compress the volume of data but also to facilitate data search, analysis and visualization.

III. Modelization of the Temporal, Zone and use Aggregations of the Data for the Smart Metering System using the Jump Process

A. *Fundamental Principle of Modelizing the System Aggregations by the Jump Process*

The jump process can be used to model the data aggregation of the smart metering system. Indeed, consumption data are recorded at successive and regular times. In addition, the current state of a smart meter is independent of the previous state at each time step of the arrival of electricity consumption data. The jump process thus models the data aggregations as follows:

1) Calculation of the electricity consumption of a smart meter at a time $\boldsymbol{t}$

$Z_t = \sum_{n=1}^{\infty} X_n \mathbb{1}_{[T_{n-1},T_n[}(\mathrm{t})$, electricity consumption at time t.

n the number of data records.

X_n electricity consumption between T_{n-1} and T_n at time t .

$\forall n, 0 = t_1 < t_2 \ldots < t_n$.

$\forall n$, $X_{t_{n-1}}$ and X_{t_n} are independent.

$$\mathbb{1}_A(t) = \begin{cases} 1\ if\ t \in A \\ 0\ otherwise \end{cases}$$

$T_n - T_{n-1} = H$, H being the consumption data recording time step for a given meter. In the case of the jump process $T_{n-1} - T_n$ may not be constant unlike the special case of smart meters.

2) Calculation of the cumulative electricity consumption of a smart meter so far $\boldsymbol{t}$

$$Y_t = \sum\nolimits_{n=1}^{\infty} X_n \mathbb{1}_{T_n < t}$$

Y_t , the total electricity consumption up to time t.

Y_t is a jump process, X_n being the different values of the smart metering system.

B. *Temporal Aggregation of the Measurement*

1) Case of temporal aggregation of data from smart meters.

a) Temporal aggregation of data from a set of smart meters in the metering system at time t.

Let k be the index of any smart meter and Z_t^k be the electricity consumption at time t.

$Z_t^k = \sum_{n=1}^{\infty} X_n^k \mathbb{1}_{[T_{n-1}^k,T_n^k[}$ (t) , X_n^k being the electricity consumption of the smart meter k between T_{n-1}^k and T_n^k at time t.

Let N be the number of smart meters in the metering system and S_N be the total electricity consumption for all these meters at time t.

$S_N = \sum_{k=1}^{N} Z_t^k$ represents the data aggregation for all N smart meters at time t.

b) Temporal aggregation of data from a set of smart meters in the metering system up to the instant $\boldsymbol{t}$.

Let k be the index of any smart meter and Y_t^k be the total electricity consumption up to time t.

$$Y_t^k = \sum\nolimits_{n=1}^{\infty} X_n^k \mathbb{1}_{T_n^k < t}$$

Let N be the number of smart meters in the metering system and R_N be the total electricity consumption for all of these meters at time t.

$R_N = \sum_{k=1}^{N} Y_t^k$ represents the aggregation of data from all N smart meters up to the moment t.

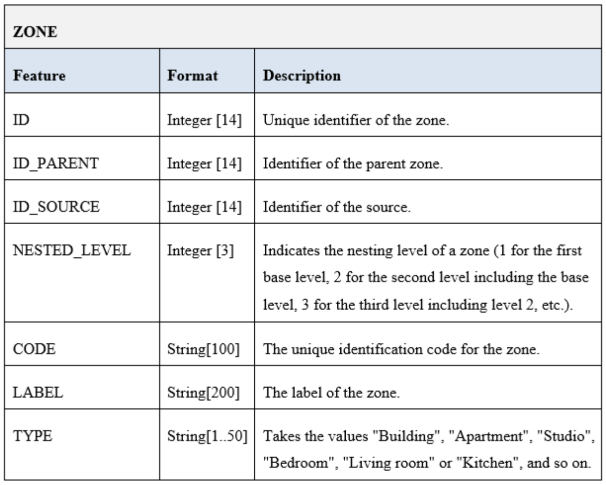
C. *Aggregation by Zone of Measurements*

This section is devoted to data aggregations in relation to zones. This step assumes that the temporal aggregations of the measurement data have already been carried out for a specific time step.

The basic data model does not allow a zone aggregation. For that, the notion of zone is necessary. The zone being a space in which one or more measurements can be counted. A zone is attached to a meter that also makes it possible to determine the customer.

It is also necessary to define the notion of parent zone. This notion is essential in the implementation of a neighborhood data model. It is also possible to locate a zone by its geographic coordinates. The difference between a zone of a residence and that of a neighborhood is made by adding the notion of parent zone for a neighborhood. Indeed, the basic data model is suitable for evaluating the power consumption of a residential zone (house, apartment, building, district, etc.).

The "ZONE" table takes into account the characteristics of a building and a district and is presented as follows in Fig. 4:



ZONE		
Feature	**Format**	**Description**
ID	Integer [14]	Unique identifier of the zone.
ID_PARENT	Integer [14]	Identifier of the parent zone.
ID_SOURCE	Integer [14]	Identifier of the source.
NESTED_LEVEL	Integer [3]	Indicates the nesting level of a zone (1 for the first base level, 2 for the second level including the base level, 3 for the third level including level 2, etc.).
CODE	String[100]	The unique identification code for the zone.
LABEL	String[200]	The label of the zone.
TYPE	String[1..50]	Takes the values "Building", "Apartment", "Studio", "Bedroom", "Living room" or "Kitchen", and so on.

Fig. 4. Representation of a Zone.

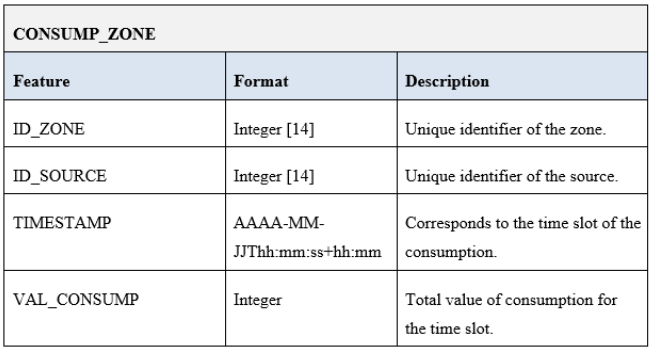
In evaluating the consumption of a building, it is necessary to cumulate the consumption of all the zones of that building. All of the building zones are made up of building zones that do not have a parent zone. Indeed, these zones already take into account the sub-zones, which compose them.

Similarly, the consumption of a district is determined by the cumulative consumption of all the zones of this district that do not have a parent zone (level 0 zone) then come the zones with only one parent (level 1 zone) And so on.

The notion of the “SOURCE” table is also essential for the transition from the basic data model to the district data model.

1) Presentation of zone aggregation structures: Below is the presentation of the “CONSUMP_ZONE” table as shown in Fig. 5. Calculating the consumption of all the zones will facilitate the restitution of the consumption of a building or even a district. Indeed, a building and a district can be considered as a set of independent or combined zones through which the measurements are distributed.

The identifier "ID_SOURCE" has been added to the table "CONSUMP_ZONE" to both take into account the size of the building and the district but also to avoid duplicating zones with the same identifiers and belonging to different sources.



CONSUMP_ZONE		
Feature	**Format**	**Description**
ID_ZONE	Integer [14]	Unique identifier of the zone.
ID_SOURCE	Integer [14]	Unique identifier of the source.
TIMESTAMP	AAAA-MM-JJThh:mm:ss+hh:mm	Corresponds to the time slot of the consumption.
VAL_CONSUMP	Integer	Total value of consumption for the time slot.

Fig. 5. Representation of Total Electricity Consumption over a Slot Time for a Zone.

2) Case of data aggregation by zone from smart meters.

a) Aggregation by zone of data from a set of smart meters in the metering system at the instant $\boldsymbol{t}$

Let k be the index of any smart meter and $Z_t^k(i)$ the electricity consumption at time t in zone i .

$Z_t^k(i) = \sum_{n=1}^{\infty} X_n^k(i)\,\mathbb{1}_{[T_{n-1}^k, T_n^k[}$ (t) , $X_n^k(i)$ being the electricity consumption of the smart meter k between T_{n-1}^k and T_n^k at time t in zone i.

Let N be the number of smart meters in the metering system and $S_N(i)$ the total electricity consumption for all these meters at time t in zone i.

$S_N(i) = \sum_{k=1}^{N} Z_t^k\,(i)$ represents the aggregation of data for all N smart meters at time t in zone i.

b) Aggregation by zone of data from a set of smart meters in the metering system up to time $\boldsymbol{t}$**.**

Let k be the index of any smart meter and $Y_t^k(i)$ the total electricity consumption up to time t in zone i.

$$Y_t^k(i) = \sum_{n=1}^{\infty} X_n^k(i)\,\mathbb{1}_{T_n^k < t}$$

Let N be the number of smart meters in the metering system and $R_N(i)$ the total electricity consumption for all of these meters at time t in zone i.

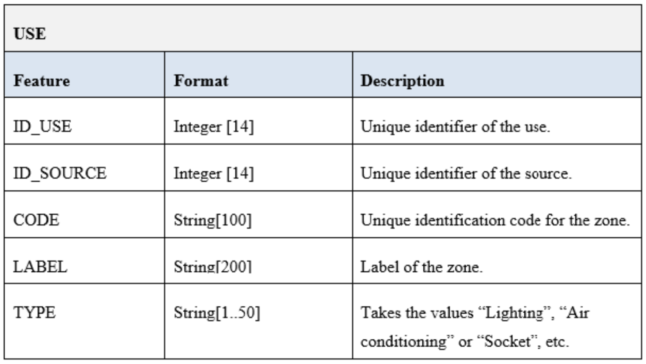
$R_N(i) = \sum_{k=1}^{N} Y_t^k(i)$ represents the aggregation of data from all N smart meters up to time t in zone i.

D. Aggregation by use of Measurements

In this section, the presentation of aggregations in relation to uses is highlighted. These aggregations are limited in space as they are attached to a zone, which can be a building or even a neighborhood. Indeed a district can be described as a set of nested zones or not.

1) Presentation of aggregation structures for an use; Measurement data can be matched to uses. Indeed, the consumption data from each measurement comes from a well-defined type of measure (lighting, air conditioning, a socket, etc).

Each source has a set of uses associated with it as displayed in Fig. 6. The “USE” table therefore takes a reference to the “SOURCE” table.

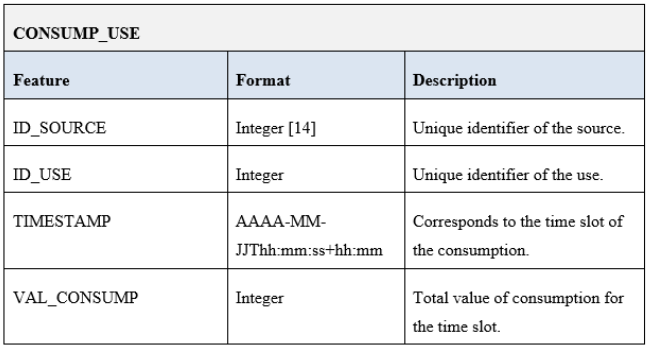


USE		
Feature	**Format**	**Description**
ID_USE	Integer [14]	Unique identifier of the use.
ID_SOURCE	Integer [14]	Unique identifier of the source.
CODE	String[100]	Unique identification code for the zone.
LABEL	String[200]	Label of the zone.
TYPE	String[1..50]	Takes the values “Lighting”, “Air conditioning” or “Socket”, etc.

Fig. 6. Representation for an use of Electricity.

Each measurement is associated with an use. A reference to the "SOURCE" table has therefore been added to the "USE" table, which makes it possible to identify the use for all measurements.

The use information is only define at the source level but not at the smart meter level. Aggregation by use for a building or a district requires the information of the field "ID_SOURCE at the level of the "USE" table to link each measurement to its source of correspondence as observed in Fig. 7. The "ID_SOURCE" field also makes it possible to resolve conflicts in the case of identical identifiers of measurements belonging to different sources.



CONSUMP_USE		
Feature	**Format**	**Description**
ID_SOURCE	Integer [14]	Unique identifier of the source.
ID_USE	Integer	Unique identifier of the use.
TIMESTAMP	AAAA-MM-JJThh:mm:ss+hh:mm	Corresponds to the time slot of the consumption.
VAL_CONSUMP	Integer	Total value of consumption for the time slot.

Fig. 7. Representation of Total Electricity Consumption over a Slot Time for an use.

2) Case of data aggregation by use from smart meters.

a) Aggregation by use of the data of a set of smart meters in the metering system at time t.

Let k be the index of any smart meter and $Z_t^k(j)$ the electricity consumption at time t for use j.

$Z_t^k(j) = \sum_{n=1}^{\infty} X_n^k(j)\,\mathbb{1}_{[T_{n-1}^k, T_n^k[}$ (t) , $X_n^k(j)$ being the electricity consumption of the smart meter k between T_{n-1}^k and T_n^k at time t for use j.

Let N be the number of smart meters in the metering system and $S_N(j)$ the total electricity consumption for all of these meters at time t for use.

$S_N(j) = \sum_{k=1}^{N} Z_t^k\,(j)$ represents the data aggregation of all N smart meters at time t for use j.

b) Aggregation by use of the data of a set of smart meters in the metering system up to time $\boldsymbol{t}$.

Let k be the index of any smart meter and $Y_t^k(j)$ the total electricity consumption up to time t for use j.

$$Y_t^k(j) = \sum_{n=1}^{\infty} X_n^k(j)\,\mathbb{1}_{T_n^k < t}$$

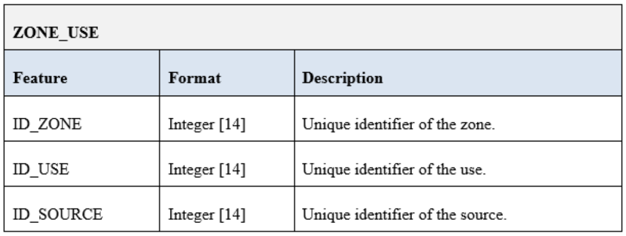
Let N be the number of smart meters in the metering system and $R_N(j)$ the total electricity consumption for all of these meters at time t for use j.

$R_N(j) = \sum_{k=1}^{N} Y_t^k(j)$ represents the aggregation of data from all N smart meters up to time t for use j.

E. Aggregation by Zone and by use of the Measurements

A set of uses constitute the link between a zone and its measurement data. For this purpose, the definition of the ZONE_USE table is necessary. This table contains the references of all uses and measurements attached to each zone.

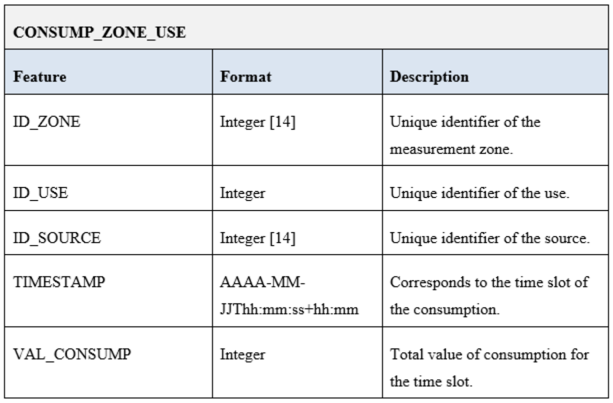
1) Presentation of aggregation structures for a zone and an use: In a zone containing sub-zones, it is necessary to group together all the uses according to the types of uses and sub-zones. This will make it possible to present the consumption of uses for a given zone as a whole but also to highlight the detail of the consumption of uses for a given zone according to its sub-zones. Fig. 8 shows the link between a zone and an use.



ZONE_USE		
Feature	**Format**	**Description**
ID_ZONE	Integer [14]	Unique identifier of the zone.
ID_USE	Integer [14]	Unique identifier of the use.
ID_SOURCE	Integer [14]	Unique identifier of the source.

Fig. 8. Representation of the Relationship for a Zone according to an Electricity use.

Fig. 9 is a view of the table containing consumption by zone and by use:



CONSUMP_ZONE_USE		
Feature	**Format**	**Description**
ID_ZONE	Integer [14]	Unique identifier of the measurement zone.
ID_USE	Integer	Unique identifier of the use.
ID_SOURCE	Integer [14]	Unique identifier of the source.
TIMESTAMP	AAAA-MM-JJThh:mm:ss+hh:mm	Corresponds to the time slot of the consumption.
VAL_CONSUMP	Integer	Total value of consumption for the time slot.

Fig. 9. Representation of the Relationship of Total Electricity Consumption over a Period for a Zone according to an Electricity use.

2) Case of data aggregation by zone and by use.

a) Aggregation by zone and by use of the data of a set of smart meters in the metering system at time t.

Let k be the index of any smart meter and $Z_t^k(i,j)$ the electricity consumption at time t in zone i and for use j.

$Z_t^k(i,j) = \sum_{n=1}^{\infty} X_n^k(i,j)\,\mathbb{1}_{[T_{n-1}^k, T_n^k[}$ (t) , $X_n^k(j)$ being the electricity consumption of the smart meter k between T_{n-1}^k and T_n^k at time t in zone i and for use j.

Let N be the number of smart meters in the metering system and $S_N(i,j)$ the total electricity consumption for all of these meters at time t in zone i and for use j.

$S_N(i,j) = \sum_{k=1}^{N} Z_t^k\,(i,j)$ represents the aggregation of data from all N smart meters at time t in zone i and for use j.

b) Aggregation by zone and by use of data from a set of smart meters in the metering system up to time t.

Let k be the index of any smart meter and $Y_t^k(i,j)$ the total electricity consumption up to time t in zone i and for use j.

$$Y_t^k(i,j) = \sum_{n=1}^{\infty} X_n^k(i,j)\,\mathbb{1}_{T_n^k < t}$$

Let N be the number of smart meters in the metering system and $R_N(j)$ the total electricity consumption for all of these meters at time t in zone i and for use j.

$R_N(i,j) = \sum_{k=1}^{N} Y_t^k(i,j)$ represents the aggregation of data from all N smart meters up to time t in zone i and for use j.

IV. Analysis of user Consumption Behavior

This section focuses on the analysis of the behavior of electricity consumption by users. To do this, A study is carried out on the effects of consumer's daily actions on the use of each device connected to its smart meter.

Let M be the number of devices connected to a given smart meter with index k. Then consider l the index of any device connected to this smart meter k.

Let T_l^n be the duration of consumption for device l in the interval $[T_{n-1}, T_n[$ and let C_l be the consumption per unit of time for this device.

The consumption of the device l in the interval $[T_{n-1}, T_n[$ can then be represented by $\mathrm{C}_l^n = \mathrm{C}_l \times T_l^n$.

The duration of consumption T_l^n is random because the interval $[T_{n-1}, T_n[$ may or may not occur, which makes it possible to deduce that T_l^n therefore follows a continuous law. In addition, the duration which separates two consecutive realizations in a series of independent realizations is modeled by an exponential law of parameter θ. T_l^n is therefore an exponential variable.

The consumption of electricity between T_{n-1} and T_n is independent of the consumption before T_{n-1} and that after T_n.

The duration between any two realizations in a series of independent realizations follows a gamma law with parameter (a, θ), a designates the number of intervals between the first realization and the last realization; and θ denotes the average number of realizations per unit of time. θ is estimated over time by following the consumption history linked to a smart meter.

If the duration of consumption T_l^n is performed only once in the interval $[T_{n-1}, T_n[$, this duration of consumption T_l^n follows an exponential law.

If the duration of consumption T_l^n is carried out several times in the interval $[T_{n-1}, T_n[$, T_l^n follows a gamma law.

$\mathrm{X}_n^k = \sum_{l=1}^{M} C_l^{n,k} = \sum_{l=1}^{M} C_l^k \times T_l^n$ with $C_l^{n,k}$ the consumption read by the meter k for the device l in the interval $[T_{n-1}, T_n[$ [and C_l^k the consumption per unit of time for the device l relatively to the characteristics of the smart meter k.

V. Discussion

The present work is part of an approach to design a formal framework for storing data from smart meters. This study also aims to present an approach for resizing large data from smart meters in order to facilitate their use. Likewise, the processing methods applied to data are presented.

This work shows the importance of data aggregation in the management of smart meters data and proposes a method of aggregating this data, based on the jump process. In addition, the storage structures for raw data and aggregated data are described as well as the mechanisms for recording and updating this data. An analysis of the behavior of electricity consumption is also carried out based on the actions of electricity consumers and the characteristics of the various electrical devices connected to their smart meters.

The research carried out did not make it possible to identify similar work on the aggregation of data from smart meters, on an attempt to present the structures of their databases and even less on the details of the processing applied to their data. It is indeed all these reasons that motivated and led to the completion of this work. This publication aims to promote an understanding of how smart meters work and to provide a basis for further research.

However, there is a great deal of research on the analysis of the behavior inherent in the consumption of electricity from data compiled by existing systems [8, 43]. The publication [8] provides an overview on algorithms and applications applicable to smart meter data. The article [43] shows that the individual predictability of user consumption can be determined with a high degree through time models of energy demand analysis. Likewise, the analysis of electricity consumption in the present publication is distinct from those encountered in similar studies including [8, 10, 12]. It is indeed atypical and is based on the determination of the mathematical law likely to explain the phenomenon of electricity consumption.

It is therefore necessary to underline that the work carried out within the framework of this publication covers several aspects of smart meters data management and provides contributions that can be explored in future publications.

VI. Conclusion

This work made it possible to define the framework for using data from smart meters. It was used to design a data model based on a comparative study of the functionalities of smart meters. The specificity of the data in a metering system has led to the implementation of data aggregation methods to facilitate data processing and analysis. The complexity of the data management of smart metering systems is highlighted by the proposal for a comprehensive approach that includes a concrete case of implementation.

The limitations of this work lie in the lack of an implementation of the metering system. Likewise, this does not make it possible to test the performance of the various algorithms put in place. However, issues relating to the implementation of an intelligent metering system as well as that of data processing are presented. The proposed design makes it possible to process the data collected at the level of a house and a neighborhood.

The originality of this work lies in the presentation of a mechanism for processing centralized data from smart meters, the aggregation of this data by the jump process and the determination of the law determining the behavior of the users on electricity consumption.

The lack of information on smart metering systems makes any comparative study difficult. Indeed, the implementation of these systems remains closed even if the functionalities are well known to the users. The future work will address the issues of factors influencing electricity consumption such as weather conditions, communication, billing, data correction and incident management on smart meters.

References

[1] Yazid Hambally Yacouba, Amadou Diabagaté, Abdou Maiga, Adama Coulibaly. Multi-agent system for management of data from electrical smart meters.

[2] Jaime Lloret, Jesus Tomas, Alejandro Canovas, Lorena Parra. An Integrated IoT Architecture for Smart Metering. Instituto de Investigación para la Gestión Integrada de zonas Costeras. Universidad Politécnica de Valencia, Spain.

[3] Xiufeng Liu, Per Sieverts Nielsen. A Hybrid ICT-Solution for Smart Meter Data Analytics. Energy, Volume 115, Part 3, 15 November 2016, Pages 1710-1722.

[4] Yi Wang, Qixin Chen, Tao Hong, Chongqing Kang. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid (Volume: 10 , Issue: 3 , May 2019).

[5] Iana Vassileva, Fredrik Wallin, Erik Dahlquist. Analytical comparison between electricity consumption and behavioral characteristics of Swedish households in rented apartments. Applied Energy Volume 90, Issue 1, February 2012, Pages 182-188.

[6] Ya Wu, Li Zhang. Evaluation of energy saving effects of tiered electricity pricing and investigation of the energy saving willingness of residents. Energy Policy Volume 109, October 2017, Pages 208-217.

[7] Xiaochen Zhang; Santiago Grijalva; Matthew J. Reno. A Time-Variant Load Model Based on Smart Meter Data Mining. 2014 IEEE PES General Meeting | Conference & Exposition, DOI: 10.1109/PESGM.2014.6939365, Electronic ISBN: 978-1-4799-6415-4.

[8] Xiufeng Liu; Lukasz Golab; Ihab F. Ilyas. SMAS: A smart meter data analytics system. 2015 IEEE 31st International Conference on Data Engineering, DOI: 10.1109/ICDE.2015.7113405, Electronic ISBN:978-1-4799-7964-6.

[9] Markus Weiss; Adrian Helfenstein; Friedemann Mattern; Thorsten Staake. Leveraging smart meter data to recognize home appliances. 2012 IEEE International Conference on Pervasive Computing and Communications, DOI: 10.1109/PerCom.2012.6199866.

[10] Ming Dong; Paulo C. M. Meira; Wilsun Xu; Walmir Freitas. An Event Window Based Load Monitoring Technique for Smart Meters. IEEE Transactions on Smart Grid (Volume: 3, Issue: 2, June 2012), DOI: 10.1109/TSG.2012.2185522.

[11] Xiufeng Liu, Per Sieverts Nielsen. A Hybrid ICT-Solution for Smart Meter Data Analytics. Energy, Volume 115, Part 3, 15 November 2016, Pages 1710-1722, https://doi.org/10.1016/j.energy.2016.05.068.

[12] Peng Xu, Jingchun Shen, Xingxing Zhang, Xudong Zhao , Yingchu Qian. Case Study of Smart Meter and In-home Display for Residential Behavior Change in Shanghai, China. Energy Procedia Volume 75, August 2015, Pages 2694-2699.

[13] Juan I. Guerrero, Antonio García, Enrique Personal, Joaquín Luque, Carlos León. Heterogeneous data source integration for smart grid ecosystems based on metadata mining. Expert Systems with Applications Volume 79, 15 August 2017, Pages 254-268.

[14] Jinsong Liu, Xiaolu Li, Member, IEEE, Dong Liu, Member, IEEE, Hesen Liu, and Peng Mao. Study on Data Management of Fundamental Model in Control Center for Smart Grid Operation. IEEE Transactions on Smart Grid (Volume: 2, Issue: 4, Dec. 2011), DOI: 10.1109/TSG.2011.2160571.

[15] Yi Wang, Qixin Chen, Chongqing Kang. Overview of Smart Meter Data Analytics. Springer Link, 25 February 2020.

[16] Yuxuan Wang ; Fengji Luo ; Zhaoyang Dong ; Ziyuan Tong ; Yichen Qiao. Distributed meter data aggregation framework based on Blockchain and homomorphic encryption. IET Cyber-Physical Systems: Theory & Applications (Volume: 4 , Issue: 1 , 3 2019), DOI: 10.1049/iet-cps.2018.5054.

[17] Mohamed Saleem Haja Nazmudeen ; Au Thien Wan ; Seyed M. Buhari. Improved throughput for Power Line Communication (PLC) for smart meters using fog computing based data aggregation approach. 2016 IEEE International Smart Cities Conference (ISC2), DOI: 10.1109/ISC2.2016.7580841.

[18] Zhenyu Guo ; Z. Jane Wang ; Ali Kashani. Home Appliance Load Modeling From Aggregated Smart Meter Data. IEEE Transactions on Power Systems (Volume: 30 , Issue: 1 , Jan. 2015).

[19] Xiaodi Wang, Yining Liu, Kim-Kwang Raymond Choo. Fault Tolerant Multi-subset Aggregation Scheme for Smart Grid. IEEE Transactions on Industrial Informatics, 05 August 2020, DOI: 10.1109/TII.2020.3014401.

[20] Toshichika Shiobara ; Peter Palensky ; Hiroaki Nishi. Effective metering data aggregation for smart grid communication infrastructure. IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society.

[21] J. Shanmugasundaram, K. Indhumathi, S. Sivaranjani. Deep Learning Approach cum Aggregated Smart Meter Data Based Residential Energy Load Modeling. Iconic Research and Engineering journals, february 2020, volume 3, issue 8, issn: 2456-8880.

[22] Anastasia Ushakova, Slava Jankin Mikhaylov. Big data to the rescue? Challenges in analysing granular household electricity consumption in the United Kingdom. Energy Research & Social Science, Volume 64, June 2020, 101428.

[23] Souhaib Ben Taieb, James W. Taylor & Rob J. Hyndman. Hierarchical Probabilistic Forecasting of Electricity Demand With Smart Meter Data. Journal of the American Statistical Association, 30 Mar 2020.

[24] Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings, Volume 49, June 2012, Pages 591-603.

[25] Hamed Chitsaz, Hamid Shaker, Hamidreza Zareipour, David Wood, Nima Amjady. Short-term electricity load forecasting of buildings in microgrids. Energy and Buildings, Volume 99, 15 July 2015, Pages 50-60.

[26] Pedro Gouveia, Júlia Seixas, Ana Mestre. Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling. Energy, Volume 141, 15 December 2017, Pages 108-122.

[27] Ilze Laicane, Dagnija Blumberga, Andra Blumberga, Marika Rosa. Evaluation of Household Electricity Savings. Analysis of Household Electricity Demand Profile and User Activities. Energy Procedia, Volume 72, June 2015, Pages 285-292.

[28] Anna Carolina Menezes, Andrew Cripps, Dino Bouchlaghem, Richard Buswell. Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. Applied Energy Volume 97, September 2012, Pages 355-364.

[29] Z. Song, X. Zhang, C. Eriksson. Data Center Energy and Cost Saving Evaluation. Energy Procedia 75 (2015) 1255 – 1260.

[30] Tamar Krishnamurti, Daniel Schwartz, Alexander Davis, Baruch Fischhoff, Wändi Bruine de Bruin, Lester Lave, Jack Wang. Preparing for smart grid technologies: A behavioral decision research approach to understanding consumer expectations about smart meters. Energy Policy, Volume 41, February 2012, Pages 790-797.

[31] Jacopo Torriti. People or machines? Assessing the impacts of smart meters and load controllers in Italian office spaces. Energy for Sustainable Development Volume 20, June 2014, Pages 86-91.

[32] Gordon Rausser, Wadim Strielkowski and Dalia Streimikien. Smart meters and household electricity consumption: A case study in Ireland. Energy & Environment, November 22, 2017, DOI: 10.1177/0958305X17741385.

[33] I. Laicāne, A. Blumberga, M. Rosa and D. Blumberga. Assessment of changes in households' electricity consumption. Agronomy Research 11 (2), 335–346, 2013.

[34] Slavisa Aleksic, Vedad Mujan. Exergy-based life cycle assessment of smart meters. IEEE, 14 July 2016, DOI: 10.1109/ELEKTRO.2016.7512075.

[35] Christian Beckel, Leyna Sadamori, Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. e-Energy '13: Proceedings of the fourth international conference on Future energy systems, May 2013, Pages 75–86, https://doi.org/10.1145/2487166.2487175.

[36] Bruna Di Silvio, Vittorio Cesarotti, Vito Introna . Evaluation of electricity rates through characterization and forecasting of energy consumption: A case study of an Italian industrial eligible customer. International Journal of Energy Sector Management, ISSN: 1750-6220, Publication date: 22 May 2007.

[37] A K Imtiaz; Norman B Mariun ; M M R Amran ; M Saleem ; N I A Wahab ; Mohibullah. Evaluation and Forecasting of Long Term Electricity Consumption Demand for Malaysia by Statistical Analysis. 2006 IEEE International Power and Energy Conference, DOI: 10.1109/PECON.2006.346658.

[38] V. F. Kadankov and T. V. Kadankova. on the Distribution of the Time of the First Exit From an Interval and the Value of a Jump Over the Boundary for Processes with Independent Increments and Random Walks. Ukrainian Mathematical Journal, Vol. 57, No. 10, 2005.

[39] Clifford A. Ball and Walter N. Torous. A Simplified Jump Process for Common Stock Returns. The Journal of Financial and Quantitative Analysis , Mar., 1983, Vol. 18, No. 1 (Mar., 1983), pp. 53-65.

[40] Yong Li, Depeng Jin, Zhaocheng Wang, Pan Hui, Lieguang Zeng, and Sheng Chen. A Markov Jump Process Model for Urban Vehicular Mobility: Modeling and Applications. IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 13, NO. 9, SEPTEMBER 2014.

[41] Lydia Chabane, Raphaël Chétrite and Gatien Verley. Periodically driven jump processes conditioned on large deviations. Laboratoire de Physique Théorique (UMR8627), CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France, Laboratoire J A Dieudonné, UMR CNRS 7351, Université de Nice Sophia Antipolis, Nice 06108, France (Dated: October 30, 2019).

[42] Alexander Sikorski, Marcus Weber, Christof Schütte. The Augmented Jump Chain - a sparse representation of time-dependent Markov jump processes. Zuse Institute Berlin, November 2020.

[43] Adrian Albert, Ram Rajagopal. Smart Meter Driven Segmentation: What Your Consumption Says About You. IEEE Transactions on Power Systems (Volume: 28, Issue: 4, Nov. 2013).

Assessing Data Sharing's Model Fitness Towards Open Data by using Pooled CFA

Siti Nur’asyiqin Ismael[1], Othman Mohd[2], Yahaya Abd Rahim[3]
Faculty of Information and Communication Technology
UTeM, Melaka, Malaysia

***Abstract*—This study demonstrates the step-by-step procedure to perform Pooled Confirmatory Factor Analysis (CFA) in the measurement part of Structural Equation Modelling (SEM). CFA is crucial for the SEM measurement model to obtain the acceptable model fit before modeling the structural model. There are two techniques in CFA; individual CFA and Pooled-CFA. Usually, Pooled-CFA is done due to the high number of constructs and items. If the model is too complicated and has so many constructs and items, then it is recommended to perform Pooled-CFA to simplify the model's looks yet easy to understand. The perception of Malaysia Technical University Network (MTUN) academics on data sharing towards open data was analysed by using pooled-CFA. There are three main constructs: data sharing with its 4 sub-constructs; (technological factor, organizational factor, environmental factor, and individual factor), mediator construct (open data licenses), and open data construct was analyzed in this research. Furthermore, second-order constructs' factor loadings towards their corresponding sub-constructs were investigated. This research collected the primary data of 442 respondents using a stratified random sampling technique. This paper will explain the theoretical framework before revealing the results of Pooled-CFA on data sharing towards open data.**

***Keywords*—Pooled CFA; data sharing; open data; measurement model; validity**

I. Introduction

Open data initiatives have become ubiquitous in every country. According to [1], Malaysia has embarked on the open government data framework by The Malaysian Administrative Modernisation and Management Planning Unit (MAMPU) in the year 2015. The initiative is then leveraged to be implemented at the ministries' and agencies' levels. It seems crucial to have an open data framework within the higher education environment as [2] has mentioned that the higher institutions play a significant role and are among the most significant contributors that support the citizen's needs in the education world. In the meantime, [3] has stated that the data producer is reluctant to share data might because it possesses challenges at many levels such as cultural, ethical, financial, and technical. Adding to these challenges, [4] has highlighted that the reluctance of data sharing perhaps due to disinterest from the universities. Thus, this study employs quantitative techniques; survey to Malaysia Technical University Network (MTUN) academics. There were 442 feedbacks received and there was a need to perform Confirmatory Factor Analysis (CFA) to confirm the factor that influence MTUN academics on data sharing.

This research aims to identify the factor influence MTUN academics on data sharing and analysing the open data license, which will act as a mediator between data sharing and open data. This paper will explain in details the theoretical framework developed for this research, the component in structural equation modeling (SEM), the determination of sample size, the fitness indexes of technological, organizational, environmental, and individual construct that determine data sharing and how the procedure of pooled-CFA is done. The reliability and validity indexes also will be measured to indicate the acceptance state.

II. Literature Review

Open data might change the relationship between the government and the public in terms of transparency [5]. This intention can be perceived by accessing the government's data through an open format datasets form. Furthermore, [6] has emphasized that open data enactment will address the existing legal challenges. The challenges include the scope of accessing the data and data ownership.

Since open data has been announced, it has created a sensation worldwide. In [7], the statement is supported by highlighting the potential of open data to improve organization services to the public. Besides that, citizen participation is encouraged in open data towards having a transparent government. As in [8], the approach to embark on open data will be different as every country has a different governance structure, and the organization has its policies regarding open data.

Ministry of Education Malaysia (MOE) has specified that Malaysia's Higher Education Institutions (HEIs) are categorized as Public Universities, Private Higher Educational Institutions, Polytechnics, and Community Colleges [9]. As for this research's scope, MTUN universities include 4 public universities (UMP, UTEM, UTHM, and UNIMAP) where it focuses on the Technical and Vocational Education and Training (TVET) approaches.

In Malaysia, according to [9], 12 National Key Economic Areas (NKEA) have been identified under the government's Economic Transformation Programme (ETP). In [10], it is highlighted that the programme has demanded an additional 1.3 million TVET workers by 2020. This demand has strengthened the need to have MTUN open data framework as this initiative will help the potential worker make an informed decision from the data shared.

As in [11], the open data demand keeps growing in public universities. The demand ensues due to the open data capabilities of removing the barriers to reuse and redistributing the data. On the other hand, it will help the public to make informed decisions. As the results in [2], the economy's deception can be reduced, and universities' accountability can be expected while embarking on open data. To conclude, the data shared are valuable in creating innovations towards having a better university in the future.

This research endeavors to develop an open data framework for MTUN academics. Before that, the factors that influence data sharing, the roles of open data license as a mediator, and the indexes of open data components are measured. This paper will explain in detail the results of factors that influence data sharing towards open data by running pooled-CFA.

III. Theoretical Framework

This study integrates the technological, organizational, and environmental (TOE) framework with the Theory of Planned Behavior (TPB) to determine factors that influence data sharing. According to [12], The TOE framework is an organization-level theory that explains organization structures from 3 perspectives. These 3 perspectives are technological, organizational, and environmental. These contexts were adopted and integrate with TPB theory that examines individuals' perspectives and were analysed as the factors that contribute to the data sharing.

The TPB theory has been useful and considered one of the most influential models in predicting social behaviors [13], [14]. The TOE framework and TPB theory integration were used to develop the MTUN open data theoretical framework as Fig. 1.

Fig. 1 shows how the technological, organizational, and environmental sub-constructs are derived from the TOE framework. Whereby the individual sub construct is derived from TPB theory. From the framework, it can be seen that all of these 4 sub-constructs (technological, organizational, environmental, and individual) are the factors that contribute to the data sharing construct. The open data licenses (ODL) will act as the mediator between data sharing and open data (MTUN_OD). This paper will explain how the pooled-CFA is conducted and how the constructs and items were analyzed by using IBM SPSS AMOS software (version 2.4).

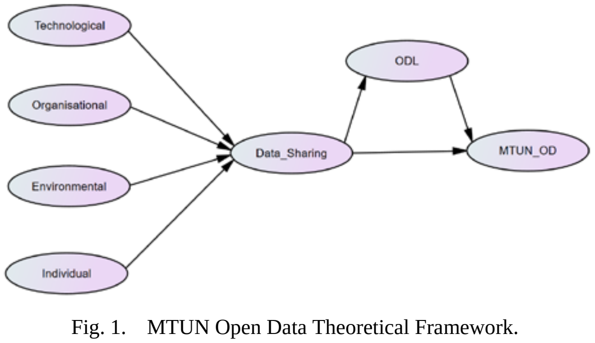


Fig. 1. MTUN Open Data Theoretical Framework.

IV. Structural Equation Modelling

Structural equation modeling (SEM) is a powerful, multivariate technique that has been used widely in scientific investigations to assess and evaluate multivariate causal relationships. According to [15], [16], sometimes, it is also called a statistical methodology where the confirmatory approach is used to analyse the structural theory.

There are 2 main components in SEM: measurement and structural models. These 2 components are used to examine variables in different ways. The measurement model section will relate the measured variables to the latent variables. On the other hand, the structural model section relates the latent variables to one another. SEM combines 2 statistical methods; confirmatory factor analysis (CFA) and path analysis. There are 2 CFA techniques: Individual CFA and Pooled-CFA. This paper focused on the SEM measurement model, assessed through the CFA.

The first step to run pooled-CFA was having to perform the individual CFA for each construct. In [17], [18] has mentioned that the second-order construct was validated using the CFA procedure separately before it’s been simplified into first-order constructs to reduce the model's complexity. As [19] suggested, the pooled-CFA for all constructs was important to perform to assess the discriminant validity among the model's constructs. Thus, this study analyzed the feedback from the MTUN academics survey of data sharing towards open data.

There was 1 main construct (second-order construct) involved in this study: data sharing with its sub-constructs: technological factor, organizational factor, environmental factor, and individual factor. According to [20], that second-order CFA is employed in this study as it involved the assessment of a second-order variable's factor loadings towards its corresponding sub-constructs. By running a second-order CFA, the relationships of data sharing towards its sub-constructs were examined as well. The ODL construct and MTUN_OD construct were identified as first-order constructs as they do not have the sub-construct and was analyzed directly without the need to simplifying it anymore.

V. Material and Method

A. Preface

This study's population target covers MTUN academics from various educational backgrounds and working experiences. Based on [21], the total population of MTUN academics in 2018 was 3818. According to [22], as Table I, the sample needed for this study was 351 for the population of 4000.

As per shown in Table I, the population sample obtained was 442, which was higher than the number of samples required. A total of 442 respondents were chosen randomly. The comprehensive questionnaire for the field study was constructed which was derived from an exploratory factor analysis (EFA). The EFA was executed by using IBM SPSS software. The data collection for the field study is done by distributing the questionnaire to MTUN academics using a stratified random sampling technique.

TABLE I. DETERMINING SAMPLE SIZE FOR A FINITE POPULATION

N	*S*	*N*	*S*	*N*	*S*
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	357

N = number of population
S = Sample Size

IBM statistical package for social science (SPSS) and IBM SPSS analysis of moment structures (AMOS) version 24.0 were used to build and analyze the model in this study.

B. *Confirmatory Factor Analysis*

Confirmatory factor analysis (CFA) is a method of factor analysis, most commonly used in social research. It is usually used to examine the consistency of a construct with a researcher's understanding of that construct's factor. CFA's objective is to examine whether the data fit a hypothesized measurement model. This hypothesized model is based on theory or previous analytic research. In CFA, several things need to be tested: reliability, validity, and unidimensionality of the measurement model. The results must meet the stated requirement before modeling the structural model. According to [18], [23], the theorized model must pass 3 types of validities: Construct Validity, Convergent Validity, and Discriminant Validity. The details of validity and reliability indexes are shown in Table II.

Table II shows the validity and reliability test that need to be passed. The construct validity is assessed through the fitness indexes of the measurement model. The convergent validity is assessed by computing the Average Variance Extracted (AVE). The Discriminant Validity is evaluated by developing the Discriminant Validity Index Summary.

Adding to this, several fitness indexes need to be examined as well to evaluate the model fitness. Absolute fit, incremental fit, and parsimonious fit are three types of model fit categories. Below are the fitness of indexes as shown in Table III.

As shown in Table III, [24] has mentioned that the names of indexes that are frequently reported in many research are Root Mean Square Error Approx (RMSEA), Comparative Fit Index (CFI), and Chi-square/degrees of freedom (Chisq/df).

C. *Discriminant Validity*

The discriminant validity needs to be assessed to ensure no construct redundancy occurs in the model. Construct redundancy might occur when any pair of constructs in the model are highly correlated. This redundancy also can happen when one or more constructs assess the same variable. In other words, discriminant validity tests whether the concepts of measurements that are not supposed to be related are unrelated. According to [24], if the redundancy occurs, that particular redundant items in a model need to be deleted. The deletion should start from the lowest value of factor loading until the model is fit.

Besides that, correlation coefficients are used to measure the strength of the relationship between 2 variables. As mentioned in [25], it is also acted as evidence of discriminant validity. A correlation between variables indicates that if one variable changes in value, the other variable tends to change in a specific direction. The variables should not be highly correlated to each other, or else the multi-collinearity problem will exist. Besides, [24] has highlighted that the correlation value among the exogenous variables should not exceed 0.85 to achieve the variables' discriminant validity.

D. *Summary*

There are 2 techniques of CFA in SEM's measurement model: Individual CFA and Pooled-CFA. Individual CFA runs each unobserved construct in the research individually; whereas Pooled-CFA runs all construct simultaneously [26]. Before performing Pooled-CFA, the individual CFA for all constructs need to be done separately. The results must achieve the indexes' fitness as Table II and Table III to make them reliable and validated. The AVE's results were recalculated to get the mean score and were used in Pooled-CFA.

TABLE II. VALIDITY AND RELIABILITY INDEXES

Name of Category	**Name of Index**	**Level of Acceptance**	**Literature**
Convergent Validity	Average Variance Extracted	AVE ≥ 0.5	Zainudin (2015)
Internal Reliability	Cronbach Alpha	A ≥ 0.5	Zainudin (2015)
Construct Reliability	Composite Reliability	CR ≥ 0.6	Zainudin (2015)

TABLE III. FITNESS OF INDEXES

Name of category	**Name of index**	**Level of acceptance**
Absolute Fit Index	RMSEA	RMSEA < 0.1
	GFI	GFI > 0.90
Incremental Fit Index	AGFI	AGFI > 0.90
	CFI	CFI > 0.90
	TLI	TLI > 0.90
	NFI	NFI > 0.90
Parsimonious Fit Index	Chi-sq/df	Chi-Square/ df < 5.0

VI. Results and Discussions

A. Individual CFA

The analysis started with performing Individual CFA. It ran the latent construct one after another to achieve the required model fitness. The CFA can only be performed if the constructs have more than 3 items with no model identification problem. Fig. 2 shows that all these 4 constructs (technological, organizational, environmental, and individual) have met the initial requirement to run CFA. All of the constructs must achieve the fitness indexes required.

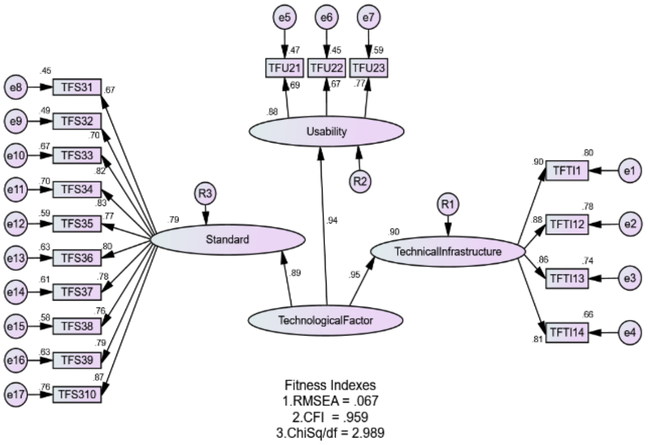


Fig. 2. The CFA Results for Technological Factor Construct.

Fig. 2 shows that the technological factor construct has 3 components; technical infrastructure (4 items), usability (3 items), and standard (10 items). The model fitness of the technological factor construct was overall met the fitness indexes. The value for RMSEA shown was .067, the CFI was .959, and Chisq/df was 2.989.

For the Convergent Validity (CV) assessment, the study needs to calculate the AVE. According to [19], [26], the construct achieved the CV if its AVE exceeds the threshold value of 0.5. Besides in [24], there was a need to compute the CR, and the value should exceed the threshold value of 0.6 for this reliability to achieve. The AVE and CR for the primary constructs and their respective components were computed and presented in Table IV.

Table IV shows that each item's factor loading was high, which above 0.6. The CR value for the technological factor was 0.948, and AVE was 0.859. Meanwhile, the CR value for technical infrastructure was 0.921 and AVE was 0.745. In addition to that, the CR value for usability was 0.754 and AVE was 0.506. Meanwhile, the CR value for a standard was 0.940, and AVE was 0.610.

From these results, we can conclude that technological factors construct together with its components and items have met the CR's requirement, which must above 0.6, and AVE, which must above 5.0. Fig. 3 shows the CFA results for the organizational factor construct.

The organizational factor construct has 4 components; norms (10 items), data sharing policy (3 items), governance (3 items), and resources (5 items). The model fitness of organizational factor constructs was overall met the fitness indexes. The value for RMSEA shown was .067, the CFI was .945, and Chisq/df was 2.988. Table V shows the AVE and CR for the organizational factor construct.

TABLE IV. The AVE and CR for Technological Construct

Construct	Items	Factor Loading	CR (>0.6)	AVE (>0.5)
Technological Factor	Technical Infrastructure	0.95	0.948	0.859
	Usability	0.94		
	Standard	0.89		
Technical Infrastructure	TFTI1	0.90	0.921	0.745
	TFTI2	0.88		
	TFTI3	0.86		
	TFTI4	0.81		
Usability	TFU21	0.69	0.754	0.506
	TFU22	0.67		
	TFU23	0.77		
Standard	TFS31	0.67	0.940	0.610
	TFS32	0.70		
	TFS33	0.82		
	TFS34	0.83		
	TFS35	0.77		
	TFS36	0.80		
	TFS37	0.78		
	TFS38	0.76		
	TFS39	0.79		
	TFS310	0.87		

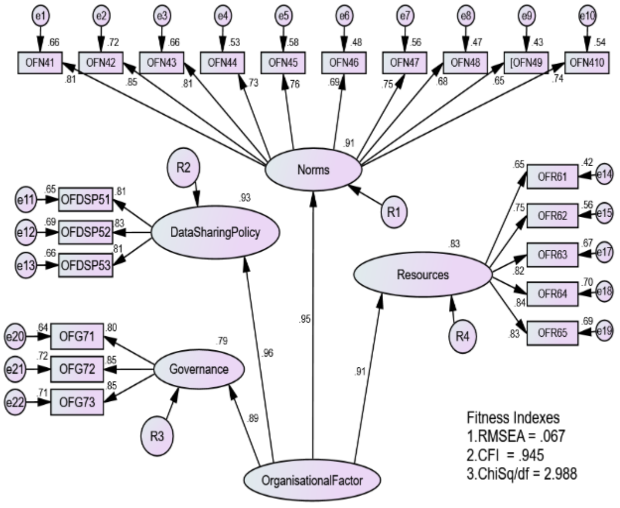


Fig. 3. The CFA Results for Organisational Factor Construct.

TABLE V. THE AVE AND CR FOR ORGANISATIONAL CONSTRUCT

Construct	**Items**	**Factor Loading**	**CR (>0.6)**	**AVE (>0.5)**
Organisational Factor	Norms	0.95	0.961	0.861
	Data sharing policy	0.96		
	Governance	0.89		
	Resources	0.91		
Norms	OFN41	0.81	0.927	0.562
	OFN42	0.85		
	OFN43	0.81		
	OFN44	0.73		
	OFN45	0.76		
	OFN46	0.69		
	OFN47	0.75		
	OFN48	0.68		
	OFN49	0.65		
	OFN410	0.74		
Data sharing policy	OFDSP51	0.81	0.857	0.667
	OFDSP52	0.83		
	OFDSP53	0.81		
Resources	OFR61	0.65	0.886	0.610
	OFR62	0.75		
	OFR63	0.82		
	OFR64	0.84		
	OFR65	0.83		
Governance	OFG71	0.80	0.872	0.695
	OFG72	0.85		
	OFG73	0.85		

Based on Table V, it can be concluded that the CR value for the organizational factor was 0.961, and AVE was 0.861. The CR value for norms was 0.927, and AVE was 0.562. The CR value for the data sharing policy was 0.857, and AVE was 0.667. Meanwhile, The CR value for resources was 0.886, and AVE was 0.610. Finally, the CR value for governance was 0.872 and AVE was 0.695.

From these results, it can be concluded that the organizational factor constructs and their components and items have met CR's requirement, which must above 0.6, and AVE, which must above 5.0. Fig. 4 shows the CFA results for the environmental factor construct.

The environmental factor construct has 2 components; data sharing culture (3 items) and research practice (3 items). The model fitness of the environmental factor construct was overall meet the fitness indexes. The value for RMSEA shown was .066, the CFI was .990, and Chisq/df was 2.925. Table VI shows the AVE and CR for the environmental factor construct.

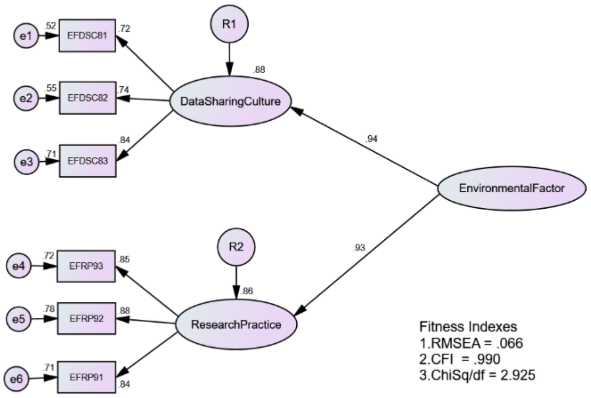


Fig. 4. The CFA Results for Environmental Factor Construct.

TABLE VI. THE AVE AND CR FOR ENVIRONMENTAL CONSTRUCT

Construct	**Items**	**Factor Loading**	**CR (>0.6)**	**AVE (>0.5)**
Environmental Factor	Data sharing culture	0.94	0.933	0.874
	Research Practice	0.93		
Data sharing culture	EFDSC81	0.72	0.812	0.591
	EFDSC82	0.74		
	EFDSC83	0.84		
Research practice	EFRP91	0.84	0.892	0.734
	EFRP92	0.88		
	EFRP93	0.85		

Table VI shows that each item's factor loading was high, which above 0.6. The CR value for the environmental factor was 0.933, and AVE was 0.874. The CR value for data sharing culture was 0.812, and AVE was 0.591. The CR value for research practice was 0.892, and AVE was 0.734. It can be concluded from these results that the environmental factor construct and its components and items have met CR requirements, which must above 0.6 and AVE must above 5.0. Fig. 5 shows the CFA results for the individual factor construct.

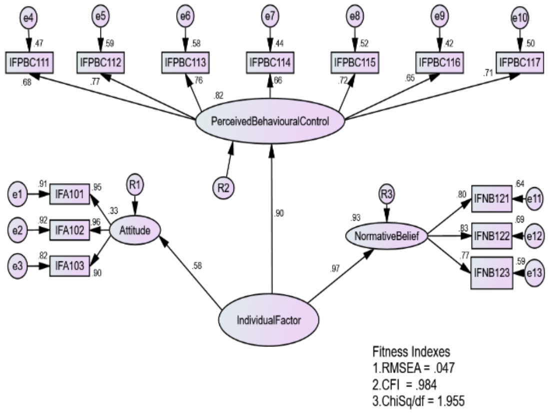


Fig. 5. The CFA Results for Individual Factor Construct.

The individual factor construct has 3 components, which are attitude (3 items), perceived behavioral control (7 items), and normative belief (3 items). The model fitness of the individual factor construct was overall met the fitness indexes. The value for RMSEA shown was .047, the CFI was .984, and Chisq/df was 1.955. Table VII shows the AVE and CR for the individual factor construct.

Table VII shows each item's factor loading was high above 0.6. The CR value for the individual factor was 0.868, and AVE was 0.696. Meanwhile, the CR value for attitude was 0.956, and AVE was 0.878. In addition to that, the CR value for perceived behavioral control was 0.875 and AVE was 0.502. The CR value for normative belief was 0.842, and AVE was 0.641. From these results, we can conclude that individual factors construct together with its components and items have met CR's requirement, which must above 0.6, and AVE, which must above 5.0. Fig. 6 shows the CFA results for the open data license factor construct.

In Fig. 6, the ODL construct has 5 items. Thus, the model fitness of the ODL construct was overall met the fitness indexes. The value for RMSEA shown was .053, the CFI was .995, and Chisq/df was 2.257. Table VIII shows the AVE and CR for ODL construct.

Table VIII shows that each item's factor loading was high, which above 0.6. The CR value for ODL was 0.898, and AVE was 0.639. Thus, from these results, we can conclude that ODL construct and items have met CR requirements that must above 0.6 and AVE, which must above 5.0. Fig. 7 shows the CFA results for the open data (MTUN_OD) construct.

TABLE VII. THE AVE AND CR FOR INDIVIDUAL FACTOR CONSTRUCT

Construct	Items	Factor Loading	CR (>0.6)	AVE (>0.5)
Individual Factor	Attitude	0.58	0.868	0.696
	Perceived Behavioral Control	0.90		
	Normative Belief	0.97		
Attitude	IFA101	0.95	0.956	0.878
	IFA102	0.96		
	IFA103	0.90		
Perceived Behavioral Control	IFPBC111	0.68	0.875	0.502
	IFPBC112	0.77		
	IFPBC113	0.76		
	IFPBC114	0.66		
	IFPBC115	0.72		
	IFPBC116	0.65		
	IFPBC117	0.71		
Normative Belief	IFNB121	0.80	0.842	0.641
	IFNB122	0.83		
	IFNB123	0.77		

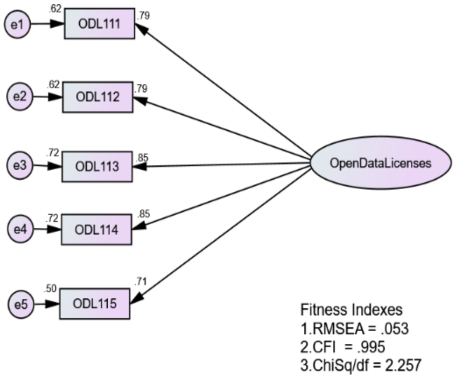


Fig. 6. The CFA Results for Open Data Licenses Factor Construct.

TABLE VIII. THE AVE AND CR FOR OPEN DATA LICENSES CONSTRUCT

Construct	Items	Factor Loading	CR (>0.6)	AVE (>0.5)
Open Data Licenses	ODL111	0.79	0.898	0.639
	ODL112	0.79		
	ODL113	0.85		
	ODL114	0.85		
	ODL115	0.71		

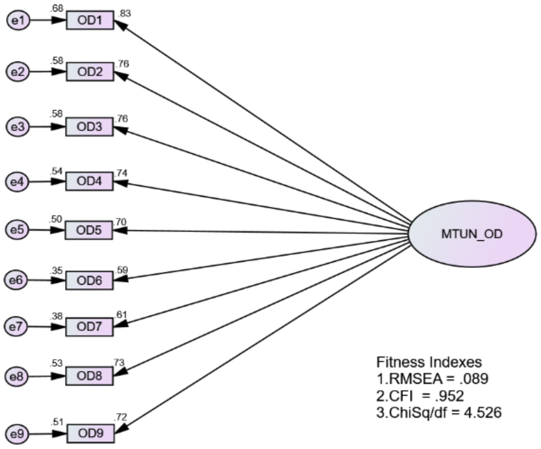


Fig. 7. The CFA Results for Open Data Construct.

In Fig. 7, the MTUN_OD construct has 9 items. The model fitness of the MTUN_OD construct was overall met the fitness indexes. The value for RMSEA shown was .089, the CFI was .952, and Chisq/df was 4.526. Table IX shows the AVE and CR for the open data construct.

Table IX shows each item's factor loading was high, above 0.6. The CR value for the open data construct was 0.909, and AVE was 0.528. Thus, from this result, we can conclude that MTUN_OD constructs and their items have met CR's requirement, which above 0.6, and AVE must above 5.0.

An overall, the technological construct, organizational construct, environmental construct, individual construct, ODL construct, and MTUN_OD construct has met the fitness indexes and passed the measurement of AVE and CR.

TABLE IX. THE AVE AND CR FOR OPEN DATA CONSTRUCT

Construct	Items	Factor Loading	CR (>0.6)	AVE (>0.5)
MTUN_OD	OD1	0.83	0.909	0.528
	OD2	0.76		
	OD3	0.77		
	OD4	0.75		
	OD5	0.71		
	OD6	0.60		
	OD7	0.63		
	OD8	0.74		
	OD9	0.72		

The study needed to simplify the overall measurement model from the first-order construct and pool them together to undergo the CFA procedure at once. This procedure is called Pooled-CFA.

B. Pooled-CFA for all Measurement Model of Constructs

In this pooled-CFA model, as suggested in [24], the measurement model for the second-order constructs was validated using the CFA procedure separately and simplified into first-order constructs to reduce complexity. The reason to perform the pooled-CFA was to assess the discriminant validity among constructs in the model [17], [19], [23], [26]. Fig. 8 shows the pooled-CFA that consists of data sharing (DS) construct, ODL, and MTUN_OD in 1 model.

The result of the analysis shown 3 types of values; fitness indexes for all constructs in the model, the factor loading for every component to the main construct, and the correlation between constructs.

In determining the fitness indexes, the values should meet the threshold as shown in Table III. As in [17], [19], [23], [26] have highlighted that the factor loading for every item should not less than 0.6 and the correlation coefficient of any two constructs should not exceed 0.85. The multicollinearity problem will occur if the correlation between any two constructs exceeds 0.85. In this study, none of the values found to be greater than 0.85. Thus, the multicollinearity problem does not arise.

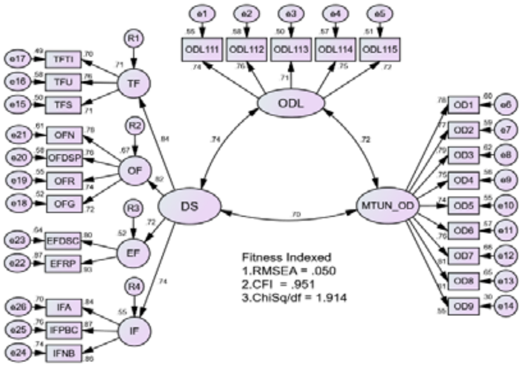


Fig. 8. The 3 Constructs are Pooled Together for the Pooled-CFA Procedure.

The Pooled CFA has merged 3 constructs. From Fig. 8, the model looks much more straightforward and easy to understand. The pooled CFA was also performed to avoid violating regression assumptions. The correlation between DS to ODL was 0.74. Then, the correlation between ODL to MTUN_OD was 0.72, and the correlation between DS to MTUN_OD was 0.70. Thus, no multicollinearity occurs as the correlation between each construct was below 0.85. Besides, pooled CFA's model fitness was overall met the fitness indexes. The value for RMSEA shown was .050, the CFI was .951, and Chisq/df was 1.914. Table X shows the AVE and CR for Pooled-CFA.

TABLE X. THE AVE AND CR FOR POOLED-CFA

Construct	Items	Factor Loading	CR (>0.6)	AVE (>0.5)
Data Sharing	Technological Factor	0.84	0.862	0.611
	Organizational Factor	0.82		
	Environmental Factor	0.72		
	Individual Factor	0.74		
Technological Factor	Technical Infrastructure	0.70	0.767	0.524
	Usability	0.76		
	Standard	0.71		
Organisational Factor	Norms	0.78	0.837	0.563
	Data Sharing Policy	0.76		
	Resources	0.74		
	Governance	0.72		
Environmental Factor	Data Sharing Culture	0.80	0.858	0.752
	Research Practice	0.93		
Individual Factor	Attitude	0.84	0.850	0.655
	Perceived Behavioural Control	0.87		
	Normative Beliefs	0.86		
ODL	ODL111	0.74	0.855	0.542
	ODL112	0.76		
	ODL113	0.71		
	ODL114	0.75		
	ODL115	0.72		
MTUN_OD	OD1	0.78	0.922	0.570
	OD2	0.77		
	OD3	0.79		
	OD4	0.75		
	OD5	0.74		
	OD6	0.76		
	OD7	0.81		
	OD8	0.81		
	OD9	0.55		

Based on the AVE and CR values in Table X, the study found that all AVE and CR exceed their threshold values of 0.5 and 0.6 respectively. Thus, the study can conclude that the AVE and CR for all latent constructs in the model have been achieved.

Proceed to the next validity test, the study needs to measure discriminant validity. The table discriminant validity index summary is developed as shown in Table XI. The diagonal values in bold were the square root of the AVE of the respective constructs while the other values indicate the correlation coefficient between the pair of the individual constructs.

TABLE XI. The Discriminant Validity Index Summary for All Constructs

Construct	Data Sharing	ODL	MTUN_OD
Data Sharing	**0.782**		
ODL	0.74	**0.740**	
MTUN_OD	0.70	0.720	**0.755**

TABLE XII. The Assessment of Normality for All

Variables	Skew	CR	Kurtosis	CR
IFA	-.759	-5.886	1.261	4.892
IFPBC	-.616	-4.774	1.022	3.964
IFNB	-.809	-6.276	1.309	5.076
EFDSC	-.326	-2.531	-.213	-.827
EFRP	-.528	-4.096	.541	2.099
OFN	-.381	-2.958	.106	.412
OFDSP	-.429	-3.328	-.071	-.275
OFR	-.061	-.470	-.254	-.986
OFG	-.188	-1.460	-.339	-1.316
TFTI	-.247	-1.914	-.009	-.036
TFU	-.066	-.513	-.420	-1.628
TFS	.024	.188	-.144	-.560
OD9	-.483	-3.744	.690	2.676
OD8	-.353	-2.737	.201	.779
OD7	-.390	-3.022	-.052	-.202
OD6	-.154	-1.196	-.189	-.731
OD5	-.160	-1.242	-.054	-.209
OD4	-.069	-.532	-.422	-1.637
OD3	-.143	-1.113	-.288	-1.115
OD2	-.080	-.622	-.252	-.977
OD1	-.375	-2.907	.280	1.085
ODL115	-.115	-.889	-.158	-.611
ODL114	-.184	-1.424	.009	.035
ODL113	-.009	-.071	-.239	-.928
ODL112	-.265	-2.057	-.099	-.385
ODL111	-.035	-.272	-.285	-1.106
Multivariate			57.862	14.406

Table XI shows the discriminant validity index summary for all constructs. The discriminant validity has been achieved when the diagonal values (in bold) are higher than any other values in its row and column. Since SEM employs the parametric statistical approach of modeling, the study needs to assess all items' normality distribution measuring their respective constructs. According to [17], [19], [23], [24], [26], the value of skewness should fall within the range of -1.5 to 1.5 to make it normally distributed.

Table XII shows the values of skewness for all components in the model fell within the range between -1.5 and 1.5. It means that the distribution does not depart from normality and there were no outliers' data. Thus, the data distribution meets the normality distribution requirement for employing parametric statistical analysis in SEM.

VII. Conclusion

Data sharing in this study that was defined through the combination of technological, organizational, environmental, and individual components. The components were derived from the literature review. However, in this study, the exact components that form data sharing were investigated through the process of survey distribution to MTUN academics that were then be confirmed through CFA. The investigations were then extended to the ODL construct and MTUN_OD construct. As for this research purposes, this study examined the factor influence data sharing and the impact of data sharing on ODL construct and MTUN-OD construct.

In a conclusion, Pooled-CFA is recommended to perform on a complicated model in making it simpler to analyse and easy to understand. The model is considered complicated when it involves many second-order constructs and items. There are 3 important types of validities in this study; CR, Cronbach Alpha, and AVE. The CR is important in this study as its measure of internal consistency in scale items. On the other hand, the AVE is important to employ in this study to confirm that the construct should correlate with related variables but it should not correlate with dissimilar, unrelated ones. In determining the value of CR and AVE for each construct, the analysis results of Pooled-CFA are recalled and it can be concluded that 4 components influence data sharing; technological, organizational, environmental, individual construct.

The technical factor has a CR value of 0.948; which above the minimum accepted value of CR; 0.6 and 0.859; which above the minimum accepted value for AVE; 0.5. The fitness indexed for this construct was achieved with the value of RMSEA was 0.067, which less than 0.1 to make it accepted. CFI was 0.959, which above 0.9 to make it accepted and Chi-sq was 2.989, which less than 5.0 to make it accepted. Meanwhile, the result of organizational factors shown the CR value of 0.961 and 0.861 for AVE. The fitness indexed for this construct was achieved with the value of RMSEA was 0.067, which less than 0.1 to make it accepted. CFI was 0.945, which above 0.9 to make it accepted and Chi-sq was 2.988, which less than 5.0 to make it accepted.

On the other hand, the environmental factor has a CR value of 0.933; and 0.874 for AVE. The fitness indexed for this construct was achieved with the value of RMSEA was 0.066, which less than 0.1 to make it accepted. CFI was 0.990, which above 0.9 to make it accepted and Chi-sq was 2.925, which less than 5.0 to make it accepted. In the meantime, the result of the individual factor shown the CR value of 0.868; and 0.696 for AVE. The fitness indexed for this construct was achieved with the value of RMSEA was 0.047, which less than 0.1 to make it accepted. CFI was 0.984, which above 0.9 to make it accepted and Chi-sq was 1.955, which less than 5.0 to make it accepted.

Besides that, the ODL construct has a CR value of 0.898 and 0.639 for AVE. The fitness indexed for this construct was achieved with the value of RMSEA was 0.053, which less than 0.1 to make it accepted. CFI was 0.995, which above 0.9 to make it accepted and Chi-sq was 2.257, which less than 5.0 to make it accepted.

Furthermore, the result of the MTUN_OD construct shown the CR value of 0.909 and 0.528 for AVE. The fitness indexed for this construct was achieved with the value of RMSEA was 0.067, which less than 0.1 to make it accepted. CFI was 0. 952, which above 0.9 to make it accepted and Chi-sq was 4.526, which less than 5.0 to make it accepted.

An overall, the technological construct, organizational construct, environmental construct, individual construct, ODL construct, and MTUN_OD construct for CFA distinctively has met the fitness indexes and passed the measurement of AVE and CR.

To ensure the overall fitness indexes of the model, this study employed Pooled-CFA. From the result of pooled CFA, the fitness indexed for this overall model was achieved with the value of RMSEA was 0.050, which less than 0.1 to make it accepted. CFI was 0.951, which above 0.9 to make it accepted and Chi-sq was 1.914, which less than 5.0 to make it accepted.

Based on Table XI, it can be shown that the model achieved discriminant validity when the diagonal values (in bold) are higher than any other values in its row and column. As stated, the data sharing value for discriminant validity was 0.782, ODL was 0.740, and MTUN-OD was 0.755. It indicates that each of the constructs was measure distinctively and not related to each other.

Finally, for the normality test, the distribution does not depart from normality and there were no outliers' data. Thus, the data distribution meets the normality distribution requirement for employing parametric statistical analysis in SEM which will be discussed in the next paper.

In a conclusion, the step-by-step to do pooled-CFA must start with performing an individual CFA for every constructs to make it simpler and easy to understand for the complicated model. All the results must follow the table of indexes (Table II and Table III) to indicate that the results are reliable and validated.

These results of CFA will be used to be modeled in SEM. However, for future work, it is advisable to add 1 more component to be measured; data quality which should be determined under technological construct.

Acknowledgment

I want to express my appreciation to the Research Group of Information Security Forensics and Computer Networking (INSFORNET) and Human Computing Centre (HCC-ISL), Centre of Advanced Computing and Technology (C-ACT) Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka for allowing me to conduct this research. I am incredibly thankful to the Public Service Department (JPA) Scholarship and the Ministry of Education (MOE) for funded and support.

References

[1] S. N. Ismael, O. Mohd, and Y. A. Rahim, “Implementation of open data in higher education: A review,” J. Eng. Sci. Technol., vol. 13, no. 11, pp. 3489–3499, 2018.

[2] M. Krumova, “Higher Education 2 . 0 and Open Data : a Framework for University Openness and Co-creation Performance,” pp. 562–563, 2017.

[3] A. S. Figueiredo, “Data Sharing: Convert Challenges into Opportunities,” Front. Public Heal., vol. 5, no. December, pp. 1–6, 2017.

[4] K. Shamash, J. P. Alperin, and A. Bordini, “Teaching Data Analysis in the Social Sciences: A case study with article level metrics,” Open Data as Open Educ. Resour. Case Stud. Emerg. Pract., pp. 50–56, 2015.

[5] Reale, Giuseppe, “Opportunities and Differences of Open Government Data Policies in Europe,” Athens J. Soc. Sci., vol. Volume 1, no. Number 3, pp. 195–206, 2014.

[6] E. Tran and G. Scholtes, “Open Data Literature Review,” 19th Annu. BCLT/BTLJ Symp. Open Data Addressing Privacy, Secur. Civ. Rights Challenges 2, pp. 1–33, 2015.

[7] J. Manyika, M. Chui, P. Groves, D. Farrell, S. Van Kuiken, and E. A. Doshi, “Open Data: Unlocking Innovation and Performance with Liquid Information,” McKinsey, no. October, p. 24, 2013.

[8] M. W. AlRushaid and A. K. J. Saudagar, “Measuring the Data Openness for the Open Data in Saudi Arabia e-Government - A Case Study,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 12, pp. 113–122, 2016.

[9] Mi. of E. Malaysia, Malaysia Education Blueprint 2015-2025 (Higher Education). 2015.

[10] StudyMalaysia.com, “TVET in Malaysia.” [Online]. Available: Technical and vocational education and training (TVET) in Malaysia - StudyMalaysia.com.

[11] X. Shacklock, From Bricks to clicks. The Potential of Data and Analytics in Higher Education. 2016.

[12] C.-Y. Chiu, S. Chen, and C.-L. Chen, “An integrated perspective of TOE framework and innovation diffusion in broadband mobile applications adoption by enterprises,” Int. J. Manag. Econ. Soc. Sci., vol. 6, no. 1, pp. 14–39, 2017.

[13] L. Chen and X. Yang, “Using EPPM to Evaluate the Effectiveness of Fear Appeal Messages Across Different Media Outlets to Increase the Intention of Breast Self-Examination Among Chinese Women,” Health Commun., vol. 34, no. 11, pp. 1369–1376, 2019.

[14] I. Ajzen, “The Theory of Planned Behavior,” pp. 179–211, 1991.

[15] S. Bauldry, “Structural Equation Modeling,” Int. Encycl. Soc. Behav. Sci. Second Ed., pp. 615–620, 2015.

[16] S. Abrahim, B. A. Mir, H. Suhara, F. A. Mohamed, and M. Sato, “Structural equation modeling and confirmatory factor analysis of social media use and education,” Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019.

[17] Z. Awang, U. Sultan, and Z. Abidin, “Modeling and Analyzing Second Order Model in Structural Equation Modeling The Second Order Confirmatory Factor Analysis (CFA),” no. August 2015, 2016.

[18] Z. Awang, A. Afthanorhan, M. Mamat, U. Sultan, and Z. Abidin, “The Likert scale analysis using parametric based Structural Equation Modeling (SEM),” Comput. Methods Soc. Sci., vol. 4, no. 1, pp. 13–21, 2016.

[19] Jawaria Nasir, Rashidah Mohamad Ibrahim, M. A. Sarwar, and Raja Irfan Sabir, “Impact Of High Involvement Work Practices On Employee

Performances In Health Sector, Pakistan," J. Manag. Theory Pract., vol. 1, no. 2, pp. 22–32, 2020.

[20] I. Gede Mahatma Yuda Bakti and S. Sumaedi, "An analysis of library customer loyalty:The role of service quality and customer satisfaction, a case study in Indonesia," Libr. Manag., vol. 34, no. 6–7, pp. 397–414, 2013.

[21] MOE, "Statistik Pendidikan Tinggi 2018 : Public Universities," pp. 8–45, 2018.

[22] R. Walker, "Another round of globalization in San Francisco," Urban Geogr., vol. 17, no. 1, pp. 60–94, 1996.

[23] Z. Awang, A. Afthanorhan, M. Mohamad, and M. A. M. Asri, "An evaluation of measurement model for medical tourism research: The confirmatory factor analysis approach," Int. J. Tour. Policy, vol. 6, no. 1, pp. 29–45, 2015.

[24] Awang, "Overview of Structural Equation Modeling (SEM)," A Handb. SEM, pp. 1–17, 2012.

[25] N. Ahmad and A. Sabri, "Assessing the unidimensionality, reliability, validity and fitness of influential factors of 8th grades student's mathematics achievement in Malaysia," Int. J. Adv. Res., vol. 1, no. 2, pp. 1–7, 2013.

[26] W. Mohamad, A. Bin, and W. Afthanorhan, "International Journal of Asian Social Science Pooled Confirmatory Factor Using Structural Equation Modeling On Volunteerism Program : A Step By Step Approach Sabri Ahmad Ibrahim Mamat Contribution / Originality," vol. 4, no. 5, pp. 642–653, 2014.

Towards the Development of a Brain Semi-controlled Wheelchair for Navigation in Indoor Environments

Hailah AlMazrua[1], Abir Benabid Najjar[2]
College of Computer and Information Science
King Saud University, Riyadh
Saudi Arabia

***Abstract*—Several technological advancements emerged providing the technical assistance supporting people with special needs in tackling their everyday tasks. Particularly, with the advancements in cost-effective Brain-Computer Interfaces (BCI), they can be very useful for people with disabilities to improve their quality of life. This paper investigates the usability of low-cost BCI for navigation in an indoor environment, which is considered one of the daily challenges facing individuals with mobility impairment. A software framework is proposed to control a wheelchair using three modes of operations: brain-controlled, autonomous and semi-autonomous, taking into consideration the usability and safety requirements. A prototype system based on the proposed framework was developed. The system can detect an obstacle in the front, right and left sides of the wheelchair and can stop the movement automatically to avoid collation. The usability evaluation of the proposed system, in terms of effectiveness, efficiency and satisfaction, shows that it can be very helpful in the daily life of the mobility impaired people. An experiment was conducted to assess the usability of the proposed framework using the prototype system. Subjects steered the wheelchair using the three different operation modes effectively by controlling the direction of motion.**

Keywords—Usability; wheelchair navigation; indoor navigation; mobility impairment; obstacle avoidance; obstacle detection; path planning; BCI; brain-computer interaction

I. Introduction

Independent mobility is an important aspect in the quality of life for individuals with mobility impairments. Though the needs of many individuals with mobility impairments can be satisfied with traditional manual or powered wheelchairs, a part of the impaired community finds it difficult and sometimes impossible to use the wheelchairs independently. This part comprises of individuals with low vision, visual field reduction, spasticity, tremors, or cognitive deficits [1]. These individuals have to depend on another person to push them while they are on the wheelchair, as they often lack the independent mobility to control a powered wheelchair due to the nature of their disability. In the U.S, almost 10% of the legally blind individuals also have a mobility impairment [3], which makes many of them hesitate to visit unfamiliar places since they have no information about the new environment and its accessibility conditions.

Navigation tasks are regarded as one of the critical challenges facing individuals with mobility impairments. According to the Environmental Protection Agency (EPA), the average American spends 93% of their life indoors [2]. The need to consider indoor navigation is even higher in the Saudi local context, mainly for individuals with some cognitive, visual, or physical impairments since 3.73% of the population have some form of functional disability [31].

Moreover, the safety, easiness and usability features in the assistive systems are considered a crucial requirement considering the special situation of mobility-impaired individuals. However, the currently available navigational assistive systems lack these features [32]. Highly impaired individuals, using the powered wheelchair, require an autonomous wheelchair for navigation [4]. Therefore, developing a smart wheelchair that would transport the mobility-impaired individuals to their desired destination without their direct control would significantly improve their quality of life, taking into account that such wheelchair has to plan a quick and a safe path even when faced with an obstacle. Thus, providing the mobility-impaired individuals with some level of independence, by not relying on another person for assistance, as the wheelchair would maneuver through the obstacles by itself. Also, providing the wheelchair's rider the ability to control the wheelchair according to his/her preference is a must. Some riders would like to have a full control on how the wheelchair moves, while others prefer to just sit back and make the electric wheelchair moving autonomously.

Furthermore, the obstacle detection is a key component for any autonomous system, in order to ensure the safety of the individual driving it. Therefore, many research studies, according to [36], have considered integrating different types of sensors in their autonomous systems, in order to detect and avoid the faced obstacles. These sensors can either be ultrasonic, infrared, computer vision type sensors, laser sensors, or a combination of different types.

This paper focuses on the usability engineering of an obstacle avoidance system by combining the new technologies with path optimization techniques taking into consideration the special needs of the target users throughout the development cycle.

II. Background

Assistive technologies are developed according to two approaches: (1) developing special hardware devices designed for disabled people or (2) using existing hardware devices and integrating specific applications to improve some aspects of the individuals’’ daily life. The latter approach is what will be used in this paper.

Smart wheelchairs can either be autonomous, non-autonomous or a mix of both types (semi-autonomous). In autonomous wheelchairs, the TetraNauta [5], and Kanazawa University [6] wheelchairs include a computer for processing along with different types of sensors. They tackled specific problems such as: obstacle avoidance, local environment mapping, and path navigation. With the autonomous control feature, the system analyses the environment, plans a navigation path, detects an obstacle, makes decisions, and controls the wheelchair’s movements [7]. Moreover, several prototypes of smart wheelchairs have been developed and many research papers have been published in this area like in [8] and [9]. However, most of them have hardware and software architectures developed specifically for a particular wheelchair model and usually they tend to be difficult to configure in order to be used by the physically impaired individuals [1][8].

The capability and nature of the user’s disability might become a restriction when operating a wheelchair. Therefore, the spectrum of automation in the brain- controlled wheelchairs (BCW) relate to the level of dependence on the human rider to guide the wheelchair, as compared to the wheelchair guiding itself. Many research studies [35] have investigated the inclusion of different operation modes to the brain-controlled wheelchair (BCW). These modes range from low-level (manual), High-level (autonomous), and shared-control (semi-autonomous). Below is a detailed explanation for each mode:

- Low-level navigation: where the user is in a complete control of the wheelchair as it directly obeys the user’s commands and does not move by itself. The wheelchair is controlled through simple navigation commands, such as “move forward” or “turn left” Also it incorporates basic collision avoidance supports as stopping the wheelchair when obstacles are encountered. Using this navigation mode, users can navigate and perform any path they want, along with having control of the specific movements. In this mode, the system does not assist in the execution of the selected command.
- High-level navigation: where the users have a rough control of the BCW by selecting high-level commands such as “take me to the living room” or “leave this room.” The BCW must have some kind of intelligence so that the specific path to the selected destination is transparent to users, i.e., the user does not select specific low- level commands. Also, it can incorporate basic collision avoidance (stopping the wheelchair when obstacles are encountered) and obstacle avoidance (planning a new path to avoid the obstacle) supports.
- Shared-control navigation: where both the user and the system share the control of the BCW. This can be done in two ways: i) users generate low-level commands, while the system assists the navigation with features such as obstacle avoidance, or maximum likelihood command execution; and ii) users can switch between a low- and a high-level navigation mode. The user in this mode oversees the navigation and issues high-level commands while the wheelchair executes the motions. This enables the user to still remain in charge of the decision-making process, but with less involvement in the execution.

III. Related Work

Recently, considerable amount of research that tackle the operation modes to the BCW have been reported in literature. In the following we report the work that has been done categorized based on the different modes of navigation:

A. *Low-Level Control (Manual)*

The authors in [10] designed a manual BCI-based wheelchair that can be steered by only the users’ brain signals. The designed system utilizes three mental tasks that are turning left, right, and go forward. These commands were achieved through the left and right motor imageries to turn left and turn right, respectively, and feet motor imagery to go forward. Similarly, the authors in [11] developed a manually BCI controlled wheelchair using an onboard computer that is responsible for processing and classifying the captured EEG signals to generate wheelchair steering commands. The onboard computer sends through a serial port the generated commands to the wheelchair's motor drivers. The wheelchair can either move forward, turns left or right, or stops.

On the other hand, Li et al. in [12] presented a hybrid brain/muscle interface to manually control a wheelchair. The authors argue that BCI based on P300 or SSVEP can cause the fatigue and dryness of the eye, and then lead to user's inattention. The developed system distinguishes four user's commands, go forward, turning right, turning left, and stop. These commands are captured from the: (1) two mental states corresponding to the wheelchair’s motion of turning left and turning right, (2) the EMG signal captured from the user’s gritting his/her left teeth and right teeth that correspond to the motion go forward and stop respectfully. The developed wheelchair is equipped with various types of sensors to perceive the environment's context. These sensors are vision camera, and a laser rangefinder (sonar).

The authors in [13], also, developed a hybrid BCI system that combines two types of BCI paradigms (P300 potential and SSVEP) to improve the performance of asynchronous control to instantly and accurately distinguish the control and idle states needed to steer the wheelchair. And considering that the system controls the wheelchair manually, the authors designed the graphical user interface (GUI) to display four groups of buttons, and each group has one large button in the center and eight small buttons surrounding it. When the user concentrates on one group of buttons, both P300 potential and SSVEP can be evoked al the same time. The flickering buttons in each group invoke SSVEP, and the flashing of the four large buttons evoke P300 potential. In order for the authors to produce a go or stop commands in wheelchair control this method was used.

Moreover, in [14] Diez et al. developed a wheelchair that can be controlled through a BCI based SSVEP signals. The designed system can discriminate five classes: top, bottom, left, right and undefined. The detected stimuli can be translated to the first four classes, while the undefined class is chosen when no stimuli are detected. The wheelchair moves manually based on the identified class. For example, it moves forward if the BCI detected the top stimulus, turns left if the left stimulus was detected, turns right if the right stimulus was detected, and finally stops if the bottom stimulus was detected. Moreover, when no stimulus was detected (undefined class) the wheelchair, for safety reasons, stops as well. The system provides the user with online feedback indicating the detected stimulus by translating it to its proper steering direction using a blue arrow on the screen. And in the case of an undefined class, a red circle is shown in the center of the screen.

The authors in [15] developed a BCI based electric wheelchair control system. The proposed wheelchair enabled users to steer the wheelchair in four directions forward, backward, turn left or right by utilizing the eyes closing signal for more than one second and without any pre-training. Users wore an EEG acquisition cap that has four lights corresponding to the four directions. These light flashes in a clockwise loop and lasts for a fixed period of time. Moreover, when the user wants to select a specific direction, he/she closes their eyes as soon as the desired direction flashes.

Moreover, the authors in [16] built upon their previous work of [17] where they combined two types of BCI paradigms (motor imagery and P300 potentials) and controlled the wheelchairs direction and speed. The direction control was achieved through two commands (turning left and right), and the speed control was done by controlling the acceleration and deceleration. The authors' new extension combined motor imagery, P300 potentials, and eye blinking to achieve forward, backward, and stop control of a wheelchair respectively. Their work combined with their previous resulted in having the users choose and navigate from seven steering commands.

Cao et al. in [18] developed a hybrid BCI system that combines two BCI paradigms (motor imagery and SSVEP) to concurrently control the speed and direction of a wheelchair. The proposed system manually steers the wheelchair by providing eight commands for the users choose and navigate from. These commands are turn left, turn right, drive forward, accelerate, decelerate, drive at the uniform velocity, and turn on and off the switch. Similarly, the authors in [19] argue that their proposed system allows the user to implement different types of commands in parallel. The proposed combine SSVEP and MI tasks to develop a new hybrid BCI method. It utilizes two-class MI and four-class SSVEP tasks, in which the user imagines moving his/her left or right-hand and focuses on one of four oscillating visual stimuli simultaneously.

In [20], the authors developed a manual control wheelchair navigation system based on a hidden Markov model (HMM). The developed system steers the wheelchair by capturing the electrooculography (EOG) signal originating from the user's eyeball and eyelid movements. A feature extraction was used to determine whether the eyes are open or closed and whether the eyes are gazing to the left, right, or center. These features are used as inputs to the HMM which generates commands for navigating the wheelchair accordingly. The wheelchair is equipped with a proximity sensor to avoid obstacles and it can move forward and backward in three directions.

Varona-Moya et al. in [21] enhanced an electric wheelchair, by incorporating multiple sensors and emulating its analog 2-axis joystick with a custom-built control board. The enhanced system receives the BCI navigation commands through a TCP connection and then transforms them into low-level movement commands that are fed to the wheelchair. Moreover, a real-time map of the area surrounding the wheelchair was created using the incorporated eleven ultrasonic rangefinders (sonar). Updating the grid-map in real time was achieved using a sonar model, which upon the detection of an obstacle at a given distance; it updates all the grid cells within the obstacle detected vertex.

In [22] Đumić and Kevrić developed a manually controlled wheelchair. The user controls the direction of the wheelchair through a BCI headset that detect an eye blinking action. There are four directions the user can choose from in order to steer the wheelchair, which are: left, right, forward, and backwards. The actual wheelchair was enhanced by using a microcontroller to control the joystick by servo motors.

B. *High-Level Control (Autonomous)*

Define The authors in [23] designed a brain-controlled wheelchair, which interacts with the user using a simple interface. The navigation system proposes a semantic map that integrates the navigation points, semantic targets, and a local 3D map. The semantic targets provide the recognized objects' type, outline and its functionality information, for example, if the object was identified to be a table, then its associated information is that it can be docked. The local 3D map provides a traversable navigation point. The user chooses one of the navigation points from the semantic map as the destination goal using a brain-computer interface (BCI).

In [24], Ng et al. developed a BCI controlled wheelchair based on the steady-state visual evoked potential (SSVEP) paradigm. The proposed system takes the desired destination from the user through a BCI signal and communicates it to the wheelchair navigation system to plan a path autonomously while avoiding obstacles on the way to the destination. The responsibility of controlling the wheelchair is switched from the user to the navigation software, which reduces the number of BCI commands needed to steer the wheelchair to the desired destination.

Zhang et al. in [25] developed an autonomous wheelchair where the user selects a destination from the map using one of the BCI paradigms (motor imagery or P300). Based on the selected destination, the navigation component plans a short path and steers the wheelchair to the desired goal. Furthermore, the user can choose to stop the wheelchair by issuing a stop command using the BCI whenever he/she wants. The authors claim that their system reduces the user's mental burden substantially.

C. Shared Control (Semi-autonomous)

The authors in [26] introduced a shared control architecture that combines the user's intention along with the precision of a powered wheelchair. Their system combines BCI with a shared control architecture that permits users to produce dynamic and simple navigation directions, as opposed to users being seated and relying on a predefined path for most of the navigation time. The shared controller decides what actions ought to be taken, based on the user's input (turning left or right) while taking into account the context of the environment, which was perceived using ten sonar sensors and two webcams. The proposed system was evaluated against four healthy experienced BCI users.

Lopes et al. in [27] propose a robotic assistive navigation wheelchair that integrates the Brain-Computer Interface (BCI) technology, as the Human– Machine Interface (HMI). In their paper, the authors proposed a two-layer collaborative control approach that takes into account both the user and the machine commands to guide and maneuver the wheelchair. The first layer in their proposed approach is a virtual-constraint layer. It is responsible for enabling/disabling the user commands, based on the environment's context. For example, user commands are enabled in the situation of multiple directions that was caused by newfound obstacles in the environment. The second layer is responsible for matching the user's intended BCI commands to a suitable steering command, considering the client capability to control the wheelchair and situation awareness of potential directions at a given location.

Moreover, in [28] the authors proposed a new shared-control approach for a brain- controlled wheelchair. The shared controller switches between two controlling agents, the BCI control agent, and an autonomous control agent taking into consideration both the context of the environment and the intention of the user. The architecture of the proposed approach consists of four layers: the human- machine interface, the global motion planning, the local motion planning, and the motion control. The global motion planning layer comprises of the knowledge database that was built by the SLAM method to store information about the location of obstacles as well as the kinematics of the wheelchair. This layer receives an output from the BCI and determines the path to the goal by using the NRPF algorithm. It also calculates the distances between the wheelchair and the goal as well as the distances between the wheelchair and the obstacle. Moreover, the distance to the obstacle decides the mode of control. For example, the autonomous control agent is activated once an obstacle is detected using a laser scanner, which results in enabling the local path-planning module to return a collision-free path by employing an improved potential field method. Finally, the motion control layer receives direction commands from the local motion-planning layer. The received commands determine the movement of the wheelchair.

Chen et al. in [29] proposed a hybrid BCI scheme based on brain electrophysiology (EEG) signals, a shared control system of a bionic manipulator is designed and a motion planning of the wheelchair. The wheelchair is equipped with an obstacle avoidance system that is composed of 8 photoelectric sensors distributed across the wheelchair. The proposed system is comprised of the human-robot interface and EEG signals that are used to identify the steering command expressed by the user's motion imagery, which is achieved by imaging motions of the left and right legs, left and right hands and so on. The user's expressed commands are translated into the corresponding control actions, which are sent to the actuator of the arm joint motor to realize the motion control of the bionic manipulator.

Moreover, researchers in the field have used a number of different metrics to evaluate their proposed the BCW. However, there are a limited number of research that employ the usability evaluation metrics. For example, from the reviewed literature only [20] and [14] evaluated their BCW using usability, learnability, and measured the user’s experience. The rest of the reviewed papers evaluated their BCW in terms of how accurate their system completed the task and the time it took among other factors.

Furthermore, prior to the test session, most studies performed a training phase to reduce the effect of unfamiliarity to the technology used. The orientation session aims to introduce the participants to the subject in general and to the system in particular. Some studies considered the users’ familiarity with the test BCI technology. Besides, the profile and number of test participants varied among the studies: Lopes et al. [27] recruited 11 participants one of which has a cerebral palsy and motor impairment, Diez et al. [14] recruited 13 participants, with one being a paraplegic participant. While another study in [24] evaluated the system against 37 healthy participant. Other studies in [10]–[13], [15], [16], [18], [21]–[23], [25], [26], [28], [29] recruited lesser number of test participants ranging between one to eight participants.

In summary, the problem of choosing one signal or another depends on a number of factors like presence or absence of a graphical user interface (GUI), the number of commands, and the need to implement a continuous control of the wheelchair. The P300 requires a GUI to be present and has a discrete control mechanism as opposed to the ERD/ERS. Furthermore, we noticed that a few of the reviewed research handled the semi- autonomous (shared level) navigation and combining it with obstacle detection and employing a path planning algorithm to find the shortest path to the goal. Remarkably, most the research (63.16% of the papers) uses a manual (low level) navigation system without adding an obstacle detection and avoidance mechanism. On the other hand, the P300 used mostly high-level navigation commands.

Moreover, some of the reviewed high-level BCWs focus solely on the BCI interface without mentioning the path planning technique used. Similarly, some works like [23], [24], and [25] did not consider applying an obstacle detection and avoidance technique, this might be attributed to the fact that they are building their system to navigate in a static environment without taking into consideration the changes that might suddenly occur in the environment.

It should be noted that the evaluation criteria used in the different studies were heterogeneous, where no standard evaluation metrics were followed. When taking the type of signals into account the BCW that were based on the ERD/ERS or SSVEP paradigms all gave a great importance to the evaluation of accuracy in selection the commands, along with the time it was required to complete the path. On the other hand, half of the BCW adapting the P300 paradigm evaluated the selection time, which is defined as the time it takes the user to select the desired command. However, the selection time cannot be used when evaluating a BCW that is based on ERD/ERS signals, due to the difficulty of exactly knowing when the user starts the selection process.

This paper extends this line of research and proposes a software framework integrating the three different operating modes. A prototype system based on this framework was implemented including three components: the navigation and path planning, the obstacle detection, and the user interface. Finally, the usability of the proposed system was evaluated against standardized evaluation metrics.

IV. Proposed Framework Design and Architecture

This paper proposes to design and develop a software framework for a BCI controlled wheelchair that gives the user the ability to choose one of the three navigation methods: manual/direct control (low level control), semi- autonomous (shared control), and autonomous (high level control).

In the manual/direct (low level control), the user controls the wheelchair by only his thoughts. In the semi-autonomous (shared control), the control takes the user's intention into account while doing the planning and the wheelchair follows the orders feed to it from a planner. The user, however, might express his/her intention to control the wheelchair and steer it by merely using the BCI. When the user's steering intention conflicts with the planner's instructions, the control system will replan the task. Finally, in the autonomous (high level control) the user will only have to select the destination and let the system plan a safe path to the destination while avoiding any obstacles on the way. However, in all the three navigation methods, the wheelchair will be equipped with obstacle detection sensors in order to detect and avoid the obstacle.

A. High-level Framework Architecture

As a typical navigation system, the framework consists of four main components: the indoor positioner component, the navigator component, the obstacle detection component, and the user interface component. The architecture of the whole system is represented in Figure 1.

1) Positioning component: The indoor positioner component gathers relevant information concerning the navigated environment and the user's current position and makes this information available for the other component to use. These components include the navigation component that generates the navigational directions based on the positioner information and the interface component that outputs related contextual information based on the positioner information.

2) Navigation component: Navigation including path planning is a key component to any navigational system, since it is responsible for computing the optimal path to reach the destination selected by the user, taking into account the user preferences including the route length, number of turn and safety priorities. The computed path is then translated to commands fed into the wheelchair to guide the user through the environment. Moreover, Path planning environments can either be static or dynamic, and according to them, two different path-planning approaches exist: local and global path planning. Global planning assumes a prior knowledge of the environment, a predefined map and fixed obstacle and then computes the overall path to destination. This approach assumes a static environment and is usually used for indoor navigation [30]. On the other hand, local planning aims to locally plan an obstacles-free path and determine how to navigate around obstacles. The local planning is usually used in robotics since the construction and maintenance of a global map may become computationally complex for a robot. This component also includes the shared controller which receives commands from the user and the machine, evaluates the situation and the current environment and then makes the appropriate decision. This decision is then translated into commands via the Path Translator to control the wheelchair.

3) Obstacle detection and avoidance component: The obstacle detection component, detects the existence of an obstacle, notifies the user and takes the appropriate action, which is identifying the obstacle on the map and then communicats with the navigator component to plan an alternative path to avoid the obstacle in order to reach to the goal destination. It should be noted that the control is passed when an obstacle is detected to so the wheelchair can manuver around the obstacle autonomously.

4) User interface component: Specifying a component for handling the interaction with the user and ensuring an instant response to user's requests is essential for any system. This component gets the preference of the user on how to plan a path to navigate the environment to reach the desired destination and then communicates with the navigation component so that the user's preference and destination point are fed into the path-planning algorithm. Moreover, the user interface component notifies the user when an obstacle is detected.

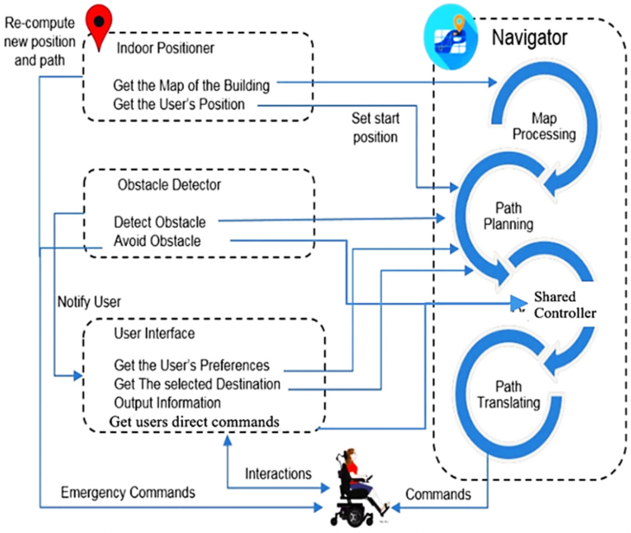


Fig. 1. Wheelchair High-level Framework Architecture.

B. Shared Control System Architecture

As the main focus is to study and assess the usability of the shared control navigation, 3 components from the aforementioned components were considered: namely Navigation, Obstacle Detection and User Interface. The architecture of the proposed shared control system was designed in layers to interact synchronously so that each component can communicate independently to send or receive information. The architecture presented in FIGURE 2 is structured in four layers: Human Machine Interface (HMI), Global Motion Planning, Local Motion Planning, and Motion Control.

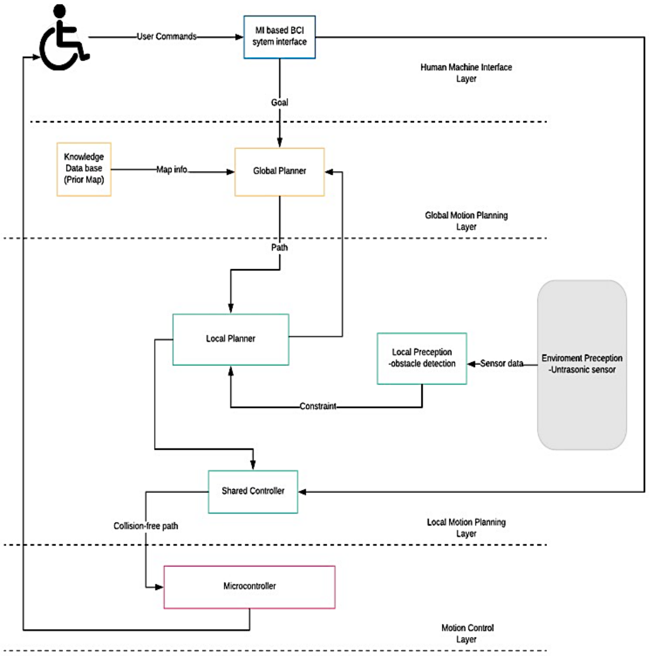

User Commands
MI based BCI sytem interface
Human Machine Interface Layer
Goal
Knowledge Data base (Prior Map)
Map info
Global Planner
Global Motion Planning Layer
Path
Local Planner
Local Preception -obstacle detection
Sensor data
Enviroment Preception -Untrasonic sensor
Constraint
Shared Controller
Collision-free path
Local Motion Planning Layer
Microcontroller
Motion Control Layer

Fig. 2. The Proposed Shared Control System Architecture.

The first layer implements the input interface of the User Interface component. In this layer, an MI-based BCI is used to provide the user intent, which are the issued steering commands. The Navigator component was implemented across three layers. The global planner determines the path to a predefined goal, based on a priori grid map information. The local planner plans a new path to avoid new detected obstacles in the environment. The shared-controller determines the set of appropriate moves to reach a predefined goal combining both the user and the machine commands. The obstacle detection is implemented in the third layer, along with the local planner, and is performed out based on ultrasonic sensor information.

C. Implemented System Components

The implemented components include the BCI interface, navigation component and obstacle detection.

- BCI Interface Component:

For BCI, hybrid BCI that uses both motor imagery and P300 potential approach was adapted in this system. The idea of hybrid BCI is to activate and control the system by imagining performing specified tasks. First, the device has to be trained in order to recognize the brain patterns of a user (motor imagery). Users were trained in two mental commands: neutral, and push. The choice of having only two mental commands rather than associating each task with a command for itself; is due to the user's difficulty of distinguishing different commands and thus increasing his/her burden.

The neutral command is trained by asking the user to stay idle and think of nothing for two seconds, in which the user does not perform any command. And the push command is trained by asking the user to imagine pushing the cube for four seconds. These commands, however, have to be trained multiple times in order for the commands to have a high confidence.

Furthermore, the P300 was used to control which action of direction the wheelchair has to take (right, left, forward, backward). For example, to turn right, the user imagines push movement while focusing on the right arrow flashing button on the graphical user interface (GUI).

- Obstacle Detection Component

The wheelchair will be equipped with Ultrasonic sensors mounted on its front part in order provide a safety mechanism against collision. If the measured distance to the obstacle is smaller than 50cm, the map is updated with the new obstacle and the local planner will plan a path to avoid the obstacle.

- Navigation Component

o ***Vector Map***

A two- dimensional vector map of the environment will be manually constructed and pre- processed to define the obstacles' locations so that the generated path can avoid intersecting with them later. The map will be segmented into a grid of equal sized cells (1 meter of each cell). It is assumed that all doors and windows on the wall are to be closed and are considered as normal walls. This assumption implies that the wheelchair cannot travel through them.

o ***Path Planning***

The proposed system includes two path-planning algorithms: (1) the global path planner finds the shortest path from initial location to the goal, (2) the local path planner to plan a path around the obstacle to avoid it. The goal (endpoint) is assumed to be selected by the user.

The D*Lite, global path planning algorithm is used to move a robot equipped with an ultrasonic sensor for detecting obstacles in order to move from the start point to a goal point assuming that the environment and all obstacles are known, and the size of each cell equals the size of the robot. For the local planning part, a local repair strategy called path splicing was used. The path splicing strategy finds a path nearby and assumes the path farther away need not be recomputed until we get closer to it. So instead of recalculating the entire path, the first M steps of the path are calculated.

o ***Shared Controller***

The shared controller receives commands from two agents, the user and the machine. The user issues BCI commands using the BCI headset. While the machine commands are issued from the application. However, instead of directly executing the user's commands, the shared control component evaluates the situation first. The current environment, perceived through ultrasonic sensors, is taken into considerations.

The shared controller has two levels of support that are only initiated when the situation calls for them. The two levels of support are collision avoidance and obstacle avoidance that will be activated near obstacles to prevent collisions.

The collision avoidance is thought of as an emergency stop. For example, when the user moves the wheelchair too close to an obstacle, the velocity will be decreased until it comes to a full stop. The ultrasonic sensors, mounted on the wheelchair, are used to determine when to activate this behavior. The activation threshold was set at 0.4m to maintain the safety of the user. If the ultrasonic sensors detect obstacles within this threshold, the collision avoidance behavior will be activated. However, unlike the previous behavior, the obstacle avoidance employed the use of the local planning strategy in order to steer the robot away from the obstacle. This behavior takes both the input of the user and the environment into consideration, to properly assist the local planning strategy.

D. Deployment And Navigation Modes

The proposed system consists of three main nodes. The first node is the Emotiv insight headset to capture EEG signal from the user. The second node hosts the software application that receives the captured EEG signal from the headset and converts it into commands. It also plans an obstacles free path to the destination. The third node is the microcontroller to control the wheelchair movement. However, due to limited funding, the wheelchair could not be acquired and was replaced by a prototyped robot. The robot is a Boe-Bot robot equipped with an Arduino UNO as the microcontroller to operate two motors and several ultrasonic sensors.

When using the Emotiv insight headset, the user was seated and asked to focus his attention to the command he wished to instruct. It is important to note that to know that every person has unique EEG signals. Therefore, every participant has to go through a training session before proceeding with the system. Furthermore, the system's nodes are able to share messages between them via Bluetooth. The software application component can control the robot through sending commands as a string of characters. The commands used in the system are: forward (F), turn 90 degree to the left (L), turn 90 degree to the right (R) , and stop (S).

V. Pilot Usability Evaluation

The (ISO 9241-11: 2018) identifies effectiveness, efficiency and satisfaction as major attributes of the usability. To evaluate the usability of the proposed system, test tasks were designed to assess the metrics for measuring the required criteria. Hence, the effectiveness was evaluated in terms of task completion, the efficiency was measured based on completion time and workload, and the satisfaction was assessed using the System Usability Scale (SUS).

A. Usability Evaluation Materials

A prototype based on a car robot (Boe-Bot) was used to simulate the wheelchair. The robot car is equipped with an ultrasonic sensor to detect obstacles. The assembled robot is shown in Figure 3.

1) Moreover, the Emotiv Insight [33] was used as the BCI headset to read and transmit EEG signals. Additionally, the BCI interface has been designed as a visual oddball paradigm. The paradigm comprises four steering commands, encoded by the following symbols: FORWARD, RIGHT90, LEFT90 and STOP, as shown in Figure 4. These symbols flash randomly. At a given moment, the relevant steering event is the symbol mentally selected by the user, which corresponds to the direction he/she wants to follow, and all other flashing symbols are discarded and considered as a non-relevant-event.

The map for room used for the evaluation is shown in Figure 5. The room is an 8x8 m^2, where the green circle is the starting point and the purple circle is the final destination (goal).

B. Usability Evaluation Method

Each experimental session was designed to last for a duration of about one hour, during which subjects were asked to control the robot using the Emotiv headset. A detailed flow is presented in Figure 6.

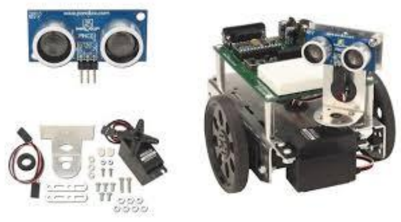


Fig. 3. The Assembled Robot.

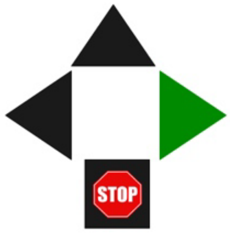


Fig. 4. The P300 GUI.

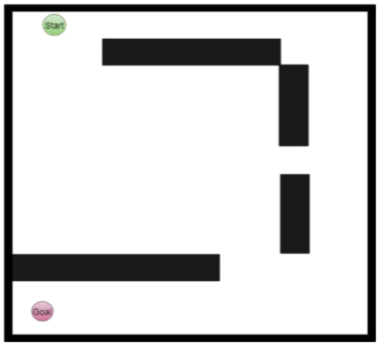


Fig. 5. Experimental Room Mock-up.

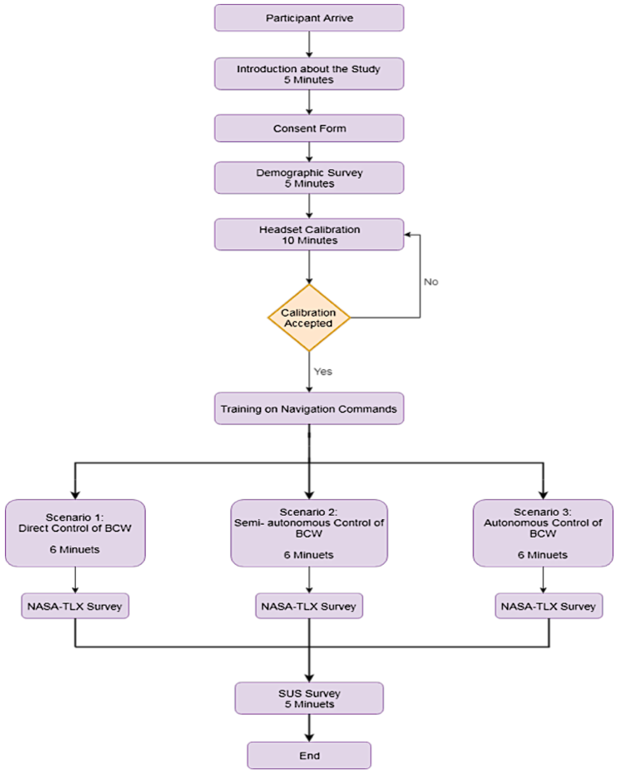


Fig. 6. The Experimental Session Procedure.

Each session began with asking the participant to fill out a demographic questionnaire and sign a consent form. Then, the Emotiv headset was mounted on the participant for calibration. Calibration involves mounting the headset on the participant and ensuring a good signal for each of the five electrodes (green color). In some cases, a fair signal (yellow color) is accepted. Following the calibration and before the online session, each participant was asked to perform an off-line training to control wheelchair, this training was done until the participant's signal recognition accuracy reaches above 85 percent.

After the calibration and off-line training, the participants were asked to perform three real-time navigation scenarios (direct control, semi-autonomous, and autonomous control) for a duration of six minutes each. The navigation scenarios all took place in a structured known environment that included both static mapped obstacles and an obstacle/s new to the environment.

The first navigation scenario, denoted by scenario 1, was to directly control the wheelchair. The participant is instructed to focus on the GUI that has four symbols flash randomly. These symbols correspond to the four steering commands (forward, turn right 90 degree, turn left 90 degree, and stop). When the participant chooses to go forward the robot continues performing this command till the users issues another command or the robot stops due to facing an obstacle. Moreover, in the second navigation scenario, denoted by scenario 2, where the participants share the responsibility of controlling the wheelchair with system. The wheelchair autonomously moves from the starting point to the destination while avoiding obstacles. However, the participants can interfere anytime and change the direction of the wheelchair and steer it as he/she wishes. Once the participant is done with commanding the wheelchair, the system takes back the control and navigates the wheelchair back to the destination. Finally, in the third navigation scenario, denoted by scenario 3. The participant is instructed to sit still while the wheelchair is navigating the environment to reach to the final destination (goal) while avoid the new introduced obstacle. It is also worth noting that the order of the three navigation scenarios was counter-balanced across the participants as shown in TABLE 1

TABLE I. THE SEQUENCE OF THE 3 NAVIGATION SCENARIOS FOR EACH PARTICIPANT

ID	Sequence of Navigation Scenarios
P1	Scenarios 1, 2, 3
P2	Scenarios 1, 3, 2
P3	Scenarios 2, 1, 3
P4	Scenarios 3, 2, 1
P5	Scenarios 2, 3, 1

C. Evaluation Participants

TABLE 2 illustrates the demographics of the participants taking part in the usability evaluation. The participants were mainly recruited through the use of social networks and word of mouth. All the participants took part in an initial calibration task. This initial calibration was required in order to have an acceptable accuracy of the acquired brain signals, which would enable the participants to control the BCW in the navigation task.

TABLE II. THE DEMOGRAPHICS OF THE PARTICIPANTS TAKING PART IN THE USABILITY EVALUATION

ID	Gender	Age	BCI Experience	Degree of Motor Disability
P1	Female	29	No	None
P2	Male	22	No	None
P3	Male	24	No	Fractured leg
P4	Female	66	Yes	None
P5	Female	25	No	None

D. Evaluation Results

The participants were asked to perform the navigation scenario they are presented with. The scenario is considered completed if the participant performs the navigation in less than 6 minutes. Moreover, all the participants have completed the navigation scenarios in a short time period for the direct control scenario (scenario 1) except P1. These results range from 4.9 to 6.2 minutes. It can be seen that P4 completed the scenario in short time which is attributed to the fact that the participant has prior knowledge on how to use the BCI. However, P1 got the highest completion time, which indicates that the user faced some difficulty using the BCI.

Furthermore, it can be seen from FIGURE 7 that the participants' completion time improved and this can be attributed to the learning effect. In addition, it was observed that the sequence of the navigation scenarios that the participants were supposed to follow affected the completion time. P1 and P2 for example took scenario 1 which is the direct control first and that helped them get familiar with how BCW work and got a first-hand experience on how to control the prototype via a sequence of commands. Also, according to the evaluation model this would result in a high efficiency of the proposed framework.

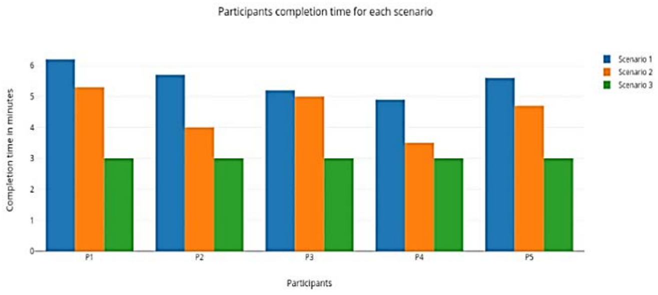


Fig. 7. Navigation Completion Time Periods among Participants for different Three Scenarios.

Moreover, the overall results from the NASA-TLX [34] workload survey showed that the participants' perceived the BCW as having an average physical demand, although, the BCW experiments does not require movement. Therefore, we believe this average score in physical demand was due to fatigue from sitting during the set-up and calibration period, and the increased test length. We believe the increase in mental demand and effort in scenario 1 and 2 was because the BCW experiments required people to focus their attention on making selections, compared to scenario 3 where the participants sat still and let the system do the steering. Nevertheless, it should be noted that in the autonomous navigation, participants reported they felt some frustration, which could be attributed to the fact that the participants might have wanted some navigation control over the prototyped BCW.

The overall analysis reported in FIGURE 8 shows that the workload perceptions increased significantly when the participants were dealing with scenario 1, which is when the workload demand increased.

Moreover, regarding the accuracy in selection the correct command. It was noted that previous experience in using the BCI can affect the accuracy results. P4, which had a previous experience in using a BCI application performed better and had a high accuracy in selecting the right command. However, with more training it is expected that the other participants can perform the same and improve their selection accuracy.

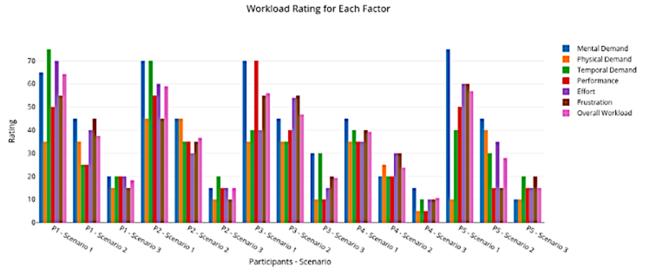


Fig. 8. Workload Rating for each Factor for each Participant/Scenario.

Additionally, according to the SUS questionnaire results, in that were filled by the participants after testing the system to measure the likability of the solution provided. It was found that participants' opinions were diverse in regard to their experience of using the prototyped BCW. While one participant found it difficult to use the system, two found the experience relatively easy and two were neutral in this regard. This applies to their view of the system and whether it can be described as an easy-to-use system or not, despite their own experience with it. Three of the participants believe that they don't require any specialized help in order to be able to use the system. Only two of the participants think that training on the system is required before using it.

Moreover, the majority of the participants agreed the stimuli were easy to distinguish and understand. Moreover, steering the prototyped BCW in an autonomous navigation mod was considered to be easy, similarly, steering the prototyped BCW in a semi- autonomous mode. On the other hand, three of the participants found that steering the prototyped in a direct control mode was difficult. Overall, the participants showed an interest in using the BCW once it becomes available in the market.

Table 3 and Figure 9, that were filled by the participants after testing the system to measure the likability of the solution provided. It was found that participants' opinions were diverse in regard to their experience of using the prototyped BCW. While one participant found it difficult to use the system, two found the experience relatively easy and two were neutral in this regard. This applies to their view of the system and whether it can be described as an easy-to-use system or not, despite their own experience with it. Three of the participants believe that they don't require any specialized help in order to be able to use the system. Only two of the participants think that training on the system is required before using it.

Moreover, the majority of the participants agreed the stimuli were easy to distinguish and understand. Moreover, steering the prototyped BCW in an autonomous navigation mod was considered to be easy, similarly, steering the prototyped BCW in a semi- autonomous mode. On the other hand, three of the participants found that steering the prototyped in a direct control mode was difficult. Overall, the participants showed an interest in using the BCW once it becomes available in the market.

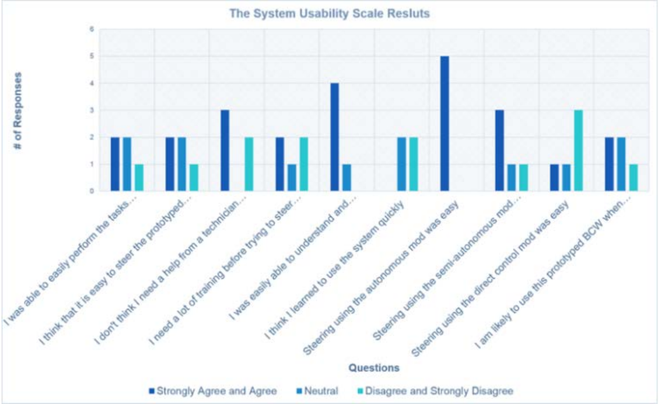


Fig. 9. System Usability Scale (SUS) Results.

TABLE III. SUS RESULTS (1: STRONGLY AGREE, 5: STRONGLY DISAGREE)

Question	**1**	**2**	**3**	**4**	**5**	**Mean Opinion Score**
I was able to easily perform the tasks required using the BCI	0	2	2	1	0	2.2
I think that it is easy to steer the prototyped BCW	1	1	2	1	0	2.6
I think that I need a help from a technician person in order to be able to steer the prototyped BCW	1	1	0	1	2	3.4
I need a lot of training before trying to steer the prototyped BCW	0	2	1	2	0	3
I was easily able to understand and distinguish the application's stimuli	2	2	1	0	0	1.8
I think I learned to use the system quickly	0	0	2	2	1	3.8
Steering using the autonomous mod was easy	3	2	0	0	0	1.4
Steering using the semi-autonomous mod was easy	3	1	1	0	0	1.6
Steering using the direct control mod was easy		1	1	2	1	3.6
I am likely to use this prototyped BCW when it becomes available in the apps market		2	2	1	0	2.8

VI. CONCLUSION

The work of this research was motivated by number of factors in order to improve the quality of life for individuals with mobility impairments since most of the smart Wheelchairs developed have hardware and software architectures that are specific for the wheelchair model developed and are usually very difficult to configure in order for the physically impaired individuals to start using them.

In this work, we proposed a framework by engineering three components (navigation and path planning, obstacle detection, and user interface) taking into consideration the usability and safety requirements to develop a brain- controlled wheelchair for mobility-impaired individuals to help them navigate their way seamlessly in an indoor environment. However, number of issues were encountered resulting in number of limitations. Among these limitations was that the framework was developed using a prototype rather than a real wheelchair, which is attributed to the limited and late funding. On the usability aspect, the small number and lack of variance within the evaluation sample might, though it is considered as an initial evaluation.

In the future we aim to test the proposed framework with a real wheelchair in addition to increasing the number and variance of the evaluation sample.

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by King Abdulaziz City for Science and Technology (KACST) by grant number (1-17-02-001-0031).

REFERENCES

[1] J R. C. Simpson, "Smart wheelchairs: A literature review," J. Rehabil. Res. Dev., vol. 42, no. 4, p. 423, 2005.

[2] N. E. Klepeis et al., "The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants," J. Expo. Anal. Environ. Epidemiol., vol. 11, no. 3, pp. 231-252, Jun. 2001.

[3] R. Simpson et al., "A prototype power assist wheelchair that provides for obstacle detection and avoidance for those with visual impairments," Journal of neuroengineering and rehabilitation, vol. 2, p. 30, Nov. 2005, doi: 10.1186/1743-0003-2-30.

[4] R. A. M. Braga, M. Petry, A. P. Moreira, and L. P. Reis, "Concept and Design of the Intellwheels Platform for Developing Intelligent Wheelchairs," in Informatics in Control, Automation and Robotics: Selcted Papers from the International Conference on Informatics in Control, Automation and Robotics 2008, J. A. Cetto, J.-L. Ferrier, and J. Filipe, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 191-203.

[5] "Impaired Physical Mobility – Nursing Diagnosis & Care Plan - Nurseslabs." [Online]. Available: https://nurseslabs.com/impaired-physical-mobility/. [Accessed: 10-Oct-2017].

[6] "What is assistive technology?" [Online]. Available: https://www.washington.edu/accessit/print.html?ID=1109. [Accessed: 10-Oct-2017].

[7] P. C. Garrido, I. L. Ruiz, and M. Á. Gómez-Nieto, "Support for Visually Impaired through Mobile and NFC Technology," in IT Revolutions, 2011, pp. 116–126.

[8] D. E. Hinton Sr, "Research and technological aids for people who are deaf-blind," Am. Rehabil., vol. 15, no. 2, pp. 7–11, 1989.

[9] W. Hasselbring, "Component-based software engineering," Handb. Softw. Eng. Knowl. Eng., vol. 2, pp. 289–305, 2002.

[10] J. Li, J. Liang, Q. Zhao, J. Li, K. Hong, and L. Zhang, "DESIGN OF ASSISTIVE WHEELCHAIR SYSTEM DIRECTLY STEERED BY HUMAN THOUGHTS," Int. J. Neural Syst., vol. 23, no. 03, p. 1350013, Jun. 2013.

[11] S. M. T. Müller, T. F. Bastos, and M. S. Filho, "Proposal of a SSVEP-BCI to Command a Robotic Wheelchair," J. Control Autom. Electr. Syst., vol. 24, no. 1, pp. 97–105, Apr. 2013.

[12] Z. Li, S. Lei, C. Su, and G. Li, "Hybrid brain/muscle-actuated control of an intelligent wheelchair," in 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2013, pp. 19–25.

[13] Y. Li, J. Pan, F. Wang, and Z. Yu, "A hybrid BCI system combining P300 and SSVEP and its application to wheelchair control," IEEE Trans. Biomed. Eng., vol. 60, no. 11, pp. 3156–3166, 2013.

[14] P. F. Diez et al., "Commanding a robotic wheelchair with a high-frequency steady-state visual evoked potential based brain–computer interface," Med. Eng. Phys., vol. 35, no. 8, pp. 1155–1164, Aug. 2013.

[15] D. Ming et al., "Electric wheelchair control system using brain-computer interface based on alpha-wave blocking," Trans. Tianjin Univ., vol. 20, no. 5, pp. 358–363, Oct. 2014.

[16] H. Wang, Y. Li, J. Long, T. Yu, and Z. Gu, "An asynchronous wheelchair control by hybrid EEG-EOG brain-computer interface," Cogn. Neurodyn., vol. 8, no. 5, pp. 399–409, Oct. 2014.

[17] J. Long, Y. Li, H. Wang, T. Yu, and J. Pan, "Control of a simulated wheelchair based on a hybrid brain computer interface," Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 2012, pp. 6727–6730, 2012.

[18] L. Cao, J. Li, H. Ji, and C. Jiang, "A hybrid brain computer interface system based on the neurophysiological protocol and brain-actuated switch for wheelchair control," J. Neurosci. Methods, vol. 229, pp. 33–43, May 2014.

[19] J. Li et al., "Evaluation and application of a hybrid brain computer interface for real wheelchair parallel control with multi-degree of freedom," Int. J. Neural Syst., vol. 24, no. 4, p. 1450014, Jun. 2014.

[20] F. Aziz, H. Arof, N. Mokhtar, and M. Mubin, "HMM based automated wheelchair navigation using EOG traces in EEG," J. Neural Eng., vol. 11, no. 5, p. 056018, 2014.

[21] S. Varona-Moya, F. Velasco-Álvarez, S. Sancha-Ros, Á. Fernández-Rodríguez, M. J. Blanca, and R. Ron-Angevin, “Wheelchair navigation with an audio-cued, two-class motor imagery-based brain-computer interface system,” in 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), 2015, pp. 174–177.

[22] D. Đumić and J. Kevrić, “BCIs for Electric Wheelchair,” in Advanced Technologies, Systems, and Applications II, 2018, pp. 833–847.

[23] Z. Wei, W. Chen, J. Wang, H. Wang, and K. Li, “Semantic Mapping for Safe and Comfortable Navigation of a Brain-Controlled Wheelchair,” in Intelligent Robotics and Applications, 2013, pp. 307–317.

[24] D. W. Ng, Y. Soh, and S. Goh, “Development of an Autonomous BCI Wheelchair,” in 2014 IEEE Symposium on Computational Intelligence in Brain Computer Interfaces (CIBCI), 2014, pp. 1–4.

[25] R. Zhang et al., “Control of a Wheelchair in an Indoor Environment Based on a Brain–Computer Interface and Automated Navigation,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 24, no. 1, pp. 128–139, Jan. 2016.

[26] T. Carlson and J. del R. Millan, “Brain-Controlled Wheelchairs: A Robotic Architecture,” IEEE Robot. Autom. Mag., vol. 20, no. 1, pp. 65–73, Mar. 2013.

[27] A. C. Lopes, G. Pires, and U. Nunes, “Assisted navigation for a brain-actuated intelligent wheelchair,” Robot. Auton. Syst., vol. 61, no. 3, pp. 245–258, Mar. 2013.

[28] J. Duan, Z. Li, C. Yang, and P. Xu, “Shared control of a brain-actuated intelligent wheelchair,” in Intelligent Control and Automation (WCICA), 2014 11th World Congress on, 2014, pp. 341–346.

[29] N. Chen, X. Wang, X. Men, X. Han, J. Sun, and C. Guo, “Hybrid BCI based control strategy of the intelligent wheelchair manipulator system,” in 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018, pp. 824–828.

[30] M. Samadi and M. F. Othman, “Global Path Planning for Autonomous Mobile Robot Using Genetic Algorithm,” in 2013 International Conference on Signal-Image Technology Internet-Based Systems, 2013, pp. 726–730.

[31] Al-Jadid M. Disability in Saudi Arabia. Saudi Med J. 2013;34(5):453-60.

[32] B. R. K. Mantha, C. C. Menassa, V. R. Kamat, and C. R. D’Souza, “Evaluation of Preference- and Constraint-Sensitive Path Planning for Assisted Navigation in Indoor Building Environments,” J. Comput. Civ. Eng., vol. 34, no. 1, p. 04019050, Jan. 2020, doi: 10.1061/(ASCE)CP.1943-5487.0000865.

[33] “EMOTIV Insight Brainwear® 5 Channel Wireless EEG Headset.” [Online]. Available: https://www.emotiv.com/insight/. [Accessed: 07-April-2021].

[34] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,” in Advances in psychology, vol. 52, Elsevier, 1988, pp. 139–183.

[35] L. Bi, F. xin’an, and Y. Liu, “EEG-Based Brain-Controlled Mobile Robots: A Survey,” Human-Machine Systems, IEEE Transactions on, vol. 43, pp. 161–176, Mar. 2013, doi: 10.1109/TSMCC.2012.2219046.

[36] S. Desai, S. Mantha, and V. Phalle, Advances in smart wheelchair technology. 2017, p. 7.

Computing Academics' Perceived Level of Awareness and Exposure to Software Engineering Code of Ethics: A Case Study of a South African University of Technology

Robert T. Hans[1]
Department of Computer Science
Tshwane University of Technology
Soshanguve, South Africa

Senyeki M. Marebane[2]
Faculty of ICT
Tshwane University of Technology
eMalahleni, South Africa

Jacqui Coosner[3]
Operations Department
Incusdata, Centurion
South Africa

***Abstract*—The need for awareness on ethical computing is increasingly becoming important. As a result this challenges all stakeholders in the software engineering profession, including educators, to improve their efforts on the awareness of professional codes of ethics which provide framework for ethical reference. However, the several compromises in the software engineering practice suggest that there are some in the profession, who are not familiar with the profession's codes of ethics and subsequently not able to practice and teach students about them. This research work investigates the extent of codes of ethics awareness by practitioners who are teaching software development courses in an academic environment. An online questionnaire with indicators for measuring awareness on software engineering code of ethics was deployed and responded to by 44 educators. Graphical, univariate and bivariate analyses were conducted on the data to determine the profile of the respondents and the extent of their level of awareness on the codes of ethics. The results indicate that majority of the lecturers (54.5 %) are not aware of software engineering codes of ethics, and those who are aware, majority of them were exposed to through self-study or personal development. Furthermore, the inclusion of codes of ethics in the learning activities is minimal as inhibited by lack of awareness and failure to apply the codes practically. This study recommends that lecturing staff as part of the professional software engineers serving as academic corps, should be placed on programmes for exposing them to professional software engineering codes of ethics. Moreover, the study calls for accreditation of software engineering courses, as it is the case with other professional engineering disciplines, to improve awareness and subsequent practical application of the codes of ethics.**

Keywords—Software engineering ethics; code of ethics; ethics awareness; ethics education; moral development ethics

I. INTRODUCTION

Awareness of the ethical implications of computing by educators in software engineering is increasingly becoming important. This is due to the responsibility educators have to prepare students to practice the trade in a way that reflects ethicality in their technical work [1]. Therefore, educators' role is significant to heighten the efforts in the teaching of ethics in technology courses [2] especially that they are also regarded as software engineers with a duty to promote the profession [3]. Improved efforts in education specifically about the ethical implications of developing software products are more required, as software has evolved to a de facto virtual resident in all areas of human lives due to the indispensable role it plays in devices and systems used by society. Software has contributed immensely to humanity, starting from promotion of human rights through social media [4], operation of machines in dangerous zones such as underground mines [5], saving of lives through computer aided surgeries [6] and usage of artificial intelligence to discover and discern patterns useful in understanding complex situations [7]. However, the list of software ethics violations such as wrongful use of software [8], deployment of badly designed and insufficiently tested software that created catastrophes such as aircraft crashes [9], [10], social media platforms used to drive misinformation for propaganda [11] even misinformation on global health pandemic such as Covid-19 [12] is continuing to haunt the province of computing ethics education. It specifically challenges the capacity of teaching software engineering ethics about the ethical awareness of those who teach software engineering and their ability to practicalise ethics into the curriculum.

Software engineering endeavours consider ethics education as an integral part of the moral development agenda, which, positively contribute to ethical computing awareness [1], [13]. For the software engineering ethics education to flourish, those involved in the teaching should be aware of the codes of ethics [3] as ethics navigation tool, infrastructure and content for teaching ethics. They should also create learning opportunities to assist students to practically internalize the codes alongside their other sources of ethical reference such as personal codes [14]. These opportunities can provide a rich experience on ethical awareness and ethical pluralism. Several frameworks for teaching ethics in computing advocates for the integration of codes of ethics into the curriculum [15]–[19]. These codes of ethics express a profession's disposition on ethics and professionalism, spells out principles to guide professionals and educates the professionals about what the society should hold them accountable for [20]. Therefore, the embracing of ethical codes by those involved in the education of software development can significantly contribute to the improvement of ethical awareness of the future graduates. It is on that note that Barnard *et al* [17] advocate that the teaching of ethics concepts and actions that underpin them should be integral part of the training of any future ICT professional. Therefore, those involved in the teaching of software development should be aware of the profession's ethics so they that can be in a position to teach the students accordingly.

The purpose of this study is to establish the extent to which university lecturers who are involved in teaching of software development courses in a South African University of Technology (UoT) are aware of software engineering ethics code. To help achieve this purpose the following research question and objective were formulated:

Research question:

What is the computing academics' level of awareness and exposure to a code of ethics specific to software engineering?

Research objective:

To determine computing academics' level of awareness and exposure to a code of ethics specific to software engineering.

This study helps to understand the extent of awareness on code of ethics in environments for teaching software development courses by measuring awareness by lecturers. Based on the awareness results, lecturers, industry trainers and employers will have better understanding of their insufficient knowledge of codes of ethics and the subsequent need for improvement. It will also help lecturers in faculties and organizations providing training on software engineering courses to realize the significance of using codes of ethics and the alignment thereof to real life incidents of ethical issues (good and bad), as well as including the codes in the constructive alignment activities related to teaching and learning of software engineering.

II. Literature Review

A. *The Need to Teach Ethics and the Role of Education on Moral Development*

The need for ethics in computing dates back to the days of the need for analysing the nature and social impacts of computing technology, and policies to clarify the ethical use of such technology [21]. However the challenge has always been the lack of practical knowledge and skills on how to apply ethics when faced with dilemmas [22]. University lecturers form an integral part of the ethics awareness program through the role they play in teaching software engineering ethics; and also serving as ethical examples. Therefore, institutions of learning should form the centre of efforts to be evaluated on their effect on moral development of students [23], especially faculty staff who participates in the teaching of ethics in educational environments.

Academic institutions like UoTs contribute to the software engineering business by providing education to graduates who work in the industry. Therefore, it is important for university lecturers to be aware of codes of ethics and include them as part of content used to orientate students about the expectations of software engineering profession in their teaching activities. Though it seems difficult and impossible, Kohlberg and Hersh [23] submit that the teacher bears the responsibility to guide the student to embrace practical moral conflicts, consider various sources for ethical reasoning, contrast personal thinking inconsistencies and inadequacies and means to address them. Several authors identify numerous instructional strategies for adoption in teaching ethics, such as online and face to face classroom discussions [24], writing of codes of ethics [25], service learning and civic engagement [26] case studies [27], evaluation of own project design [28], discussion on case studies and codes of ethics [19] and a combination of various strategies amongst others.

B. *Teachers as Facilitators of Ethical Awareness*

Research has proven the lack of required skills by ICT professionals to adequately apply moral judgements at work [29]. This challenges the adequacy and effectiveness of teaching computer ethics. Teaching staff as software engineers themselves, are obligated to promote ethical and professional approach to the practice [3], [18] through curriculum endeavours and conducting themselves as real life examples. Curriculum guidelines place special emphasis on software engineering teachers to take responsibility to expose students to codes of ethics as an instrument for developing professionalism and ethical competence [30], [31].

Though Harris and Lang [26] implore teachers to ponder on the ill-preparedness of graduates on ethics, several schools of thoughts have emerged regarding the teaching of computing ethics. There are divergent views as to the background of the teachers to be involved in the teaching of ethics and the instructional models to be adopted in teaching computing ethics [2], [32]. In response to these questions, in his essay Tavani [32] captures scholarly views of intellectuals like Jonhson, Dianne and Gotterbarn, whose works provided a detailed engagement on these questions.

Firstly, on the question of who can teach ethics better, it is believed that if the teaching of ethics is sourced from outside the faculty such as from philosophy experts, students may not seriously regard the subject as it may not appear to be mainstream [33]. Secondly, some believe that computer scientists lack knowledge in philosophy from which ethics derive existence, which may deny students opportunities to learn from the best whilst some believe that philosophy teachers lack technical knowledge in computing [32]. Monzon and Monzon-wyngaard [33] points out that both approaches are insufficient to integrate ethics approach.

Following on the question of instructional model, some schools of thought advocate for a single compound course that solely focuses on ethics whilst others prefer for the ethics content to be spread across several computing modules according to the applicability of the topic [32]. However, studies demonstrated that a multidisciplinary approach to the teaching of ethics can immensely enrich students' experience. For example, a study by Reich *et al* [34] which involved instructors from different faculties and industry experts resonated well with the course outcomes. Furthermore, Huyck *et al* [25] show as part of diversifying the teaching of ethics, that inclusion of students from various courses in ethics course generates positive outcomes. In support, Towell and Thompson [27] and Skirpan *et al* [28] respectively show that the functioning of an integrated instruction to ethics education yields success. The fact is, relevantly qualified teachers should be allocated to the teaching of software engineering ethics, sufficient time and faculty strategies should support the development of such ethical reasoning in the software engineering graduates.

To conduct ethics education instruction, the teacher needs to be aware of ethical and social issues technology is likely to confront the society with [17]. Therefore, such awareness should go beyond just the knowledge of principles prescribed by the codes of ethics, but also competently lead students through the ethical analysis learning process [35]. Furthermore, the views on ethics and an inclination to posture in a particular way to ethics is dependent on an individual's self-worth [36], hence it becomes deducible that a teacher who is ethically aware is likely to teach ethics to students better.

C. Research Advances on Software Engineering Code of Ethics Awareness

Due to the evolving ethical concerns in computing, research efforts on ethics awareness and professionalism remain relevant, as observable from previous studies conducted on this subject by [36]–[41]. Factors considered in the studies as having effect on ethics awareness on practitioners included individual characteristics such as age or maturity [42], [43], membership to a professional body [19], company commitment or leadership [44]–[46], and communication and enforcement [39] amongst others. Awareness of ethical codes helps to shape a computing professional, with their inclusion in education being the leading driver of ethical awareness [42].

There are other studies on ethical awareness that were conducted outside the academic environments and they include a study by Valentine and Barnett [37], which was carried out in a sales environment and found that awareness of codes is likely to improve organizational commitment and employees exposed to ethics codes view their work environments as being ethical. For the codes of ethics to have impact, [39] indicate that a relationship needs to exist between the codes of ethics awareness, their communication and their enforcement. Therefore, organizations need to demonstrate ethical leadership and commitment [44].

In the academic environment, a study by Cheng [47] sought to understand lecturers' perceptions toward teaching business ethics, and it found that teachers with greater self-efficacy perceive themselves as inclined to teach ethics in their domain. In relation to software engineering a study by Towell [19] found that majority of educators were aware of professional codes of ethics, and belonging to professional body appeared to be key to the awareness and promotion of such codes of ethics. A follow-up study by [27] revealed that whilst majority of educators were aware of the codes, 41% indicated that teaching of ethics was largely ignored, possibly due to lesser self-efficacy.

Several studies such as those conducted by [25], [48]–[51] studied ethical awareness on students but do not determine if those who are instructing the students, do have the ethical awareness required to teach the students and what their impact is on the students. Barnard *et al* [17] have demonstrated that computing instructors regard the teaching of computer ethics as importantly equal to the teaching of technological topics. We, therefore find it important to determine the extent to which lecturers involved in the teaching of software development are aware of software engineering codes of ethics, which are necessary and in some cases prescribed to instruct students. Furthermore, a study by Rogerson [52] shows that only 3% of professionals in the software development industry belong to a professional body, which implies that there is a possibility that there may be many university lecturers teaching in computing within the 97% who are not aware of the codes of ethics because they do not belong to a professional body.

This research study extends on the above studies as it analyses an environment of teaching software development, particularly in a UoT, in order to determine the extent to which a faculty is aware of software engineering codes of ethics.

III. Research Methodology

Online survey was utilized to collected data used in this study. The collected data formed part of a research project aimed at establishing software engineering ethical awareness climate in South African software development environments, including teaching environments. The data collection instrument was pilot tested with few targeted participants to ensure that the respondents interpreted and understood survey questions correctly. Following the pilot test of the data collection instrument, lecturers from two computing departments in a university of technology were individually invited through emails to participate in the study over a period of 6 months. A total of 103 email invites were sent out to staff members of the two departments, however, 44 participated in the online survey, resulting in a response rate of 43%.

A number of questions were posed to participants to determine their level of awareness of software engineering code of ethics. To ensure that respondents were the intended ones, only participants who chose a job description of a university computing lecturer were allowed to respond to the rest of the questions of the survey.

The study used graphical, univariate and bivariate analyses on the collected data in order to provide answers to the research question posed earlier. Due to the limited number of responses, inferential statistics, which could have enriched the findings of this study, was not be applied that. The next discussion presents the research results of the study.

IV. Research Results

A. Participants' Profile and Job Descriptions

Fig. 1 shows the profile of the 44 participants of this study. The data analysis results revealed that the majority (just over 77%) of participants were males, while females and those who chose not to specify their gender made up the remainder of the participants. On the other hand, 68.2% of the respondents were between 30 and 39 years of age, while 18.2% and 13.6%

of them were between 40 and 49 as well as between 50 and 59 years of age respectively, as shown in Fig. 2.

Seventy five percent (75%) of the participants had a post-graduate qualification; while 18.2% were holders of a degree qualification (see Fig. 3). The remaining 6.8% was split amongst participants who had a doctoral degree, a diploma and those who preferred not to divulge their qualifications. According to Fig. 4, the respondents who had more than 10 years of lecturing experience were 29.5%, followed by those who had between 3 and 5 years of work experience at 25%. The staff members who had between 1 and 2 years of work experience made up 15.9% of the participants, while the lecturers (educators) who had less than 1 year of work experience as well as the ones who had no lecturing experience were 13.6% each of the respondents.

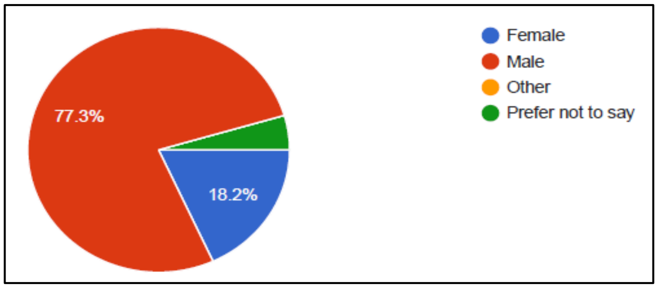


Fig. 1. Gender Distribution of Participants.

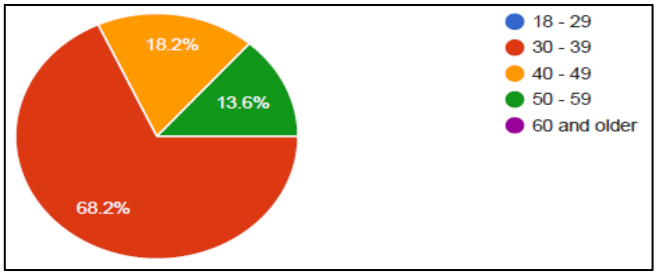


Fig. 2. Age Group of Participants.

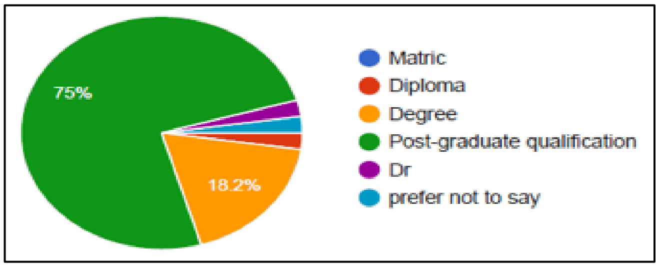


Fig. 3. Qualification Levels of Participants.

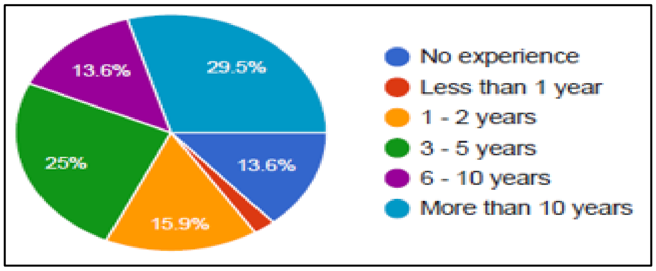


Fig. 4. Work Experience of Participants.

According to Fig. 5, the majority (52.3%) of the participants were holding lecturer positions, while those who held junior lecturer positions were 36.4%. About nine percent (9.1%) of the respondents were senior lecturers, while the remaining 2.2% preferred not to disclose their positions. Since there were participants who chose not to indicate their positions, it is therefore difficult to tell whether the two departments had participants who held professorship positions. If they did, it would have been very few individuals. Fig. 6 shows that only 11.4% of participants had membership with professional bodies, which included Institute of Electrical and Electronics Engineers (IEEE), Engineering Council of South Africa (ECSA), Institute of IT Professionals South Africa (IITPSA) and International Association of Engineers (IAENG). The notable part of these results is the lack of affiliation of any female educator to the professional organizations, as shown in Table I.

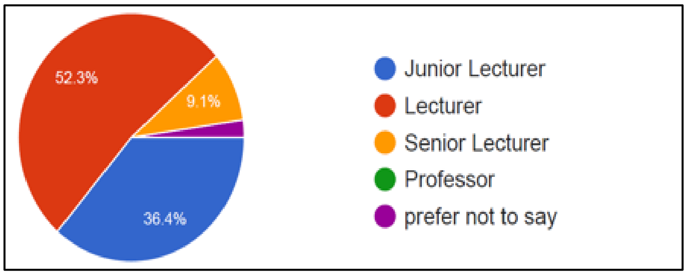


Fig. 5. Job Levels of Participants.

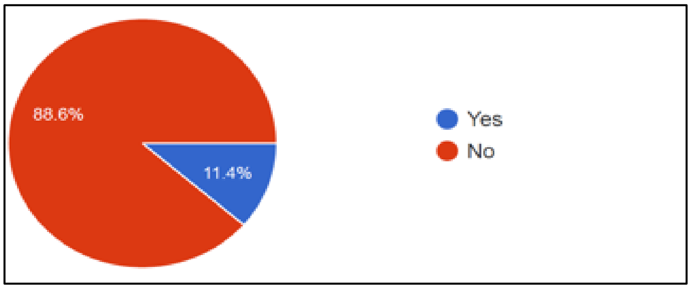


Fig. 6. Membership of Professional Bodies.

TABLE I. MEMBERSHIP OF PROFESSIONAL BODIES

	Gender	Responses	Frequency	Percentage
Membership of professional bodies	Female	No	8	100
		Yes	0	0
	Male	No	29	85.29
		Yes	5	14.71
	Prefer not to say	No	2	100
		Yes	0	0
		Total	**44**	**100**

B. The Level of Awareness and Exposure to a Code of Ethics Specific to Software Engineering

This study used the following measurements to gauge the level of awareness and exposure to software engineering code of ethics:

- Awareness of software engineering unethical incidents reported on the media.
- Previously been made aware of software engineering code of ethics, and if so,
  - How and where did the participant become aware?
  - Inclusion of software engineering code of ethics in current curriculum.
  - Presented software engineering code of ethics related topics.
  - Inclusion of code of ethics in modules offered by colleagues of respondents.
  - Discussions with students about current events related to code of ethics.
  - Code of ethics related learning outcomes in the content taught by participants.

The following discussion presents the responses of the participants, as shown in Table II (see next page), for each of the abovementioned measurements of code of ethics awareness.

TABLE II. RESPONSES OF PARTICIPANTS TO QUESTIONS REGARDING THEIR AWARENESS AND EXPOSURE TO SOFTWARE ENGINEERING CODE OF ETHICS

The level of awareness and exposure to a code of ethics specific to software engineering	**Measurements**	**Responses**	**Frequency**	**Percentage**	**Cumulative**
	Awareness of software engineering unethical incidents reported on media	No	22	50	50
		Yes	22	50	100
		Total	**44**	**100**	
	Previously been made aware of software engineering code of ethics	No	24	54.55	54.55
		Yes	20	45.45	100
		Total	**44**	**100**	
	How and where did the awareness occur	Through self-study and personal development	6	30	30
		Through their tertiary education	5	25	55
		Through short courses or workshops	4	20	75
		Through company policies	2	10	85
		Through membership to external professional bodies	2	10	95
		I don’t know or can’t remember	1	5	100
		Total	**20**	**100**	
	Inclusion of software engineering code of ethics in current curriculum	No: there is no such content in any content	9	45	45
		Yes: it is included in a compulsory course content	8	40	85
		Yes: it is included in an optional or elective course content	3	15	100
		Total	**20**	**100**	
	Presented software engineering code of ethics related topics	No	10	50	50
		Yes	10	50	100
		Total	**20**	**100**	
	Inclusion of code of ethics in modules offered by colleagues of respondents	No	7	35	35
		Yes	13	65	100
		Total	**20**	**100**	
	Discussions with students about current events related to code of ethics	No	7	35	35
		Not sure / Do not remember	5	25	60
		Yes	8	60	100
		Total	**20**	**100**	
	Code of ethics related learning outcomes in the content taught by participants	No	6	30	30
		Yes	14	70	100
		Total	**20**	**100**	

1) Awareness of software engineering unethical incidents reported on media: The respondents were split in the middle in their responses on their awareness of unethical incidents reported on the media – 50% said they were aware, whereas the other 50% claimed they were not aware.

2) Previously been made aware of software engineering code of ethics: The majority of the participants (54.5%) indicated that they were previously not made aware of the code of ethics pertaining to software engineering, while the remaining 45.5% percent said they were made aware. Those who indicated that they were made aware were asked follow-up questions, whose answers are presented next. When the results were interrogated little further, the indication was that, senior staff members (lecturers and senior lecturers) were the most ones (18 of the 27 = 67% of them combined) who were not previously aware of software engineering code of ethics, as shown in Table III.

3) How and where did the awareness occur: Thirty percent (30%) of the participants said they became aware through self-study and personal development, while 25% of the respondents said they became aware through their tertiary education (see Fig. 7). Twenty percent (20%) of the respondents said they were made aware through short courses or workshops that they attended, while 10% of each of the remaining 20% indicated that company policies and membership to external professional bodies played a role in the awareness process respectively. The remaining 5% said they could not remember or did not know.

4) Inclusion of software engineering code of ethics in current curriculum: Sixty percent (60%) of the participants indicated that the code of ethics or any topics related to it were not included in the courses/modules that the participants were involved in (45% of the participants) or if it was included, it was included in an elective module (15% of the participants). Forty percent (40%) purported that it was included in a compulsory course content.

5) Presented software engineering code of ethics related topics: The respondents were split in the middle – 50% said they once conducted a lesson on software engineering code of ethics related topic(s), while the other 50% indicated that they never did.

TABLE III. Level of Ethical Awareness by Job Level

The level of awareness and exposure to a code of ethics specific to software engineering	**Job level**	**Responses**	**Frequency**	**Percentage**
	Junior lecturer	No	6	37.50
		Yes	10	62.50
	Lecturer	No	15	65.22
		Yes	8	34.78
	Senior lecturer	No	3	75
		Yes	1	25
	Prefer not to say	No	0	0
		Yes	1	100
		Total	**44**	**100**

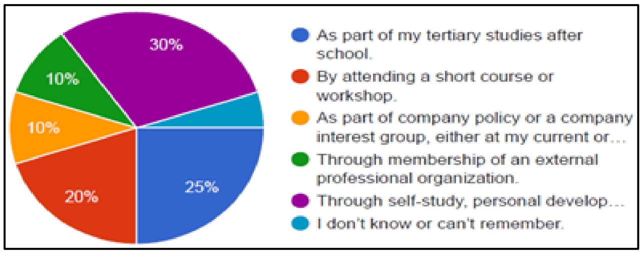


Fig. 7. Source of Ethics Awareness for Participants.

6) Inclusion of code of ethics in modules offered by colleagues of respondents: The majority (65%) of the participants said that they were aware that subjects that were offered by their colleagues had content on software engineering code of ethics, while 35% of the participants mentioned that they were unaware of such inclusion.

7) Discussions with students about current events related to code of ethics: Only 40% of the respondents could confirm that they had discussions with their students on recent (last 12 months) events related to software engineering code of ethics or code breaches. Sixty percent (60%) either never had such discussions (35%) or were not sure or could not remember (25%).

8) Code of ethics related learning outcomes in the content taught by participants: Seventy percent (70%) of the respondents mentioned that they were aware of software engineering code of ethics related learning outcomes (LOs) in the content they taught, while on the other hand, 30% of the participants were not aware of such LOs.

V. Discussion of Research Results

A. *Participants' Gender, Age and Job Profile*

The gender and age profile of participants show that the workforce in the two departments, have high gender imbalances and the workforce is relatively young. The job profile of the respondents indicates that the majority (70.5%) of them held lecturer positions and had no more than 5 years of experience and thus relatively inexperienced. These results should be expected given that the educators were relatively young as mentioned above. Almost 90% of the participants (including all female participants who identified themselves as such) were not members of any professional organization, which promote software code of ethics. Even though these figures are better than those mentioned in a study by Rogerson [52] they are still a cause for concern. This should be worrying because there are (should be) benefits (such as developing code of ethics awareness and individual development) for belonging to such professional bodies. However, the role of the professional bodies in providing code of ethics awareness to their members seems to be minimal given the fact that the results of this study show that only 10% of those who had subscription with these professional organizations received ethical awareness through them. This calls for improved code of ethical awareness campaigns from the professional organizations. Moreover, the professional organizations should have targeted drive to recruit female members.

B. The Level of Awareness and Exposure to Software Engineering Ethics Code

The survey results of this study show that more than half (54.5%) of the educators were not aware of software engineering code of ethics prior to this study, yet they are entrusted with the responsibility of educating students on the same code of ethics. Moreover, the analysis results revealed that the majority of the academics who were unaware of code of ethics held lecturer and senior lecturer positions (see Table III), which should be a concern given that these educators should be the guiding figures on issues of ethics. How can an educator, who is not aware of ethics, be able to lead a student to develop competent ethical reasoning? Furthermore, half of the academics were not aware of media reports of unethical incidents, and this should be worrying because a knowledge of such incidents should enable educators to practicalise the concept of ethics in lecture halls, thus allowing students to internalize and relate with the concept much better. Relating the incidents to students would help ease the difficulty of teaching ethics concepts to students. The awareness and knowledge of unethical incidents would not only demonstrate knowledge of violated principles prescribed by the codes of ethics, but would also assist in leading students through the ethical analysis learning process, as indicated by [35]. The lack of awareness of ethics code could mean such educators are unable to link ethics to practical day-to-day happenings that involve issues of software engineering code of ethics, resulting in failure to make students aware of their ethical responsibility as future software engineers.

There were three leading sources of awareness for the academics who were aware of software engineering code of ethics, and these were self-study and personal development (30%), tertiary education (25%) and short courses or workshops (20%). The organizational policies and professional organizations accounted for 10% each. However, even though tertiary education was amongst the three leading sources of code of ethics awareness it only accounted for a mere 25%, thus showing that tertiary education has little impact in this regard, bringing into question the level of coverage of software engineering code of ethics by our institutions. The majority (60%) of the educators indicated that the code of ethics or topics related to it were not included in the courses/modules that the participants were involved in (45% of the participants) or if it was, it was included in elective modules (15% of the participants). This finding and our observation on ethics coverage by our institutions of higher learning concur with the findings by Marebane and Hans [53] that the coverage of software engineering code of ethics by computing curricula of South African UoTs is inadequate. The preceding discussion has also brought into spotlight the lack of awareness of the code of ethics by the educators who are expected to be crusaders of the awareness campaign. It could also be possible that even though the code of ethics is covered by some modules/subjects in the institutions' curricula (65% of the educators indicated that code of ethics was covered by modules offered by their colleagues), but due to educators' lack of awareness (knowledge) of the code of ethics then the code of ethics topics do not get covered in the teaching. The results also show that of those that were aware of the code of ethics, only 50% (10 educators) of them ever lectured on the subject, meaning that 34 educators were either not aware of the code of ethics (24 educators) or never presented any lecture on code of ethics (10 educators). Another indication that code of ethics seems not to be receiving necessary attention from the educators is the fact that only 40% of those who were aware of code of ethics had a discussion in the last 12 months on code of ethics related events, the other 60% either never had a discussion or could not recall. An interesting finding of this study is that even though 70% of those who were aware of code of ethics knew about the learning outcomes in the content they taught only 50% of them ever presented a lecture on the subject, yet another signal of inadequate coverage of the topic by the educators. These findings reveal and confirm that indeed software engineering education receive marginal attention from our educators and curricula, as also purported by [18].

VI. Conclusion

This study set to answer and meet the following research question and objective respectively:

- What is the computing academics' level of awareness and exposure to a code of ethics specific to software engineering?
- To determine computing academics' level of awareness and exposure to a code of ethics specific to software engineering.

Two major findings of this study, which were presented in the previous section provided an answer to the study's research question and also met its objective. The two findings are:

- The worrying lack of awareness or exposure on software engineering ethics codes by the majority of the educators before this research study. This may be the possible cause for such lecturers not to be able to recognise the importance of including ethics in their curriculum, if they do include, they may fail to teach the principles contained in the codes sufficiently. Secondly, this lack of awareness may inhibit or disable the lecturers' ethical radar. Unfortunately, tertiary education played a minor role in educating lecturers (*who were once junior students*) on principles of software engineering code of ethics. This brings into question the level of coverage of code of ethics by tertiary institutions in their computing programmes. What makes this question even more valid and solid is that majority of the educators mentioned that software engineering code of ethics or topics related to it were either not included in the courses/modules that the participants were involved or were included in elective modules
- Half of the academics were not aware of unethical incidents that were reported on media. This could be the consequence of a disabled ethical radar, which subsequently will prevent lecturers from using such practical examples in their teaching, consequently

depriving students from learning from practical examples.

The concern that these findings highlight is the fact that the same educators who lack awareness of code of ethics are expected to be at the forefront of making students aware of the same codes they knew little about. This limits the ability to practicalise and relate the incidents to the curriculum and codes of ethics. How can they effectively fulfil their moral development educators' role of teaching ethics in technology courses, and to promote the ethicality of the profession when they lack ethical awareness? It is an imaginable expectation that academics who are entrusted with the responsibility of teaching students on the concepts of code of ethics of a profession are themselves in need of such education. To remedy the situation, institutions of higher learning should initiate or promote activities for professional code of ethics awareness for computing lecturers. Furthermore, lack of awareness by educators is likely to graduate software engineers who will fail to behave ethically and degrade the public view of the profession. Collaboration between universities and software engineering professional bodies on ethical practices and accrediting courses for software engineers can significantly improve awareness and ethical practices in the training of future software engineers.

VII. Limitations and Future Studies

The results of this study are based on self-reporting from participants and this has a possibility of biasness. It was also indicated in Section 4 that the number of responses to this study made it difficult to use inferential statistical, which could enrich the findings of the study, however this presents an opportunity for possible future study, which fulfils this requirement.

References

[1] Huff and A. Furchert, "Computing Ethics Toward a Pedagogy of Ethical Practice," Commun. ACM, vol. 57, no. 7, pp. 25–27, 2014, doi: 10.1145/2618103.

[2] C. Fiesler, N. Garrett, and N. Beard, "What dowe teach whenwe teach tech ethics? a syllabi analysis," Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE, pp. 289–295, 2020, doi: 10.1145/3328778.3366825.

[3] IEEE-CS, "Code of Ethics |IEEE-CS/ACM Joint Task Force on Software Engineering Ethics and Professional Practices," 1999. [Online]. Available: https://www.computer.org/education/code-of-ethics.

[4] S. Joseph, "Social Media, Political Change, and Human Rights Recommended Citation SOCIAL MEDIA, POLITICAL CHANGE, AND HUMAN RIGHTS," Boston College Int. Comp. Law Rev., vol. 35, no. 1, pp. 1–1, 2012, [Online]. Available: http://lawdigitalcommons.bc.edu/iclrhttp://lawdigitalcommons.bc.edu/iclr/vol35/iss1/3.

[5] T. Lecklider, "Autonomous mining equipment years ahead of car development," Evaluation Engineering, 2017. https://www.evaluationengineering.com/test-issues-techniques/technology/machine-vision/article/13015150/autonomous-mining-equipment-years-ahead-of-car-development (accessed Feb. 13, 2021).

[6] L. Joskowicz, "Computer-aided surgery meets predictive, preventive, and personalized medicine," EPMA J., vol. 8, no. 1, pp. 1–4, 2017, doi: 10.1007/s13167-017-0084-8.

[7] K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," Artif. Intell. Agric., vol. 2, pp. 1–12, 2019, doi: 10.1016/j.aiia.2019.05.004.

[8] A. Rashid, J. Weckert, and R. Lucas, "Software engineering ethics in a digital world," Computer (Long. Beach. Calif)., vol. 42, no. 6, pp. 34–41, 2009, doi: 10.1109/MC.2009.200.

[9] A. Buncombe, "Boeing 737 MAX: Company reveals new software problem detected in jets which 'must be fixed before planes can fly,'" Independent News, 2019. https://www.independent.co.uk/news/world/americas/boeing-737-max-jets-ceo-new-software-problem-a8855841.html (accessed Jun. 03, 2019).

[10] G. Travis, "How the Boeing 737 Max Disaster Looks to a Software Developer," IEEE Spectr., pp. 1–10, 2019, [Online]. Available: https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer.

[11] United Nations, "Human Rights Council Thirty-ninth session 10-28 September 2018 Agenda item 4 Human rights situations that require the Council's attention Report of the independent international fact-finding mission on Myanmar," 2018.

[12] C. Wardle and E. Singerman, "Too little, too late: Social media companies' failure to tackle vaccine misinformation poses a real threat," BMJ, vol. 372, 2021, doi: 10.1136/bmj.n26.

[13] C. Hanchey, "Yes, you can teach ethics!," J. Comput. Sci. Coll., vol. 17, no. 4, pp. 145–153, 2002.

[14] F. Ahmad, "Computer Science & Engineering Curricula and Ethical Development," in International Conference on Teaching and Learning in Computing and Engineering, 2014, pp. 220–225, doi: 10.1109/LaTiCE.2014.50.

[15] C. D. Martin and W. C. Huff, "A Conceptual and Pedagogical Framework," in Proceedings Frontiers in Education 1997 27th Annual Conference. Teaching and Learning in an Era of Change, 1997, vol. 1, pp. 479–483.

[16] C. Huff and D. Martin, "Computing Consequences: A framework for teaching ethical computing," Commun. Acm, vol. 38, no. 12, pp. 75–84, 1995.

[17] A. Barnard, C. de Ridder, L. Pretorius, and E. Cohen, "Integrating Computer Ethics into the Computing Curriculum: A Framework for Implementation," Proc. 2003 InSITE Conf., no. June, 2003, doi: 10.28945/2619.

[18] J. Jia and J. Xin, "Integration of ethics issues into software engineering management education," ACM Int. Conf. Proceeding Ser., pp. 33–38, 2018, doi: 10.1145/3210713.3210725.

[19] E. Towell, "Teaching ethics in the software engineering curriculum," in Software Engineering Education Conference, Proceedings, 2003, vol. 2003-Janua, pp. 150–157, doi: 10.1109/CSEE.2003.1191372.

[20] D. Gotterbarn, "How the new Software Engineering Code of Ethics affects you," IEEE Softw., vol. 16, no. 6, pp. 58–64, 1999, doi: 10.1109/52.805474.

[21] J. H. Moor, "What is computer ethics?," Metaphilosophy, vol. 16, no. 4, pp. 266–275, 1985, doi: 10.1007/BF00882026.

[22] J. Johnson, "Teaching Ethics to Science Students: Challenges and a Strategy," in Education and Ethics in the Life Sciences: Strengthening the Prohibition of Biological Weapons, B. Rappert., B. Rappert, Ed. Canberra: ANU E Press, 2010.

[23] L. Kohlberg and R. H. Hersh, "Moral Development : A Review of the Theory," vol. 16, no. 2, pp. 53–59, 1977, doi: 10.1146/annurev.ecolsys.3.

[24] K. Muskavitch, "Cases and goals for ethics education," Sci. Eng. Ethics, vol. 11, no. 3, pp. 431–434, 2005, doi: 10.1007/s11948-005-0011-6.

[25] M. Huyck, D. Ferguson, J. Ferrill, L. Getzler-Linn, and M. Raber, "Work in progress - Enhancing ethical awareness within undergraduate multidisciplinary teams by preparing Codes of Ethics," Proc. - Front. Educ. Conf. FIE, no. 978, pp. 21–23, 2008, doi: 10.1109/FIE.2008.4720383.

[26] A. Harris and M. Lang, "Incorporating Ethics and Social Responsibility in IS Education.," J. Inf. Syst. Educ., vol. 22, no. 3, pp. 183–190, 2011, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=10553096&AN=69713585&h=2IqrlxqAX3mY6CMHvX7E3jmFeST2qp7NoJEez765uemfcuUJ2gixNWDj%2F2l0iDZ0h1QTbbo%2FH1%2FdGOKV2HOnCQ%3D%3D&crl=c.

[27] E. Towell and J. B. Thompson, "A further exploration of teaching ethics

in the software engineering curriculum," in Software Engineering Education Conference, Proceedings, 2004, vol. 17, pp. 39–44, doi: 10.1109/csee.2004.1276508.

[28] M. Skirpan, N. Beard, S. Bhaduri, C. Fiesler, and T. Yeh, "Ethics education in context: A case study of novel ethics activities for the CS classroom," SIGCSE 2018 - Proc. 49th ACM Tech. Symp. Comput. Sci. Educ., vol. 2018-Janua, pp. 940–945, 2018, doi: 10.1145/3159450.3159573.

[29] Y. Al-Saggaf and O. K. Burmeister, "Improving skill development: An exploratory study comparing a philosophical and an applied ethical analysis technique," Comput. Sci. Educ., vol. 22, no. 3, pp. 237–255, 2012, doi: 10.1080/08993408.2012.721073.

[30] T. C. Lethbridge, R. J. Leblanc, A. E. Kelley Sobel, T. B. Hilburn, and J. L. Diaz-Herrera, "SE2004: Recommendations for undergraduate software engineering curricula," IEEE Softw., vol. 23, no. 6, pp. 19–25, 2006, doi: 10.1109/MS.2006.171.

[31] ACM/IEEE, "SE 2014: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering," Computer (Long. Beach. Calif)., vol. 48, no. 11, pp. 106–109, 2015, doi: 10.1109/mc.2015.345.

[32] H. T. Tavani, "Curriculum issues and controversies in computer ethics instruction," Int. Symp. Technol. Soc. Proc., vol. 2001-Janua, pp. 41–50, 2001, doi: 10.1109/ISTAS.2001.937720.

[33] J. E. Monzon and A. Monzon-wyngaard, "Ethics and Biomedical Engineering education : the continual defiance," in 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009, pp. 2011–2014.

[34] R. Reich, M. Sahami, J. M. Weinstein, and H. Cohen, "Teaching computer ethics: A deeply multidisciplinary approach," Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE, pp. 296–302, 2020, doi: 10.1145/3328778.3366951.

[35] Z. Khallouf, "About Integrating Ethics in The Software Engineering Curriculum Engineering Curriculum," 2007.

[36] D. M. Berry and B. Berenbach, "Ethics test results before and after ethics training: A disturbing experience," SwSTE2010 IEEE Int. Conf. Softw. Sci. Technol. Eng., pp. 70–76, 2010, doi: 10.1109/SwSTE.2010.16.

[37] S. Valentine and T. Barnett, "Ethics code awareness, perceived ethical values, and organizational commitment," J. Pers. Sell. Sales Manag., vol. 23, no. 4, pp. 359–367, 2003, doi: 10.1080/08853134.2003.10749009.

[38] J. Ballantine, M. Levy, A. Martin, I. Munro, and P. Powell, "An ethical perspective on information systems evaluation," Int. J. Agil. Manag. Syst., vol. 2, no. 3, pp. 233–241, 2000, doi: 10.1108/14654650010356149.

[39] K. Munro and J. Cohen, "Ethical Behavior and Information Systems Codes : The Effects of Code Communication, Awareness, Understanding, and Enforcement," ICIS 2004 Proc., 2004.

[40] J. R. Córdoba, "Developing ethical awareness in information systems practice: A Foucaultavian view," J. Information, Commun. Ethics Soc., vol. 4, no. 4, pp. 181–190, 2006, doi: 10.1108/14779960680000291.

[41] C. D. Martin and E. Y. Weltz, "From Awareness to Action : Integrating Ethics and Social Responsibility into the Computer Science Curriculum," ACM SIGCAS Comput. Soc., vol. 29, no. 2, pp. 6–14, 1999.

[42] S. H. Wilford and K. J. Wakunuma, "Perceptions of ethics in IS: How age can affect awareness," J. Information, Commun. Ethics Soc., vol. 12, no. 4, pp. 270–283, 2014, doi: 10.1108/JICES-02-2014-0013.

[43] L. N. K. Leonard, T. P. Cronan, and J. Kreie, "What influences IT ethical behavior intentions - Planned behavior, reasoned action, perceived importance, or individual characteristics?," Inf. Manag., vol. 42, no. 1, pp. 143–158, 2004, doi: 10.1016/j.im.2003.12.008.

[44] M. E. Brown, L. K. Treviño, and D. A. Harrison, "Ethical leadership: A social learning perspective for construct development and testing," Organ. Behav. Hum. Decis. Process., vol. 97, no. 2, pp. 117–134, 2005, doi: 10.1016/j.obhdp.2005.03.002.

[45] M. Schwartz, "The nature of the relationship between corporate codes of ethics and behaviour," J. Bus. Ethics, vol. 32, no. 3, pp. 247–262, 2001, doi: 10.1023/A:1010787607771.

[46] D. M. Mayer, M. Kuenzi, R. Greenbaum, M. Bardes, and R. (Bombie) Salvador, "How low does ethical leadership flow? Test of a trickle-down model," Organ. Behav. Hum. Decis. Process., vol. 108, no. 1, pp. 1–13, 2009, doi: 10.1016/j.obhdp.2008.04.002.

[47] P. Y. Cheng, "University Lecturers' Intention to Teach an Ethics Course: A Test of Competing Models," J. Bus. Ethics, vol. 126, no. 2, pp. 247–258, 2015, doi: 10.1007/s10551-013-1949-y.

[48] M. Masrom, Z. Ismail, and R. Hussein, "Computer ethics awareness among undergraduate students in Malaysian higher education institutions," in 19th Australasian Conference on Information Systems, 2008, pp. 628–637.

[49] M. Aliyu, N. A. O. Abdallah, N. A. Lasisi, D. Diyar, and A. M. Zeki, "Computer security and ethics awareness among IIUM students: An empirical study," in Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World: ICT Connecting Cultures, ICT4M 2010, 2010, pp. A52–A56, doi: 10.1109/ICT4M.2010.5971884.

[50] D. Bairaktarova and A. Woodcock, "Engineering Student's Ethical Awareness and Behavior: A New Motivational Model," Sci. Eng. Ethics, vol. 23, no. 4, pp. 1129–1157, 2017, doi: 10.1007/s11948-016-9814-x.

[51] L. Cilliers, "Evaluation of information ethical issues among undergraduate students : An exploratory study," South African J. Inf. Manag., vol. 19, no. 1, pp. 1–6, 2017.

[52] S. Rogerson, "Rebooting ethics education in the digital age." pp. 1–4, 2021, [Online]. Available: https://dora.dmu.ac.uk/bitstream/handle/2086/20613/Rebooting ethics education in the digital age prepublication version.pdf?sequence=2&isAllowed=y.

[53] S. M. Marebane and R. T. Hans, "International Journal of Advanced Computer Science and Applications (IJACSA)," Competency Gap Undergrad. Comput. Qualif. within South African Univ. Technol., vol. 12, no. 4, 2021.

Automating and Optimizing Software Testing using Artificial Intelligence Techniques

Minimol Anil Job
Faculty of Computer Studies, Arab Open University
Kingdom of Bahrain

***Abstract*—The final product of software development process is a software system and testing is one of the important stages in this process. The success of this process can be determined by how well it accomplishes its goal. Due to the advancement of technology, various software testing tools have been introduced in the software engineering discipline. The use of software is increasing day-by-day and complexity of software functions are challenging and there is need to release the software within the short quality evaluation period, there is a high demand in adopting automation in software testing. Emergence of automatic software testing tools and techniques helps in quality enhancement and reducing time and cost in the software development activity. Artificial Intelligence (AI) techniques are widely applied in different areas of Software engineering (SE). Application of AI techniques can help in achieving good performance in software Testing and increase the productivity of the software development firms. This paper briefly presents the state of the art in the field of software testing by applying AI techniques in software testing.**

Keywords—Software testing; artificial intelligence; testing automation; software engineering; software quality

I. Introduction

Software engineering is the creation and application of sound engineering principles to produce cost-effective software which is both reliable and functional on real machines. A well-managed development process is needed to produce a high-quality software product. Software development is a human endeavor which involves various activities. The activities are; analysis, design, implementation and testing and each of them contributes to the creation of the final product. These activities continue in the development process, and thus producing a working version of the system can be time-consuming. Software testing is one of the principal activities in software development for verifying and validating a software system. Testing assists software developers in ensuring that the developed software fulfills its intended function, as well as determining whether or not the identified problems have been solved. Since the software development life-cycle is a complex process with a crucial need to deliver a new product within the allocated time, the software testing process should be efficient and effective.

In the software industry, automation plays a critical role in increasing testing performance. There are various automation tools available to support the testing activity. Newer technologies like Artificial Intelligence (AI) and Machine Learning (ML) are constantly being used to speed up the software development process. With the advancement of AI technologies, various business domains are accepting and using AI based software. AI systems are developed based on machine learning models and techniques. AI is used to promote automation and reduce the amount of routine activities to create testing phases by applying logic, problem solving, and ML. The purpose of this paper is to present the application of AI techniques in automation of software testing and the impacts. The paper is closed with a conclusion and a viewpoint of future work to enhance the practical aspects of the AI automation tool.

II. Literature Review

Testing is an activity takes place throughout the software development life cycle. The major aim of the software testing is find errors and to ensure that the developed system satisfies the requirements of customers. Software testing a way of evaluating a system by discovering the differences between the identified requirements at the requirement engineering activity and the archived results. There are various techniques used for testing a software to ensure quality. Testing can be done as manual or automated using specific tools [20] [7]. In manual testing, the software tester follows a test plan and complete a set of test cases manually. Manual testing is time consuming and cost effective. The alternative of the manual testing is the automated execution in which a software program runs a pre-defined test cases and shows and saves the results. Automatic testing executes the test-case automation, test-case generation and result verification. [2]. Automation testing tools are used in automatic software testing and the reliability as well as performance are more than manual testing [10] [19]. Software testing using AI techniques has been adopted by software development companies all over the world. The application of AI is wide-ranging branch of computer science that deals with building smart machines which are capable of performing smart tasks with assistance of human intelligence [9].

AI has become an emerging technology which can be applied in ensuring quality assurance (QA) in organizations [26] [15]. AI developments require an approach to validation and verification of software. AI has becoming an important factor in quality assurance and testing in future. [14]. AI is a branch of science and engineering concerned with the computational analysis of intelligent behavior and the creation of intelligent machines. The term AI refers to a collection of tools, techniques, and algorithms [3] [11] [21]. Application of AI techniques are creating impacts in various domains like health, manufacturing etc. Emergence of big data and its management requires strong computing tools and AI helps in improving the efficiency of these activities. AI techniques are applied in different areas of software engineering such as requirement engineering, design and testing [24]. AI in software testing aims to make testing smarter and more efficient. AI and machine learning apply reasoning and problem solving to automate and improve testing. AI in software testing helps reduce time-consuming manual testing, so teams can focus on more complex tasks, like creating innovative new features. Faulty software is mission critical as software becomes an increasingly important component in a wide variety of systems. Since automated test generation techniques are free of the cognitive biases that have been found in human testers, they can be used for software testing [4].

Automating the execution process of the software testing cycle is the most common approach in the automation field. Automated software testing is important and adding more toolkits to make testing phases fully automated by generating test scripts is also equally important. Test cases can be generated to complete the automated testing [22]. The test execution speed will be increased by this tools and the testing process can be applied repeatedly. In manual testing method, developing test scripts is a time consuming process but if the test cases are ready the human tester can complete the testing process quickly [27] [25]. In manual testing, code visibility does not affect test code coverage and fault detection rate12]. It is possible to facilitate a process for eliciting testing requirements and creating test-suites [1]. Various algorithms are used in AI tools. Use of Bayesian probabilistic reasoning to model software reliability is an example probabilistic AI technique. This is used in the discipline of Software Engineering [18] [5] [21]. AI techniques can be applied in supporting automation and decreasing the amount tedious tasks in the software development and testing phase [17].

III. Software Testing

Software testing is one of the crucial stage in the software development life cycle and it is an important method to assess the developed software to ensure quality. Before releasing the final product, it is important to ensure that the requirements identified during the analysis stage have been accomplished as well as the product is defect free. In the context of software development, bug is a name given for the defect which means the developed system is not producing accurate results as per the requirements. Testing plays an essential part in the software development process. Various techniques can be used during the software development process to facilitate testing. Prototyping is one of the approaches in which an experimental software artefact will be developed and that will be discarded after evaluation. Early software artefacts are built on rather than discarded in the Iterative approach [23]. Another approach is using frameworks such as the dynamic systems development method (DSDM) in which best practice processes for iterative and incremental development are documented [13].

A. Categories of Testing

Software developers need to be aware and to be able to use a range of other testing methods, concepts and practices. The four distinct categories of testing are summarized in Fig. 1:

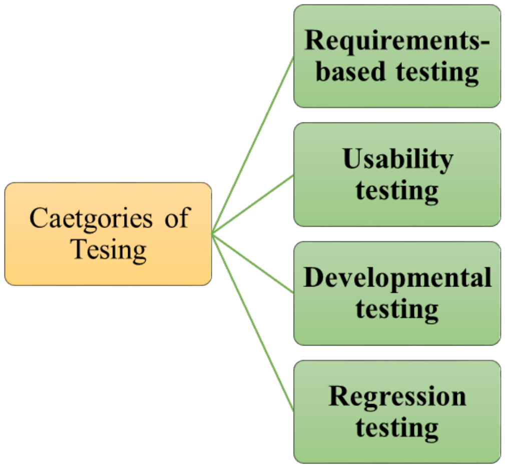


Fig. 1. Categories of Testing.

Table I presents the details of the four categories of testing shown in the above figure.

TABLE I. TYPES OF TESTING

REQUIREMENTS-BASED TESTING
➢ Checks that a system meets the customer's requirements using previously gathered or formulated testable requirements. ➢ Acceptance testing is the final stage to check that the user requirements have been satisfied. ➢ The customer formally accepts the software by ensuring the correct functioning of the acceptance tests
USABILITY TESTING
➢ User interface (UI) is an integral part in every software system. ➢ Various users other than the developers are using the system and the usability testing is essential. ➢ Due to the problems of user interface, the systems may fail. ➢ Usability testing comprises systematically trying out the user interface with intended users
DEVELOPMENTAL TESTING
➢ Refers to the testing carried out by the entire software development team. ➢ The developmental testing is done in three different levels which are unit testing, integration testing and system testing. ➢ Integration testing involves checking that all the tested units are interfaces together correctly.
REGRESSION TESTING
➢ Any form of testing done during the development or maintenance of a system to ensure that fixing one bug does not result in the introduction of new ones. ➢ are needed at unit, integration and system levels depending on the particular development method adopted ➢ Regression tests are critical during the developmental testing and in system maintenance.

B. Software Testing Techniques

Strategies for creating test cases is an important part of validation and verification. The two major strategies are black-box testing and white-box testing. The main objective of black box testing is to ensure each aspect of the customer's requirements is handled correctly by an implementation. The test cases are designed by looking at the specification of the system. The specification includes the details of the systems intended functions.

Grey-Box Testing is another method in which the application is tested with complete knowledge of the overall aspects of the system and limited knowledge of the internal functioning of the system. [12]. The major objective of white box testing is to check that the details of an implementation of the system are correct. The test cases are designed by looking at the detail of the implementation of the system to be tested to check that the system performs its intended functions correctly [8].

While comparing the possibility of automation of black-box and while-box testing techniques, automating white-box testing is easier than black-box testing. The reason is that box in black-box testing, programmer and the tests are dependent and this will affect the automation. Since grey-box testing provides the features of both black-box and white-box testing techniques, it's features and techniques are not considered in detail in this paper.

Fig. 2 depicts the five strategies used in black-box testing and four strategies used in white-box testing.

The different black-box testing techniques are summarized in Table II [17, 22].

The different approaches of White box testing are listed in Table III [17, 22].

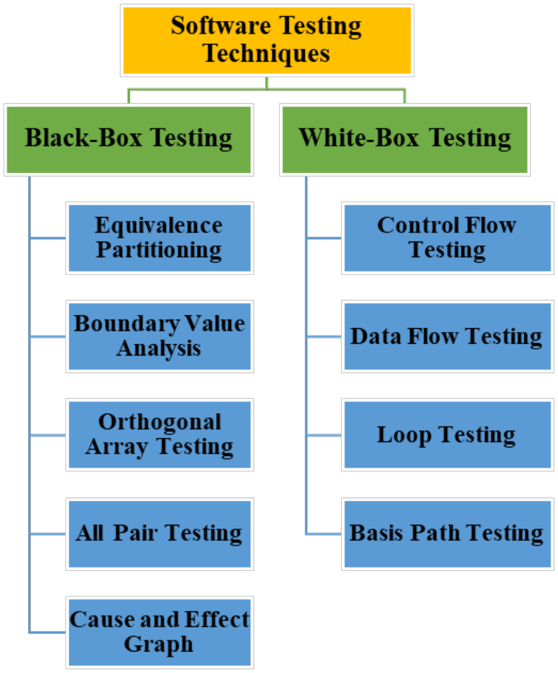


Fig. 2. Black-Box and White-Box Testing Techniques.

TABLE II. BLACK-BOX TESTING TECHNIQUES

Black-Box Testing Techniques
Equivalence partitioning
➢ This technique divides the input domain of the software unit into groups or partitions and generates test cases thus helps in reducing the number of test cases.
Boundary value analysis
➢ Errors at the boundaries of the input domain are tested in this technique. ➢ The boundary values are usually taken the minimum and maximum values in the boundaries and values just inside and outside of the boundaries.
Orthogonal Array Testing
➢ It's a strategy for problems with a limited input domain that are relatively small and cannot be tested thoroughly
All pair Testing
➢ In this method, test cases are created using all possible combinations of each pair of input parameters.
Cause and effect Graph
➢ In this approach, a graph is generated and a cause-and-effect relationship is formed. The cause is the input condition that leads to internal change in the system and the effect is the output condition.

TABLE III. WHITE-BOX TESTING TECHNIQUES

White-Box Testing Techniques
Control Flow Testing
➢ This technique uses a testing strategy in which control flow of the program is involved in the test coverage. Branch coverage, statement coverage and condition coverage are the three methods used in test coverage.
Data Flow Testing
➢ Data movement within the program is focused in this testing strategy and the test paths are determined from the positions of variable definition in the program. This technique helps to detect errors such as undefined variables etc.
Loop Testing
➢ Different types of loops such as simple loops, nested loops are used in most of the programs. Validity of the loops in the programs are tested using this technique.
Basic Path Testing
➢ Basis-path testing is based on the cyclomatic-complexity metric. The number of independent paths in a method body is computed using this metric.

IV. TEST AUTOMATION

The need to make sure that users are really receiving the value promised by the software, many powerful testing tools have emerged, which make it easy to do black box testing in an automated and repeatable manner. The automation is important because the testing process is repetitive and it is recommended to test all possible scenarios. Automating tests will increase the test coverage and improve efficiency. When the testing process is automated, the tests can be executing repeatedly, test different input values and various conditions. The testing resources and time will be reduced. To automate functional, system and acceptance tests, numerous tools are available. Some of them are Selenium, Watir and JMeter. Selenium and Watir are used to test local files using the file:// protocol supported by web browsers, however, JMeter needed the target files to be hosted on a web server.

V. Application of AI in Software Testing

Due to the increasing complexity and features, modern applications require various features in order to achieve the functional and non-functional requirements of the applications. Requirements are the information about what a system will be and do that needs to be known before development starts. Once the requirements are identified and agreed by the developer and the customer, they will be implemented in the software system. Converting the requirements into an application using appropriate tools and techniques is the role of the system developer. The application should be developed without errors and during the coding stage, the developer needs to write down and execute test cases. A software tester is a person who is in charge of putting an application through its paces and ensuring that it performs as intended. Software testers play a crucial role same as the developer in the software engineering activity as they are part of the quality assurance of the software. The software tester sends a report to the development team detailing the bugs found and the sequence of events that contributed to the mistake. As shown in Fig. 3, the Developer‘s responsibility of writing code and writing and executing test cases. The testers will follow the test plans and complete the testing tasks.

Software testing is an important process which ensures the developed system’s satisfaction from the customers. Testing helps to protect the system against any failure which may lead potential impact in organizations’ performance during the operational period. In manual testing, the software is tests are executed by the software tester to discover bugs in the system. Even though exploratory testing is possible in manual testing, it is a time-consuming process by involving human resources and test cases are executed by a human tester and software. Automated testing is considerably faster than manual testing and faster than a manual approach and uses automation tools to execute test cases. As testing is moving more and more towards automation, AI techniques are applied in software testing. AI techniques supports the automation of the testing processes in an efficient way. AI algorithms are supporting the testing environment to be more productive. The AI algorithm encourages the process and helps software testers to find the maximum number of bugs within less period of time. The system can be route to market as reliable and accurate. Fig. 4 and Fig. 5 show the functions of a tester and a developer in manual and AI enables testing environments.

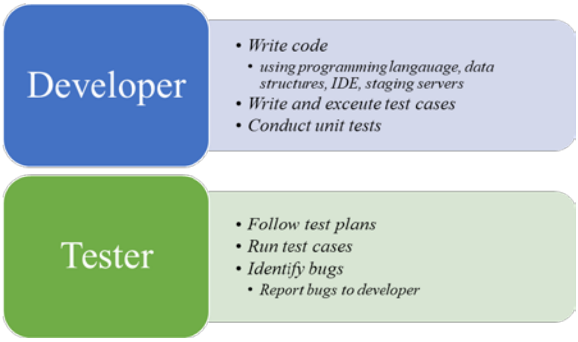

Developer
• Write code
• using programming langauage, data structures, IDE, staging servers
• Write and exceute test cases
• Conduct unit tests
Tester
• Follow test plans
• Run test cases
• Identify bugs
• Report bugs to developer

Fig. 3. Tasks: Software Developer Vs Software Tester.

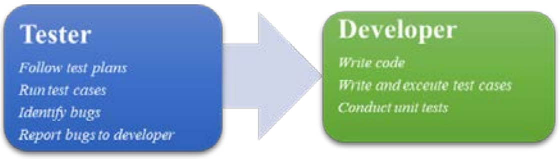


Fig. 4. Functions in a Manual Testing Environment.

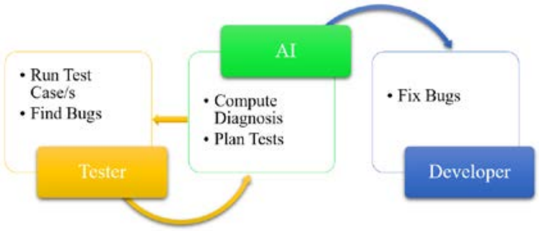


Fig. 5. Functions in an AI Enables Testing Environment.

In a manual testing environment, the tester runs the test cases according to the test plans, finds the bugs and reports them to the developer. In the AI based testing environment, the tester runs test cases and finds the bugs. The AI tool do the diagnosis and send notification to the developer to fix the bugs. The AI tool plans further tests automatically and the testers will continue running the test cases and identify bugs.

The functional and non-functional requirements of the product need to be tested by ensuring the full test coverage. In manual testing this will take time and resources and application of AI techniques will reduce the time and also will help to identify poorly designed requirements. Requirements will be thoroughly verified using AI and it will be transferred to the design stage, the next stage in the software development life-cycle. Based on the experience of the software testers, the test selection and quality will vary and the tests are created based on the requirements, use cases and user stories. Application of AI helps to automatically generate test cases and identify the tests for verification and validation of the system. Ai also help to understand the test coverage as well. Missing or inappropriate test can be identified based on the models. Ain addition to this, automatic generation of the test cases will also reduce any eluded bugs and defects also.

VI. Application of AI in Test Automation

Business organizations in various sectors are adopting AI techniques makes their operations efficient. In software development process, automation testing is widely used to improve the test efficiency. There are various ways AI techniques can support the software testing process [6] [16]. The most benefit takes from AI is the black box testing. Applications of AI techniques in software testing is listed.

Fig. 6 depicts the test automation in pyramid shape according to the test types. UI testing is placed in the highest level because and this is the most expensive testing. Test automation can be done in the lower levels both in API and Unit layers to achieve best test coverage and reduce time and cost.

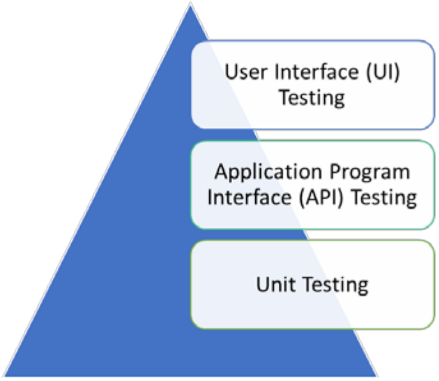

User Interface (UI) Testing
Application Program Interface (API) Testing
Unit Testing

Fig. 6. Test Automation Levels.

User Interface (U) testing: Usability tests are conducted to identify any design inconsistencies or usability issues in the user interface. To build UI tests, AI techniques use image recognition techniques to navigate through the application and visually validate UI objects and components. The test automation tools decode the Document Object Model (DOM) and related code to determine object properties in AI-based UI testing. The UI tests are done after the design and development of the software system. In the pyramid, UI tests are in the top level of the pyramid and the automation is challenging and they are the hardest to automate.

Application Program Interface (API) Testing: In API testing, validation of Application Programming Interfaces (APIs) is done by testing various software quality factors of the programming interfaces such as functionality, performance, reliability and security. Nonfunctional requirements like look-and-feel are not relevant in API testing and focuses on the business logic instead. Without any reference to the user interface, API automation can be used to validate the business logic. Automation of the API allows the software development team to start the testing activity early which helps to identify and fix errors and reducing the efforts in the AI testing.

Creating and updating unit tests: Unit tests are used to ensure that small pieces of code, such as functions or object methods, are correctly executed before they are combined with other units to form a larger functional feature. Developers spend a significant amount of time writing and maintaining unit tests, which is much less enjoyable than writing application code AI-based products for automated unit test development can be beneficial and they can be generated promptly. Automation strategy starts with unit tests. Unit tests are written by the developers using the same application development programming language. Early detection and correction of errors in the unit testing will prevents regression defects.

In addition to the above mentioned testing automation strategies AI can be applied in the following areas of the software testing also.

- AI is used as an adaptive method for detecting element-level changes that will improve the testing suites' robustness.
- aid in the prediction of incorrect test cases that result in total test failures, as well as include recommendations for resolving those issues.
- used in simulating Behavioral patterns.
    - AI techniques helps to simulate the behaviors of people in using the system by geography, devices and demographics as inputs to build test suites.
- AI algorithms are efficient in prediction and automation of test suites.
- used in script Automation.
    - Automating a test script is not required in AI because it is executed from the AI algorithm automatically.
- Helping in mining defects and based on this test suite can be defined. This speeds up the process of making informed decisions about test coverage and test suite optimization.
- Enhances number of tests and their scopes.
- AI techniques are applied in Automation Test Maintenance, Test Data Generation, early feedback in testing etc.

VII. AI Based Test Automation Tools

A. AI Bases Automation Tools

Table IV summarizes 5 AI based automation tools uses in software industry recently.

TABLE IV. AI based Automation Testing Tools

Testing Tool	Application
Testim	This AI tool is used to automate functional testing by using artificial intelligence and machine learning. This automated testing tool speeds up the authoring, execution, and maintenance of automated tests.
Functionize	A cloud based automated testing technology used for functional, performance and load testing. This automated testing tool speeds up the test creation, diagnosis, and maintenance.
Appvance	This AI enabled tool used for automating functional, performance and security testing. The AI integrated tool helps codeless test creation.
Applitools	This AI enabled tool used for User Interface (UI) testing. It provides an end-to-end software testing platform powered by Visual AI and used in Quality Assurance and test automation.
Testcraft	This AI-powered test automation platform supports regression testing and also for monitoring of web applications. The AI technology helps in eliminating maintenance time and cost.
Watir	This is an open Source tool, uses Ruby Libraries to automate tests. Watir is used for testing websites and uses Selenium for browser automation. In addition to this Watir also supports in writing stable and maintainable test scripts.

Functional tests are part of integration tests and designed to evaluate specific tasks. For example, user registration in an online system, the registration function should validate the data entry from the user. Other examples are catalogue search and online payment in an e-commerce application etc. A wide range of test values are required to perform these tests. Performance testing is used for measuring the ability of servers to respond to user demand in a networking environment. A load test is a method of evaluating a server's performance by applying a given load to the application in order to evaluate the response time for individual functions. The aim of security testing is to ensure that the application has authentication and authorization controls in place and is not vulnerable to attacks. User interfaces are usually tested for Navigation errors, presentation errors and control usage problems. Regression testing can reveal defects that recur as a result of software changes that were not planned.

B. Pactical Application: Using Watir

Watir (Web Application Testing in Ruby) has been selected as a software experimenting too to support the technical aspects of the research. l. The Watir supports almost all of the web browsers. The websites and user interfaces can be tested using Watir. The test framework by using Watir follows the following order as shown in Fig. 7.

After installing the required drivers, test scripts for opening browsers have been written in the RubyMine IDE. The test scripts are written in Ruby language. Watir is selected for the practical application because first of all, it is very easy to use open source tool. This tool is developed using Ruby and any web application can be automated irrespective of the browser it is running. There are in-built libraries in Watir to support various activities such as page performance. Fig. 8 depicts the test script for opening Chrome browser.

Fig. 9 demonstrates another sample of creation and execution of test case using Watir. A simple test has been used to test the accuracy of the code. The test script runs successfully.

Fig. 10 shows an example of a test script for the user interface testing. The sample HTML and JavaScript code is shown in Fig. 11 for the page.

The presentation of the web page is done using HTML and a JavaScript function is used for the static feature of the page code using HTML and Fig. 11 shows the code.

Fig. 12 shows the debugging screen of another sample test script with another browser, Firefox using Watir automation tool.

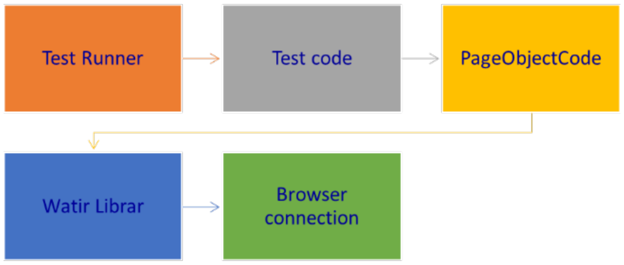
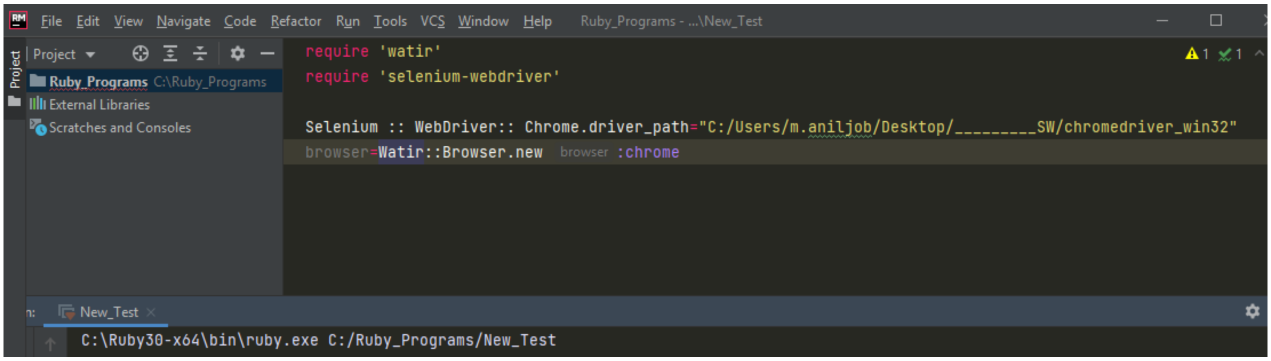


Fig. 7. The Order of Test Framework in Watir.

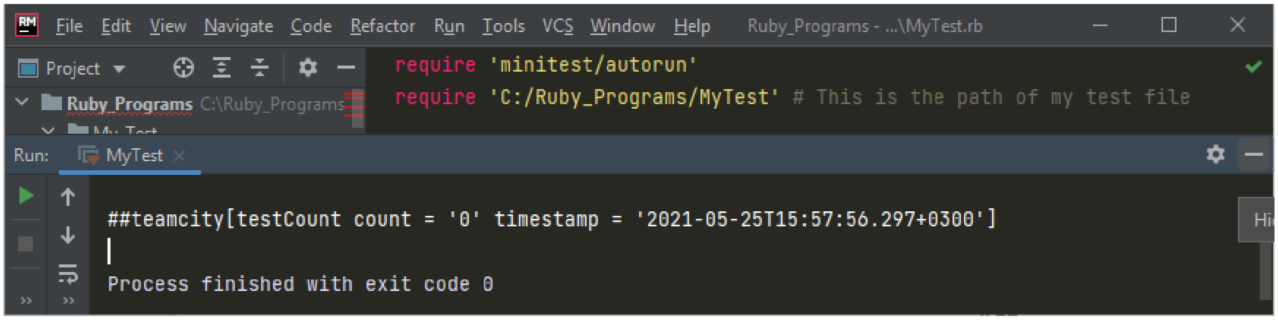

File Edit View Navigate Code Refactor Run Tools VCS Window Help Ruby_Programs - ...\New_Test
Project
Ruby_Programs C:\Ruby_Programs
External Libraries
Scratches and Consoles
New_Test

```
require 'watir'
require 'selenium-webdriver'

Selenium :: WebDriver:: Chrome.driver_path="C:/Users/m.aniljob/Desktop/_________SW/chromedriver_win32"
browser=Watir::Browser.new browser :chrome
```

```
C:\Ruby30-x64\bin\ruby.exe C:/Ruby_Programs/New_Test
```

Fig. 8. Test Script for Chrome Browser.

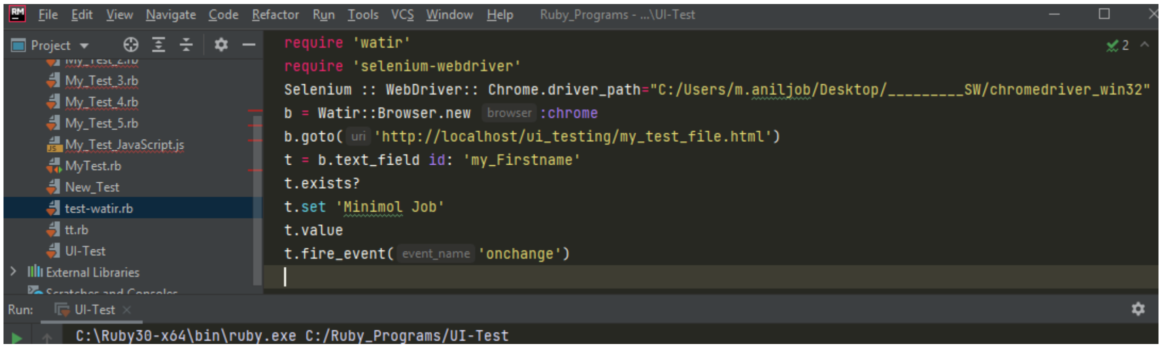

File Edit View Navigate Code Refactor Run Tools VCS Window Help Ruby_Programs - ...\MyTest.rb
Project
Ruby_Programs C:\Ruby_Programs
Run: MyTest

```
require 'minitest/autorun'
require 'C:/Ruby_Programs/MyTest' # This is the path of my test file
```

```
##teamcity[testCount count = '0' timestamp = '2021-05-25T15:57:56.297+0300']

Process finished with exit code 0
```

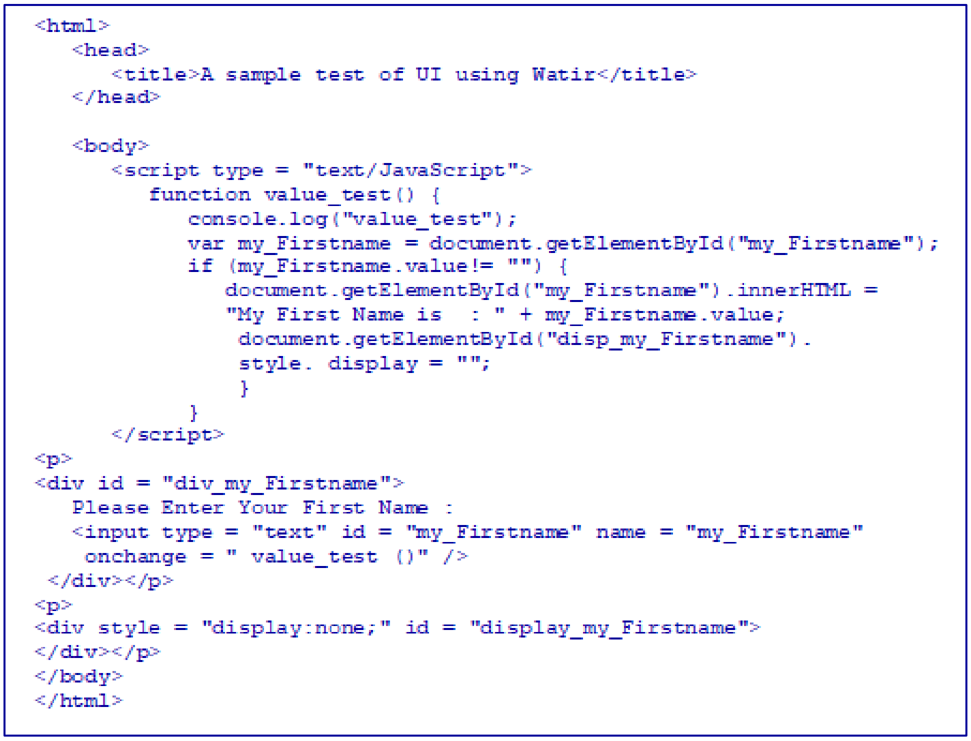
Fig. 9. Sample Test Script.



```
require 'watir'
require 'selenium-webdriver'
Selenium :: WebDriver:: Chrome.driver_path="C:/Users/m.aniljob/Desktop/_________SW/chromedriver_win32"
b = Watir::Browser.new :chrome
b.goto('http://localhost/ui_testing/my_test_file.html')
t = b.text_field id: 'my_Firstname'
t.exists?
t.set 'Minimol Job'
t.value
t.fire_event('onchange')
```

```
C:\Ruby30-x64\bin\ruby.exe C:/Ruby_Programs/UI-Test
```

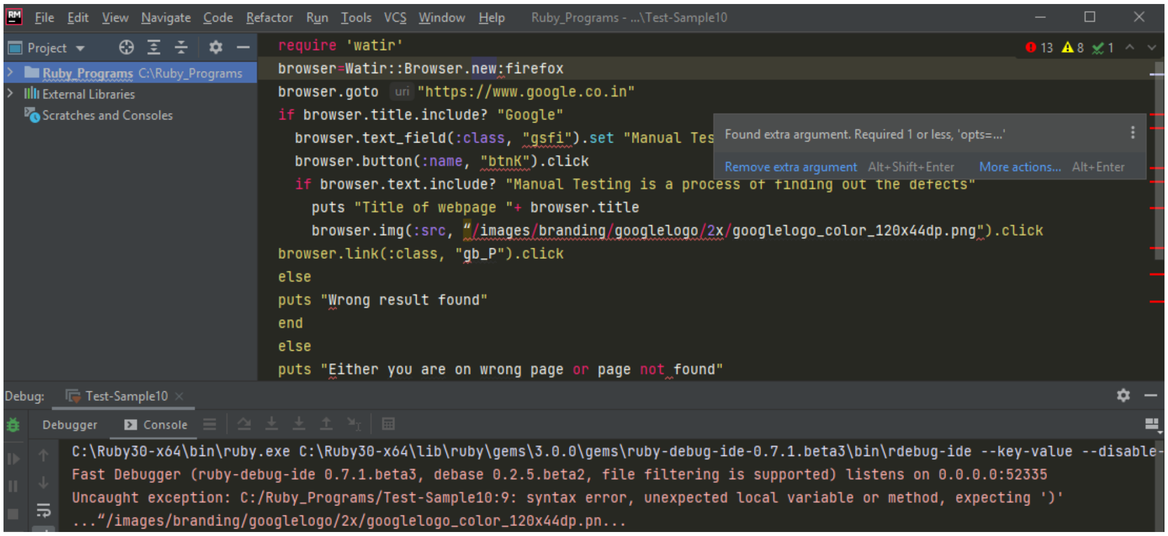
Fig. 10. Test Script for user Interface Testing.



```
<html>
   <head>
      <title>A sample test of UI using Watir</title>
   </head>

   <body>
      <script type = "text/JavaScript">
         function value_test() {
            console.log("value_test");
            var my_Firstname = document.getElementById("my_Firstname");
            if (my_Firstname.value!= "") {
               document.getElementById("my_Firstname").innerHTML =
               "My First Name is  : " + my_Firstname.value;
                document.getElementById("disp_my_Firstname").
                style. display = "";
                }
            }
      </script>
<p>
<div id = "div_my_Firstname">
   Please Enter Your First Name :
   <input type = "text" id = "my_Firstname" name = "my_Firstname"
    onchange = " value_test ()" />
 </div></p>
<p>
<div style = "display:none;" id = "display_my_Firstname">
</div></p>
</body>
</html>
```

Fig. 11. HTML and JavaScript Code for the Sample UI.


File Edit View Navigate Code Refactor Run Tools VCS Window Help Ruby_Programs - ...\Test-Sample10
Project
Ruby_Programs C:\Ruby_Programs
External Libraries
Scratches and Consoles
13 8 1
Remove extra argument Alt+Shift+Enter
More actions... Alt+Enter
Debug: Test-Sample10
Debugger
Console

```
require 'watir'
browser=Watir::Browser.new:firefox
browser.goto "https://www.google.co.in"
if browser.title.include? "Google"
  browser.text_field(:class, "gsfi").set "Manual Tes
  browser.button(:name, "btnK").click
  if browser.text.include? "Manual Testing is a process of finding out the defects"
    puts "Title of webpage "+ browser.title
    browser.img(:src, "/images/branding/googlelogo/2x/googlelogo_color_120x44dp.png").click
browser.link(:class, "gb_P").click
else
puts "Wrong result found"
end
else
puts "Either you are on wrong page or page not found"
```

Found extra argument. Required 1 or less, 'opts=...'

```
C:\Ruby30-x64\bin\ruby.exe C:\Ruby30-x64\lib\ruby\gems\3.0.0\gems\ruby-debug-ide-0.7.1.beta3\bin\rdebug-ide --key-value --disable-
Fast Debugger (ruby-debug-ide 0.7.1.beta3, debase 0.2.5.beta2, file filtering is supported) listens on 0.0.0.0:52335
Uncaught exception: C:/Ruby_Programs/Test-Sample10:9: syntax error, unexpected local variable or method, expecting ')'
...“/images/branding/googlelogo/2x/googlelogo_color_120x44dp.pn...
```

Fig. 12. Debugging Screen.

After running the test cases using it can be concluded that Watir allows easy to test file download for the user interface of a web application. Popup test alerts are provided at times as well. Another feature in Watir is, Page Performance can be measured using the performance object. Navigation, timing and memory performance can be measured by getting these details when the application is connected to the browser. Page object feature available in Watir helps reusability of the code basically in the form of classes. This feature of Watir helps the automation of the application without code redundancy. Without opening the browser, the details are obtained in the command line and thus supports the execution of the user interface test cases at the command line prompt.

VIII. Advantages of using AI in Software Testing

Fig. 13 lists the advantages of using AI tools in automating software testing. When the testing of a software system is done manually by even by highly skilled and experienced software testers there is a possibility of making mistakes and experience tiredness. Artificial Intelligent tools are constantly performing tasks effectively even repetitive tasks. Automated testing tools cab ne used by both software developers and testers. If the modifications ad editing of the program source code are cheeked thoroughly software tests will run automatically. For any unsuccessful tests, the developers will be notified and thus saving the time of the developers. AI enables automatic testing can help in improving the overall test coverage and ensure quality of the developed software. The performance of the system can be tested against the expected requirements in an efficient manner. AI enables testing can perform image and pattern recognition. This feature will help in detecting visual errors and make sure all visual elements are working properly. Software testing is a repeating process and source code may change time to time, manual testing will be a time consuming task. AI tools can perform the testing and detect errors with lesser time and AI tools are not expected to make errors and will help to generate more accurate results. Test automation can be done efficiently AI enables testing tools.

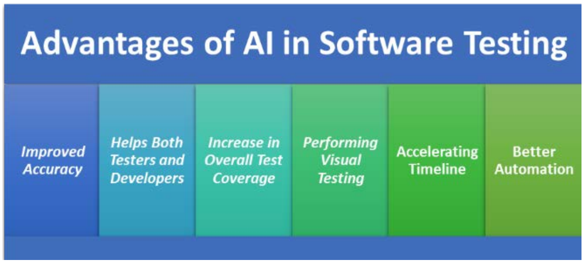

Advantages of AI in Software Testing
Improved Accuracy
Helps Both Testers and Developers
Increase in Overall Test Coverage
Performing Visual Testing
Accelerating Timeline
Better Automation

Fig. 13. AI Enabled Testing: Advantages.

IX. Conclusion and Future Work

Producing a quality software to the clients within the specified time by incorporating all the requirements is a crucial task. By using annual testing approach this will not be easy. There are various automation tools available to perform software testing. Artificial Intelligence techniques have significant impact in various stages of software development activity including software testing. Application of AI in test automation is an appropriate solution for software testing activity to produce a defect free software application. This paper presented the role of artificial Intelligence tools in software testing.

The paper also looked into various types of testing techniques for validation and verification of the software system. Selection of an appropriate AI based test automation tools is important based on the type of testing and this paper discusses the features of popular AI enables testing tools. Finally, the paper is presented the advantages of AI testing tools applications in software testing. An extension of this research paper will be presented by evaluating AI based test automation tools by taking into consideration of more technical details. Enhanced practical aspect will be covered by real implementation of the test automation of white-box testing using a mobile application.

References

[1] Agrafiotis, I., Creese, S., Goldsmith, M.: Developing a Strategy for Automated Privacy Testing Suites. In: Camenisch J., Crispo B., Fischer-Hübner S., Leenes R., Russello G. (eds.) Privacy and Identity Management for Life. Privacy and Identity 2011. IFIP Advances in Information and Communication Technology, vol. 375, pp. 32–44. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-31668-5_32.

[2] Andreas Leitner, Ilinca Ciupa, Bertrand Meyer, Mark Howard “Reconciling Manual and Automated Testing: the AutoTest Experience”. ETH Zurich CH-8092 Zürich, AXA Rosenberg Investment Management LLC Orinda, California 94563.

[3] A. M. Turing. “Computing machinery and intelligence”. In: Parsing the Turing Test. Springer, 2009, pp. 23–65. [2] J. McCarthy. “Artificial intelligence, logic and formalizing common sense”. In: Philosophical logic and artificial intelligence. Springer, 1989, pp. 161–190.

[4] Berndt, D.J., Fisher, J., Johnson, L., Pinglikar, J., and Watkins, A., “Breeding Software Test Cases with Genetic Algorithms,” In Proceedings of the Thirty-Sixth Hawaii International Conference on System Sciences (HICSS-36), Hawaii, January 2003.

[5] B. Littlewood and J. L. Verrall, “A Bayesian reliability growth model for computer software,” Applied Statistics, vol. 22, no. 3, pp. 332–346, 1973.

[6] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse, “The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users,” in Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann, Jul. 1998, pp. 256–265.

[7] Er. Rajender Bathla, Er. Shallu Bathla. “Innovative Approaches of Automated tools in Software testing & current technology as compared to Manual testing ”. Global Journal of Enterprise Information System, Vol-1(1) , 2009.

[8] Glenford J. Myers, Corey Sandler, Tom Badgett , The Art of Software Testing, 3rd Edition, 2015.

[9] Harman, M. (2012, June). The role of artificial intelligence in software engineering. In 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE) (pp. 1- 6). IEEE.

[10] Jagdish Singh, Monika Sharma.” A Comprehensive Review of Web-based Automation Testing Tools”. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3(10), October 2015.

[11] J. McCarthy. Artificial intelligence: a paper symposium: Professor Sir James Lighthill, FRS. Artificial Intelligence: A General Survey. In: Science Research Council, 1973. 1974.

[12] Khan, M. E., & Khan, F. (2012). A comparative study of white box, black box and grey box testing techniques. Int. J. Adv. Comput. Sci. Appl, 3(6).

[13] Lee Copeland, A Practitioner’s Guide to Software Test Design, 2003.

[14] M. Buenen and A. Walgude, “World quality report 2018–19,” Paris, France, Tech. Rep., 2018.

[15] Mariani, L., Hao, D., Subramanyan, R., Zhu, H.: The central role of test automation in software quality assurance. Software Quality Journal 25(3), 797–802 (2017). doi:10.1007/s11219-017-9383-5.

[16] Meziane, F. and Vadera, S., (2010). Artificial Intelligence in Software Engineering Current Developments and Future Prospects, In "Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects", IGI Global.

[17] Narayan, Vaibhav, The Role of AI in Software Engineering and Testing (June 22, 2018). International Journal of Technical Research and Applications, 2018, Available at SSRN: https://ssrn.com/abstract=3633525.

[18] N. E. Fenton, M. Neil, W. Marsh, P. Hearty, L. Radlinski, and P. Krause, “On the effectiveness of early life cycle defect prediction with Bayesian Nets,” Empirical Software Engineering, vol. 13, no. 5, pp. 499–537, 2008.

[19] Prof. (Dr.) V. N. MAURYA , Er. RAJENDER KUMAR . “Analytical Study on Manual vs. Automated Testing Using with Simplistic Cost Model ”. International Journal of Electronics and Electrical Engineering, vol .2 (1), 2012.

[20] R. M. Sharma.” Quantitative Analysis of Automation and Manual Testing.” International Journal of Engineering and Innovative Technology (IJEIT), Vol. 4(1), 2014.

[21] S. Russell et al. Artificial Intelligence: A Modern Approach. 3rd ed. Prentice Hall, 2010.

[22] S. H. Trivedi, "Software Testing Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 10, pp. 433439, 2012.

[23] Stuart Feldman, Quality Assurance: Much More than Testing, ACM Digital Library, Queue – Quality Assurance, February 2005, Vol 3, Issue 1.

[24] Tadapaneni, N. R. (2017). Different Types of Cloud Service Models. Available at SSRN 3614630.

[25] T. Kosa, M. Mernikb and T. Kosarb, "Test automation of a measurement system using a domain-specific modelling language," Journal of Systems and Software, vol. 111, p. 74–88, 2016.

[26] Van De Ven, T. et al., 2018. World Quality Report 2018-19 - Artificial Intelligence - World Quality Report 2018-19 Findings:. [Online] Available at: https://www.sogeti.com/globalassets/global/wqr-201819/wqr-201819_secured.pdf.

[27] V. Sangave and V. Nandedkar, "A Review on Automating Test Automation," International Journal of Advance Research in Computer Science and Management Studies, vol. 2, no. 12, 2014.

Cultural Events Classification using Hyper-parameter Optimization of Deep Learning Technique

Feng Zhipeng[1]
School of Culture Creativity
Hangzhou Normal University
Hangzhou, Zhejiang, P.R. China

Hamdan Gani[2]
Graduate School of Computer Systems
STMIK Handayani
Makassar, Indonesia

***Abstract*—Through digitization, maintaining and promoting cultural heritage is being strengthened. Concerning this background, this study presents a new Indonesia cultural events dataset and automatic image classification for cultural events. The dataset was developed using the Flickr image platform, and the five cultural events image was collected including the Baliem Festival, Jember Fashion Festival, Nyepi Festival, Pacu Jawi, and Pasola Festival. Further, Convolutional Neural Networks (CNN) was developed for the classification method. A comparison of CNN models (VGG16 and VGG19) using several optimization configurations was performed to get the best model. The results showed that the VGG16 with image augmentation and dropout regularization technique performed best with 94.66% accuracy. This study hoped to support the heritage's digital documentation process and preserve Indonesia's cultural heritage.**

Keywords—Cultural events; convolutional neural network (CNN); very depth convolutional network (VGG); multi-class classification

I. Introduction

Cultural events are created based on social systems or cultural wisdom passed from one generation to another [1], [2]. They have historical roots, customs, values, and beliefs influenced by many aspects such as region, social, and culture [3]. Hence, each specific ethnic group can be recognized based on their traditional cultural events. Understanding cultural values benefits maintaining cultural heritage [4]. As stated by The United Nations Educational, Scientific, and Cultural Organization (UNESCO) mission, every country is encouraged to approve the World Heritage Convention and ensure the identification, protection, and preservation of its cultural heritage [5]. Therefore, it is essential to sustain the cultural heritage in the face of rapid globalization in line with UNESCO's mission.

With the rapid development of technologies, the effort to preserve cultural heritage is being supported. By implementing digital documentation, cultural heritage can be quickly promoted and maintained. Several benefits of digital cultural heritage can make possible (a) transformation of heritage objects into the digital form [6], (b) quickly access to digital heritage [7], (c) indexation of historical heritage contents and extraction of their information [8], and (d) permanent preservation of digital objects [9]. Concerning all of those benefits, classification methods play an essential part. Classification refers to developing the classification model that will recognize the instances into categories or classes based on the training data (named supervised learning). The classification model learns from a training dataset and implements the achieved knowledge to classify new data. Therefore, documentation and classification of cultural heritage are essential since each country must save and preserve its cultural heritage.

In Indonesia's context, several issues have been discussed concerning Indonesian cultural heritage, such as foreign country claims to Indonesian regarding the cultural heritage, lack of the inter-generational transfer of knowledge in education, and lack of recognition from the local government [10]. In recent times, several efforts have been implemented to preserve and promote Indonesia's cultural heritage. For example, the Indonesian government strengthened the cultural heritage curricula in education, especially for the young generation. The Indonesian government also promoted the tagline "visit Indonesia" that has the goal to spread Indonesia's cultural events across the globe and targeted to attract visitors to Indonesia [11]. However, the preservation and promotion of cultural heritage are challenging. Indonesia is the world's largest archipelago nation, and it has one of the most varied cultural heritages with more than 300 distinct ethnic groups. Each ethnic group in Indonesia has cultural identities. Because of Indonesian culture's richness, the number of recognizable cultural events is also quite large. Thus, it needs documentation efforts to save and maintain the original cultural heritage of Indonesia.

To the best of our knowledge, no Indonesian cultural events documentation or dataset is available that describes a specific region's cultural events. One study investigated Indonesia's cultural heritage [12]. However, that study did not present Indonesia's cultural events rather than architectural heritage. Consequently, no specific Indonesian cultural events database is publicly available. Also, The Ministry of Tourism and Creative Economic of the Republic of Indonesia (Kemenparekraf RI) struggled to promote cultural events using the website https://www.indonesia.travel/, which primarily focuses on promoting various destinations in Indonesia for domestic and international tourism. However, that website does not promote cultural events.

Therefore, to support cultural heritage preservation, this study aims to present a new Indonesia's cultural events dataset and automatic image recognition for classification cultural events. Several CNN models for multi-class image classification were tested to achieve better accuracy. This paper has several contributions, specifically: (i) this study presents a new dataset of Indonesia's cultural events, and it has been made openly available to replicate this work (see link on the section availability of data and materials). The dataset would also be advantageous for researchers to consider adding a new image class to achieve the large dataset. (ii) The methodology of CNN with different hyper-parameter techniques is presented, and the results of the practical comparison are shown. Those results can be used as a benchmark for future researchers to improve the multi-class classification algorithm. In general, the proposed dataset and automatic classification system hoped can enhance an essential part of the heritage's digital documentation process and support an effort to preserve the cultural heritage.

This paper is organized as follows. Section 2 explains the materials and methods used in this study. Section 3 describes this study's results, followed by the discussion in Section 4. Finally, Section 5 explains the study's conclusions and future work.

II. Related Works

Several studies have developed cultural heritage documentation and implemented different methods to classify cultural heritage [13]. For example, in the study of architectural heritage, the authors proposed the image dataset of more than 10.000 images classified into ten classes, i.e., different architectural heritage types such as columns, domes, gargoyles, and vault [14]. This study compared the deep learning algorithms to categorize cultural heritage images. Specifically, several convolutional neural networks (CNN) were implemented, AlexNet, Inception V3, ResNet, and Inception-ResNet-v2. They achieved good accuracy on the complete training data; ResNet obtained a higher accuracy. In the fine-tuning configuration, the best accuracy was achieved for the Inception-ResNet-v2.

An early study [15] was investigated on a dataset containing 1.227 images dataset of 12 cultural heritage memorials and Pisa landmarks. The image classification was compared by using the k-nearest neighbor (k-NN) classification with different types of the feature extraction, namely Scale Invariant Feature Transform (SIFT), Speed up Robust Feature (SURF), Oriented FAST and Rotated BRIEF (ORB), and Binary Robust Invariant Scalable Keypoints (BRISK). They obtained that the local feature-based classifier achieved good accuracy; on the other hand, the best performance was reached using SIFT concerning the features.

Another study proposed 100 cultural heritage of wall painting images of reflected light for the image classification task [14]. The image dataset was involved in the reflected image, such as visible light, ultraviolet light, infrared light, and visible fluorescence. The authors used Dense SURF, spectral information, and a support vector machine algorithm. They concluded that the higher accuracy was the image in reflected ultraviolet light. Simultaneously, the dense integrating SURF and spectral information obtained the best accuracy than executing them individually.

The previous study [14] proposed an Indonesian cultural heritage dataset for image, audio, and video classification. The dataset includes 100 images, 100 audios, 100 videos, and 100 text files separated into five classes. The deep learning algorithms, Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN), were executed to classify Indonesian architectural heritage. The CNN was executed for image, audio, and video classification, while RNN was executed to classify text. The results revealed that RNN obtained the best performance regarding the accuracy, classifying 92% of the text data. Concerning CNN, the higher accuracy (76% each) reached for image and video classification, and audio acquired 57% accuracy.

In the study of archaeological sites [16], the author collected the cultural heritage dataset that included 150 images and categorized them into three classes (50 images each category): archaeological sites, frescoes, and monasteries. This study aimed to classify images using several decision tree algorithms such as J48, Hoeffding tree, random tree, and random forest. The authors determined that the random forest algorithm achieved the best performance.

Although several studies have shown advances in researching the cultural heritage for image classification, mainly those research only focused on the architectural building classification as explained in the above. A study involving cultural events or ceremonies was not still profoundly studied. Thus, studying cultural events are necessary to support the way to protect and promote cultural heritage for future generations.

III. Material and Methods

A. Materials

The scrapping image method was performed to collect Indonesia's cultural events using Python programming. The Flickr image service was used for collecting the image datasets. The Flickr images under Attribution-Noncommercial License were collected. Five cultural events were collected: Baliem Valley Cultural Festival, Jember Fashion Carnival, Nyepi the Day of Silence, Pacu Jawi, and Pasola Festival. The dataset consisted of 1.500 images and was divided into 300 images in each class. The image samples from the five classes are shown in Fig. 1. The images represent (a) Pasola, (b) Pacu Jawi, (c) Nyepi, (d) Jember Fashion, and (e) Baliem.

B. Methods

The experiments were performed in Python v.3.7 environments, and the CNN (VGG) model was developed with the Keras library. The experiments were performed under Windows 10 platform, 16 GB Graphical Processing Unit (GPU), 256 GB SSD storage, Core i7 processor 1.80 GHz, and 8 GB of RAM. All images were converted into 200 x 300 image pixels. The distribution of image classes is shown in Fig. 2.

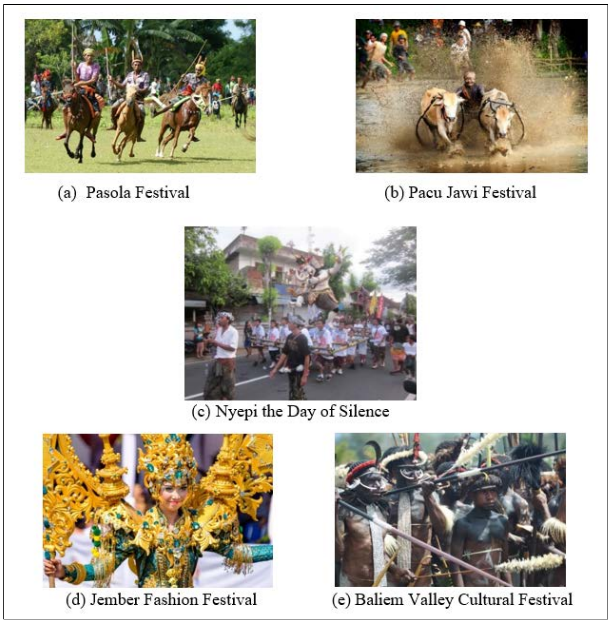

(a) Pasola Festival
(b) Pacu Jawi Festival
(c) Nyepi the Day of Silence
(d) Jember Fashion Festival
(e) Baliem Valley Cultural Festival

Fig. 1. Examples of Five Image Classes: (a) Pasola Festival, (b) Pacu Jawi Festival, (c) Nyepi the Day of Silence, (d) Jember Fashion Festival, and (e) Baliem Valley Cultural Festival.

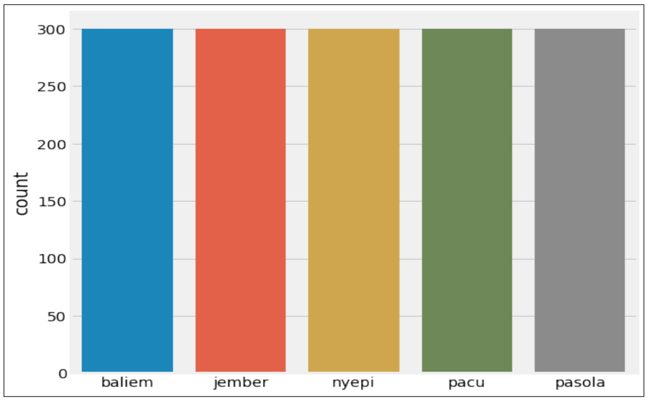


Fig. 2. Distribution of different Classes of the Dataset.

1) Convolutional Neural Network (CNN): CNN is usually applied for computer vision, as it captures images as inputs and extracts features from the images. The CNN typically contains convolutional layers (each involving several kernel sizes and filters). The convolutional layer is along with the pooling layer, decreasing data dimensionality. There are two types of pooling: max-pooling and average pooling. Max-pooling uses the maximum value from the image related to the kernel size, while average pooling utilizes the average of all the values. Once image processing is accomplished across these layers, the features from a two-dimensional matrix are converted into a vector with a flatten layer, and the achieved output is transmitted to the fully connected layer or dense layer [17].

2) Very Depth Convolutional Network (VGG): VGG's name belonged to their lab's name, the Visual Geometry Group at Oxford, and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 competition winner [18]. This architecture was designed for deep convolutional network learning. The VGG architecture is created by 3 x 3 Convolutional and MaxPooling layers, with a fully connected block at the end. In the original paper, VGG architecture has shown the depth network's effect on its performance in the large-scale image database. VGG used small (3 x 3) convolution filters in the complete architecture and presented a considerable improvement in the configurations using the depth to 16-19 weight layers [18]. Another advantage of VGG architecture is that they used many filters. The number of filters grows with the depth of the model. They start at 64 and continually increase across 128, 256, and 512 filters at the end of the model's feature extraction. VGG was named for the number of layers: the VGG16 for 16 layers and the VGG19 for 19 learned layers. The architecture for VGG16 and VGG19 is presented in Fig. 3(a) and (b). The summary of the architecture of VGG models as follows: (i) apply small convolutional filters, e.g., 3 x 3 and 1 x 1, (ii) apply max pooling with a size of 2 x 2, (iii) the stacking of convolutional layers concurrently before applying a pooling layer to identify a block, (iv) dramatic reiteration of the convolutional-pooling block pattern (v) development of intense models (16 and 19 layers). The architecture of VGG16 and VGG19 are depicted in Fig. 3. In this work, the VGG 16 and VGG19 models were used for the experiment.

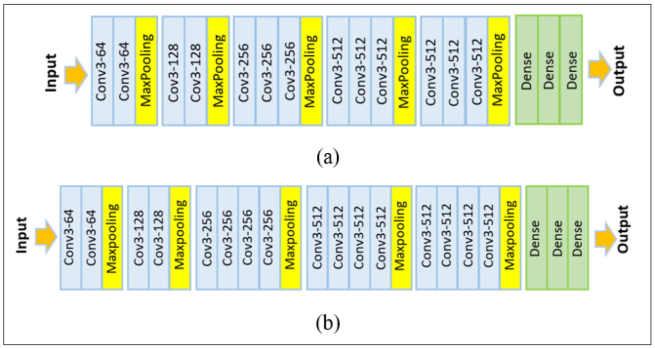


Fig. 3. The Architecture of VGG16 (a) and VGG19 (b).

3) Image augmentation optimization: The image augmentation works to create a new and unique training example. Image augmentation transforms the images' versions in the training dataset corresponding to the same class as the initial image [19]. That transformation involves several image manipulation processes, such as shifts, flips, and zooms. The current deep learning algorithms, such as CNN, can quickly learn the image features. The augmentation technique can improve the algorithm's learning process, and it is usually implemented to the training dataset, not to the test or validation dataset. In this work, the Keras deep learning library was used. That library offers several image augmentations functions via the ImageDataGenerator class. Several augmentation functions were used, such as the horizontal and vertical flip of image (randomly flip training image), rotation (randomly rotate images in the range of degrees), brightness, and zoom (randomly zoom image).

4) Dropout regularization: Dropout is an effective technique to maintain the neural network from overfitting during the training [20]. Dropout is applied by only saving a neuron active with a certain probability *p* and locating it to 0 otherwise. This condition pushes the network not to learn redundant information [20]. Consequently, this method significantly decreases overfitting and presents significant neural network improvements in supervised learning [21]. In this study, the Keras deep learning library was used for importing the dropout regularization class in the VGG model. The drop out was set on 0.5.

5) Transfer learning: In machine learning, transfer learning refers typically to a method where a model trained on a specific problem is implemented in other problems, which is a related problem [22]. Transfer learning has the advantages of reducing the training time for an algorithm model and can produce lower generalization errors. The weights in re-used or latest layers can be used as the initial point for the training process and implemented to answer the new problem. Transfer learning can be helpful when the first associated problem has many labeled data. Several high-performing models have been created for image classification on the annual ILSVRC, such as ZFNet, VGG, GoogleNet, and ResNet [17]. This competition has produced several innovative models in CNN architecture and can be implemented to transfer learning in computer vision applications. Those models have learned over 1.000.000 images for 1.000 classes and achieved state-of-the-art performance. In this study, all VGG model was developed using transfer learning and directly downloaded using the Keras library function into our python environment.

6) Configuration of VGG model: The VGG architecture included 19 weight layers, 16 convolutional layers, and 3 fully connected layers [18]. The channel number of convolutional layers starts from 64 and increases by a factor of 2 after each max-pooling layer until obtaining 512. Finally, SoftMax activation was used in the dense layer. The architecture of VGG16 is similar to VGG19; only the difference is the total number of layers (16 for VGG16). A Method for Stochastic Optimization (Adam) was implemented as an optimizer. In order to avoid overfitting, the early stopping function was implemented with configuration patience 5 and verbose 1. The number of epochs was adjusted to 30. The VGG model used 80% for training data and 20% for validation. The 150 x 150 image pixels were used as an input image for VGG models. The detailed configuration of all VGG models is presented in Table I.

7) Model evaluation: After classification was performed, several evaluation metrics were performed. Specifically, the accuracy, precision, recall, F1 score, and ROC area were used to choose the best model. The accuracy is the percentage of correct instances classified by the algorithm. Precision is the number of instances that fit the respected class and the calculated instances categorized to that class, while recall or sensitivity explains the true positive rate of prediction. The F1 score or F-measure explains the classification accuracy regarding the average precision and recall values. The F1 score values closer to 1 show a better classification accuracy. The measurement methods are calculated as follows in Eq. (1), (2), (3), and (4):

$$Accuracy = \frac{Correctly\ classified\ data}{Total\ data} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Precission = \frac{TP}{TP+FP} \tag{3}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision+ Recall} \tag{4}$$

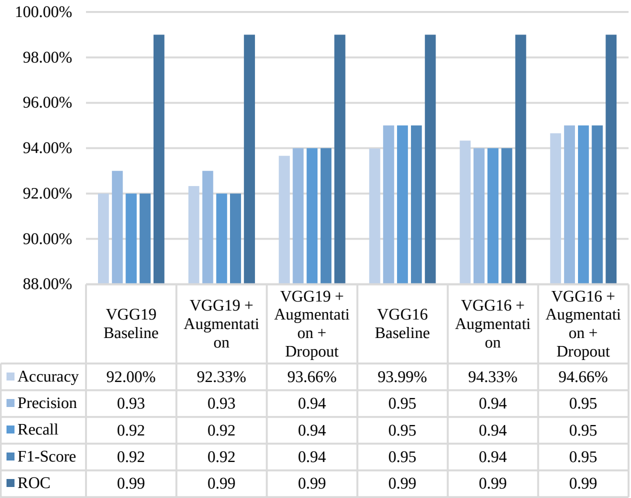
Where TP is a true positive, FN is a false negative, and FP is a false positive. Finally, the region under the ROC curve shows the proportion of true positives and false positives. This value must be close to 1, indicating a perfect prediction, as the values under 0.5 imply a random guess [23].

TABLE I. CONFIGURATION OF VGG ARCHITECTURE MODELS

VGG16 + Augmentation + Drop Out	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	Yes	Yes (0.5)	Yes (Patience 5, verbose 1)	Yes
VGG16 + Augmentation	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	Yes	Yes	Yes (Patience 5, verbose 1)	Yes
VGG16 Baseline	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	No	No	Yes (Patience 5, verbose 1)	Yes
VGG19 + Augmentation + Drop Out	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	Yes	Yes (0.5)	Yes (Patience 5, verbose 1)	Yes
VGG19 + Augmentation	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	Yes	Yes	Yes (Patience 5, verbose 1)	Yes
VGG19 Baseline	Softmax	30	ImageNet	32	Categorical Cross entropy	Adam	No	No	Yes (Patience 5, verbose 1)	Yes
Hyper-Parameter	Activation Function	Number of Epoch	Weight	Batch Size	Loss Function	Optimizer	Image Augme ntation	Drop Out	Early Stopping	Transfer Learning

IV. Experimental Results

This section explains the analysis of the obtained algorithm's performance on the cultural events dataset. Several algorithms with hyper-parameter were evaluated using the accuracy, precision, recall, F1 score, and ROC area. The performance of the algorithm can be seen in Fig. 4.



	VGG19 Baseline	VGG19 + Augmentation	VGG19 + Augmentation + Dropout	VGG16 Baseline	VGG16 + Augmentation	VGG16 + Augmentation + Dropout
Accuracy	92.00%	92.33%	93.66%	93.99%	94.33%	94.66%
Precision	0.93	0.93	0.94	0.95	0.94	0.95
Recall	0.92	0.92	0.94	0.95	0.94	0.95
F1-Score	0.92	0.92	0.94	0.95	0.94	0.95
ROC	0.99	0.99	0.99	0.99	0.99	0.99

Fig. 4. The Performance of each Algorithm

Based on the results, it showed that VGG16 performed better than VGG19. The combination of "VGG16 + Augmentation + Dropout" performed the best with 94.66% of correctly classified images, followed by "VGG16 + Augmentation" with 94.33%, and "VGG16 Baseline" with 93.99%. On the other hand, the combination of "VGG19 + Augmentation + Dropout" performed with 93.66% of correctly classified images, followed by "VGG19 + Augmentation" with 92.33%, and "VGG19 Baseline" with 92.00%. It also showed that the precision, recall, F1-score, and ROC area were better for the "VGG16 + Augmentation + Dropout" than the other VGG configurations.

After implementing the hyper-parameter optimization, the algorithms performed better than the baseline. The model performance showed a very slight increase in the model's mean accuracy, 92.33% in "VGG19 + Augmentation" compared to 92.00% with the VGG19 baseline model. It also confirmed that dropout regularization performed well. There was a very slight rise in the model's accuracy, 93.66% in "VGG19 + Augmentation + Dropout" compared to 92.33% with the "VGG19 + Augmentation".

A similar improvement is presented in the VGG16 model. The estimated performance of the "VGG + Augmentation + Dropout" model indicated a possible increase in performance compared to the baseline from 93.99% to 94.66%. This study's findings confirm that a hyper-parameter configuration (image augmentation and dropout regularization) can improve the models in line with previous works [24].

In order to detect which of the classes classified correctly, the confusion matrices were performed. The results of the confusion matrices are presented in Table II. The diagonal numbers show the correctly classified images (blue background), while other numbers in rows describe the misclassifications of images. It showed that the "VGG19 Baseline" most accurately classified the Pacu Jawi and Pasola images, while the "VGG19 + Augmentation" and "VGG19 + Augmentation + Dropout" most correctly classified the Jember and Pacu Jawi images. The "VGG16 Baseline" and "VGG16 + Augmentation" accurately classified the Pacu Jawi, Pasola, and Jember images, while the "VGG16 + Augmentation + Dropout" most correctly classified the Pacu Jawi images.

TABLE II. The Confusion Matrices for each Algorithm

Algorithm	**Baliem**	**Jember**	**Nyepi**	**Pacu Jawi**	**Pasola**	**Classified as**
VGG19 Baseline	52	0	3	0	1	Baliem
	4	56	5	0	0	Jember
	3	0	44	0	2	Nyepi
	0	0	0	65	1	Pacu Jawi
	1	0	3	1	59	Pasola
VGG19 + Augmentation	48	0	5	2	1	Baliem
	3	58	4	0	0	Jember
	2	0	47	0	0	Nyepi
	0	0	0	66	0	Pacu Jawi
	1	0	3	2	58	Pasola
VGG19 + Augmentation + Dropout	52	0	3	1	0	Baliem
	3	59	3	0	0	Jember
	3	0	46	0	0	Nyepi
	0	0	0	66	0	Pacu Jawi
	1	0	3	2	58	Pasola
VGG16 Baseline	48	1	4	2	1	Baliem
	1	62	2	0	0	Jember
	1	0	45	0	3	Nyepi
	1	0	0	65	0	Pacu Jawi
	0	1	1	0	62	Pasola
VGG16 + Augmentation	52	0	2	1	1	Baliem
	1	62	2	0	0	Jember
	4	0	43	0	2	Nyepi
	2	0	0	64	0	Pacu Jawi
	1	1	0	0	62	Pasola
VGG19 + Augmentation + Dropout	55	0	0	1	0	Baliem
	2	62	1	0	0	Jember
	5	0	41	0	3	Nyepi
	2	0	0	64	0	Pacu Jawi
	0	1	1	0	62	Pasola

In terms of misclassification, "VGG19 + Augmentation + Dropout" mostly misclassified the Nyepi images, while the rest of the other algorithms misclassified the Nyepi and Baliem images. Based on the results, Pacu Jawi images were most correctly classified among all algorithms, while Nyepi images were most misclassified. Finally, to fully show these algorithms' performance, the accuracy and loss model presented in Fig. 5. As for simplicity only "VGG16 + Augmentation + Dropout" and "VGG19 + Augmentation + Dropout" models are presented.

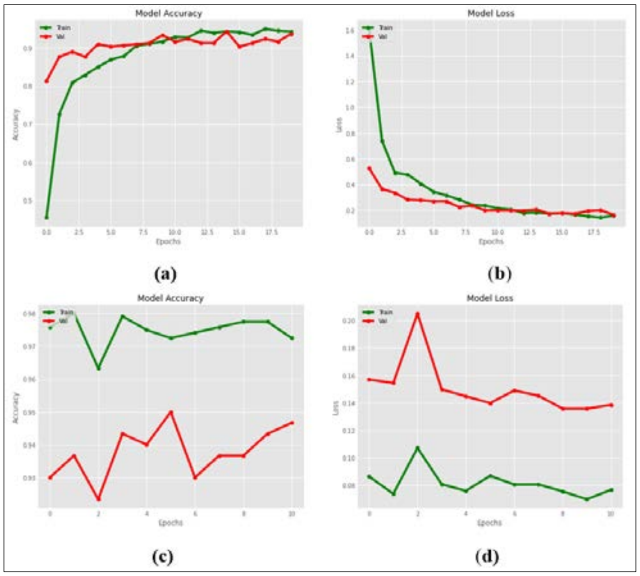


Fig. 5. (a) Accuracy for VGG19, (b) Loss Results for VGG19, (c) Accuracy for VGG16, and (d) Loss Results for VGG16.

As observed in Fig. 5, the VGG19 model performed well, with an accuracy of 93.66 % (a). Both training and validation loss had a reducing trend achieving the value of 0.20 for training and validation in the last epoch (b). On the other hand, VGG16 with 94.66% accuracy seemed to perform well and tended to increase to fit (train and validation line) further after the last epoch (c). The model showed a decreasing trend from the value of 0.16 to 0.14 (d) concerning the validation loss. The model's epoch finished on 18 epochs for VGG19 and 10 epochs for VGG16, regarding no further improvement on accuracy and avoiding the overfitting.

In a survey study, image augmentation has been proved to improve the performance of the models and enhance the limitation of datasets to take advantage of significant data capabilities [19]. In agreement with our results from performance accuracy in Fig. 4, it displayed that model using the image augmentation technique (in both VGG19 and VGG16 models) was better than the baseline model. In their experimental study, Nandini et al. [25] compared the dropout technique for image classification using several algorithms using three different image datasets. They found that the dropout regularization technique accomplished the best results in image classification compared with other classification algorithms. This finding is also in line with our results in Fig. 4 that display the dropout technique's performance, and image augmentation performs best compared to other configuration models.

V. Discussion

This study presents a new Indonesia cultural events dataset and automatic image recognition for classification cultural events. Several findings obtained from this study: (i) the convolutional neural network, with VGG16 architecture, performed well when classifying images compared to the other models (despite that other algorithms accuracy were relatively good accuracy); (ii) all classifiers performed better after adding hyper-parameter configuration (image augmentation and dropout regularization); and (iii) algorithms most correctly classified the Jember and Pacu Jawi images, while Nyepi images most frequently misclassified.

As shown in the previous section, the "VGG16 + Augmentation + Dropout" performed well in all performance measures. It obtained the highest classification accuracy compared to the other algorithms. Also, it performed the best in the other evaluation measures, such as precision, recall, F1 score, and the ROC. Furthermore, the CNN with VGG architecture model showed excellent classification accuracy. Specifically, CNN mainly performs better than other non-neural network algorithms applied for image classification tasks [17], [25]–[27].

Moreover, CNN is suitable for a large dataset. Regarding the overfitting problems, hyper-parameter tuning has been implemented with an early stopping function to avoid the overfitting problem. The Keras library's early stopping function was used to stop training when the training accuracy gets a specific threshold. Hence, the optimal model weights can be achieved and save computation time and power. Although CNN is computationally intensive, it can achieve good performance using several hyper-parameter configurations. Thus, based on this study's results, hyper-parameter configurations are a promising way to improve the algorithm's performance, especially CNN with VGG architecture.

This study could develop cultural events image classification models that could be more efficient and computationally solid. As multi-class image classification usually includes many images and needs substantial computational resources, it must be operated correctly and reliable. Therefore, developing optimized models and reliable classification methods is essential for current and future studies.

As the limitation of this study, the proposed dataset used a balanced class distribution. Thus, our proposed VGG configuration needs to test in an imbalanced classification problem to show the validity of our proposed configuration models.

As future work, this study plans to enhance the proposed image datasets, including different cultural events images from different Indonesia regions. Also, generally, the image classification generally includes large data sets, further work needs to develop large datasets of Indonesia culture's image dataset.

VI. Conclusion

This study presents a new Indonesia cultural events dataset and automatic image recognition for classification cultural events. This study compared several configurations of hyper-parameter configurations of CNN for multi-class image classification. In particular, CNN architecture, such as VGG19 and VGG16, were tested before and after the hyper-parameter configuration. Overall, the VGG19 and VGG16 achieved a good performance, but considering the hyper-parameter optimization, the VGG16 using image augmentation and dropout regularization achieved the best classification with 94.66% accuracy. Other algorithms achieved less than 94.33% accuracy. Despite that, the accuracy of other algorithms was relatively good. This study confirms that CNN with VGG architecture is a better choice for multi-class image classification, and they offer good performance for classification tasks. Finally, this study's findings hoped to support the heritage's digital documentation process and maintain cultural heritage.

ACKNOWLEDGMENT

The authors acknowledge that this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AVAILABILITY OF DATA AND MATERIALS

https://www.dropbox.com/l/scl/AAAO7Es-r_8y3Xtv0ZFkFjh4hjqTl-1GViQ

REFERENCES

[1] J. M. Hernández-Mogollón, P. A. Duarte, and J. A. Folgado-Fernández, "The contribution of cultural events to the formation of the cognitive and affective images of a tourist destination," Journal of Destination Marketing & Management, vol. 8, pp. 170–178, 2018, doi: https://doi.org/10.1016/j.jdmm.2017.03.004.

[2] B. Djibat, S. Deni, and Z. Saing, "The culture of Makayaklo in North Maluku Society: Teaching the values of building solidarity and social integration," International Journal of Critical Cultural Studies, vol. 17, no. 1, pp. 43–54, 2019, doi: 10.18848/2327-0055/CGP/v17i01/43-54.

[3] M. R. Abdulla, "Culture, Religion, and Freedom of Religion or Belief," Review of Faith and International Affairs, vol. 16, no. 4, pp. 102–115, 2018, doi: 10.1080/15570274.2018.1535033.

[4] T. Lussetyowati, "Preservation and Conservation through Cultural Heritage Tourism. Case Study: Musi Riverside Palembang," Procedia - Social and Behavioral Sciences, vol. 184, no. August 2014, pp. 401–406, 2015, doi: 10.1016/j.sbspro.2015.05.109.

[5] P. Alasuutari and A. Kangas, "The global spread of the concept of cultural policy," Poetics, vol. 82, no. March, p. 101445, 2020, doi: 10.1016/j.poetic.2020.101445.

[6] N. A. Khan, S. M. Shafi, and H. Ahangar, "Digitization of Cultural Heritage," Journal of Cases on Information Technology, vol. 20, no. 4, pp. 1–16, Oct. 2018, doi: 10.4018/JCIT.2018100101.

[7] V. Borissova, "Cultural heritage digitization and related intellectual property issues," Journal of Cultural Heritage, vol. 34, pp. 145–150, 2018, doi: https://doi.org/10.1016/j.culher.2018.04.023.

[8] G. Colavizza, M. Ehrmann, and F. Bortoluzzi, "Index-Driven Digitization and Indexation of Historical Archives," Frontiers in Digital Humanities, vol. 6, p. 4, 2019, doi: 10.3389/fdigh.2019.00004.

[9] Z. Ognjanović, B. Marinković, M. Šegan-Radonjić, and D. Masliković, "Cultural heritage digitization in Serbia: Standards, policies, and case studies," Sustainability (Switzerland), vol. 11, no. 14, 2019, doi: 10.3390/su11143788.

[10] R. Asfina and R. Ovilia, "Be Proud of Indonesian Cultural Heritage Richness and Be Alert of Its Preservation Efforts in the Global World," Humanus, vol. 15, no. 2, p. 195, 2017, doi: 10.24036/jh.v15i2.6428.

[11] N. Amanah, I. Atjo, and M. Dewi, "Indonesia Tourism Communication Strategy in Event ITB Berlin 2016," The 4th Conference on Communication, Culture and Media Studies Indonesia, no. October, pp. 10–11, 2017.

[12] R. A. Kambau, Z. A. Hasibuan, and M. O. Pratama, "Classification for multiformat object of cultural heritage using deep learning," Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018, pp. 1–7, 2018, doi: 10.1109/IAC.2018.8780557.

[13] R. Janković, "Machine learning models for cultural heritage image classification: Comparison based on attribute selection," Information (Switzerland), vol. 11, no. 1, 2020, doi: 10.3390/info11010012.

[14] J. Llamas, P. M. Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo, "Classification of architectural heritage images using deep learning techniques," Applied Sciences (Switzerland), vol. 7, no. 10, pp. 1–26, 2017, doi: 10.3390/app7100992.

[15] G. Amato, F. Falchi, and C. Gennaro, "Fast image classification for monument recognition," Journal on Computing and Cultural Heritage, vol. 8, no. 4, 2015, doi: 10.1145/2724727.

[16] R. Janković, "Classifying cultural heritage images by using decision tree classifiers in WEKA," CEUR Workshop Proceedings, vol. 2320, pp. 119–127, 2019.

[17] J. Gu et al., "Recent advances in convolutional neural networks," Pattern Recognition, vol. 77, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–14, 2015.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[20] P. Baldi and P. Sadowski, "The dropout learning algorithm," Artificial Intelligence, vol. 210, no. 1, pp. 78–122, 2014, doi: 10.1016/j.artint.2014.02.004.

[21] G. S. Nandin, A. P. S. Kumar, and C. K, "Dropout Technique for Image Classification Based On Extreme Learning Machine," Global Transitions Proceedings, pp. 0–7, 2021, doi: 10.1016/j.gltp.2021.01.015.

[22] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, A survey of transfer learning, vol. 3, no. 1. Springer International Publishing, 2016.

[23] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[24] D. Zhao, G. Yu, P. Xu, and M. Luo, "Equivalence between dropout and data augmentation: A mathematical check," Neural Networks, vol. 115, pp. 82–89, 2019, doi: 10.1016/j.neunet.2019.03.013.

[25] M. P. Véstias, "A survey of convolutional neural networks on edge with reconfigurable computing," Algorithms, vol. 12, no. 8, 2019, doi: 10.3390/a12080154.

[26] R. D. Yogaswara and A. D. Wibawa, "Comparison of Supervised Learning Image Classification Algorithms for Food and Non-Food Objects," in 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM), Nov. 2018, pp. 317–324, doi: 10.1109/CENIM.2018.8711387.

[27] F. Siraj, M. A. Salahuddin, and S. A. M. Yusof, "Digital Image Classification for Malaysian Blooming Flower," in 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, Sep. 2010, pp. 33–38, doi: 10.1109/CIMSiM.2010.92.

Towards using Single EEG Channel for Human Identity Verification

Marwa A. Elshahed
Physics Department, Faculty of Women for Arts
Sciences and Education, Ain Shams University, Cairo, Egypt

***Abstract*—Biometrics is an interesting area of research as a result of tremendous technological advances, especially in security. It is considered as an automated technology used for identification based on biological or behavioral human traits. An electroencephalogram (EEG) is the brain electrical activity signals considered as biological traits used in biometrics systems. The primary goal of this work is trying to find a single EEG channel to be used for human identification purposes. A single EEG channel recording is used for personal identity-based verification mode, which is preferred for many subjects with instant real-time system decisions. Percent residual difference (PRD) is a common quantitative measurement used to determine the human identity-based measures the distance between two signals. The proposed system sensitivity gives 100% using some single channels placed in the parietal and occipital lobes. The proposed system takes a short time in the enrolment process with an instant decision using verification mode, which is preferred with a large number of subjects. Also, using imaginary tasks is preferred for human identity verification.**

Keywords—Biometric; EEG; single channel; verification; brain lobes

I. Introduction

Biometrics is the field of studying the biological or behavioral traits of humans which could be used for personal identification. Biological traits depend on direct measures of shapes or characteristics of biological parts in the human body, such as iris, face, DNA, ECG, etc… Behavioral traits depend on the human behavioral such as gait, signature, voice, etc.

Not all human traits could be used as a biometric, it should be characterized by uniqueness, universality, collectability, permanence, performance, circumvention, and acceptability. There are some traits changes over time, so any biometric system should be updated after some time.

A Biometric system is based on identification or verification modes. The enrolment process is used to build the system database. It was the first and main step for any system. The identification mode is considered as one to many comparisons which takes the extracted trait from the claimed person and compares it with all that stored in the database to get its identity as in Fig. 1, while in the verification mode, which is considered as one to one comparison, takes the extracted trait and ID from the claimed person then only one template from the system database is compared with the extracted one using the ID as in Fig. 2. The biometric system operates in one of the two modes. Some biometric systems are considered as a unimodal system, which depends on only one biometric trait, while others are considered as a multimodal biometric system which depends on more than one human trait.

Electroencephalogram (EEG) signals are considered as physiological biometrics which represent the brain's electrical activity. It has become an interesting area of research for human identification due to the amazing progress in using sensors and wireless networks based on Wireless Body Area Network (WBAN) [1-6]. EEG signal is classified into five waves according to the frequencies based on human activity and this is illustrated in brief in Table I, while other studies classify EEG signals to six waves, the following five waves in addition to Mu waves with an overlapping frequency range [7-11].

The human brain consists of four lobes illustrated in Fig. 3 by four different colors. The frontal lobe (blue) is located in front of the brain, followed by the parietal lobe (yellow) at the top of the head, then the occipital lobe (pink) at the back of the head, while the temporal lobes (green) are located on both sides of the brain above the ears. The five senses' locations are illustrated in brain lobes as in Fig. 3.

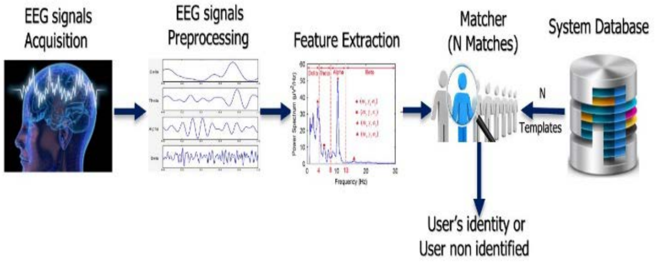


Fig. 1. Identification EEG Biometric System.

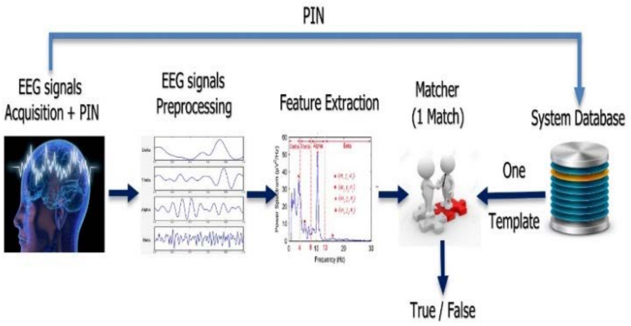


Fig. 2. Verification EEG Biometric System.

TABLE I. EEG SIGNAL CLASSIFICATION [7].

Name	Frequency band (Hz)	State
Δ WAVE	1 ~ 3	Deep sleep
Θ WAVE	4 ~ 7	Shallow sleep
Α WAVE	8 ~ 13	Relaxed state
Β WAVE	14 ~ 30	Active state
Γ WAVE	30 ~	Mental strain

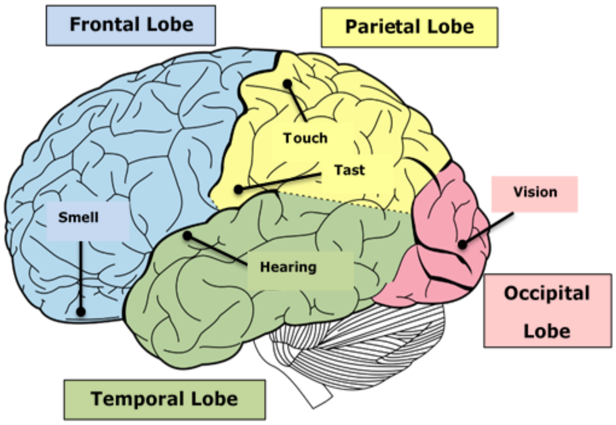


Fig. 3. The Human Brain Four Lobes.

The frontal lobe function is responsible for muscle movements, speaking, judgments, and making plans. It is described as the brain control center. Next, the parietal lobe processes information about taste and touch senses. The Occipital lobe is responsible for vision. The temporal lobe is the fourth lobe that is responsible for receiving sound from the opposite ear [12,13]. EEG signals are extracted from the human brain based on certain positions. There was an international system that illustrates these electrode or channel positions in the human scalp. This work used a database created based on the international 10-20 system as shown in Fig. 4.

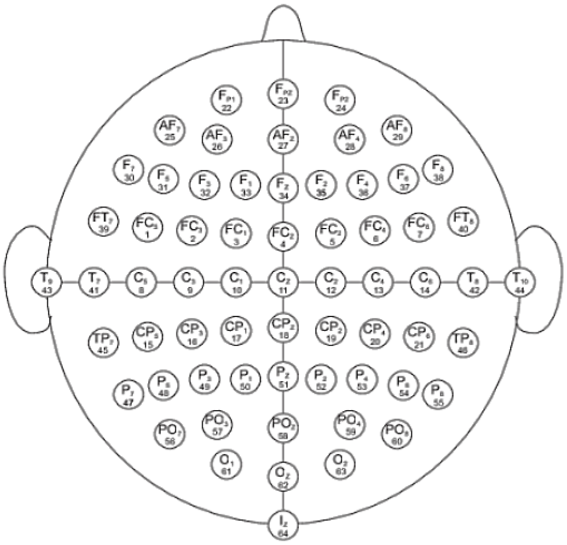


Fig. 4. 64 Electrodes Positioning based on the International 10-20 System [16].

The system depends on the relation between the cerebral cortex area and the electrode location. Also, the distances between the adjacent electrodes are equal to 10% or 20% from the whole right-left or front-back of the skill distance. The electrode name has a letter that refers to the lobe (F-frontal, P-parietal, O-O-occipital, T-temporal, C-for central positions, and Z-for midline positions) and a number that refers to the hemisphere location. The even numbers for right hemisphere positions while the odd numbers for left hemisphere positions [14, 15].

II. Related Work

Authors in [17] obtained a high system accuracy rate using proposed combined Independent Component Analysis (ICA) and AR classifiers. They used only 20 subjects based on sixteen electrode placements on the scalp according to the 10-20 international system. In [18], authors trying to get an EEG single channel to be used in the authentication. They used the same dataset which was used in this work based on open and closed eye states with 109 subjects, the obtained system accuracy in the range of 97-99%. Studying to get a single EEG channel for human identification was proposed in [19] also, which is based on generating a personal identification number (PIN) from the brain activity obtained from a single active EEG channel with 3 subjects only. The high system performance is obtained using channel Cz. New features are proposed in [20] Based on a single Fp1 single channel called the concavity and convexity features in the alpha band using 23 subjects. Authors in [21] propose a new technique for EEG human verification using three channels based on Discrete Fractional Fourier Transform (DFrFT) as a feature extraction method. Their method achieves a 0.22% Equal Error Rate (EER) with 104 subjects. In [22], an EEG biometric authentifier was proposed based on deep learning technique. Only 15 subjects were used with over 40 trials.

The experiment in [23] proposes a self-relative framework based on EEG signal for human identification using 108 subjects with closed eyes in resting state (the same dataset which was used in this work). Autoregressive is used as a feature extraction method while K-nearest neighbor is used as a classifier. They found that the openness condition gives more accurate results for identification. 19 selected channels only are used.

Most EEG biometric systems are based on identification mode. Also, they didn't test their systems for unauthorized subjects. The biometric system performance measures are neglected in calculations except the accuracy of most of them.

The objective of this research is trying to find only one EEG channel recording which could be used in a biometric system to get a personal identity with high efficiency and a short time with a large number of subjects-based verification mode.

III. Research Method

In this research, 109 subjects are used from the EEG Motor Movement/Imagery Dataset (eegmmidb) Database [16]. Table II summarizes the dataset description. The number of authorized subjects used in this work is 100 while 9 are used as unauthorized subjects. 64-channels were used for EEG signal recording based on the BCI2000 system [24].

Subjects performed six motor/imagery tasks. Four of them were repeated three runs, so each subject performed 14 runs. Task 1 for opening and closing left or right fist, Task 2 for imagining opening and closing left or right fist, Task 3 for opening and closing both fists or both feet, and Task 4 for imagining opening and closing both fists or both feet. Task 1 and task 3 are non-imaginary tasks, while task 2 and task 4 are imaginary tasks.

In this work, 12 runs for each subject were used for the four repeated tasks. The first and second runs of the four repeated tasks were used in the enrolment process while the third run was used to test the authorized subjects for each electrode. The Butterworth filter was applied to signals. The difference between two signals was calculated using the percent residual difference (PRD), which is a quantitative measurement:

$$PRD_n = \sqrt{\frac{\sum_{i=1}^{M}(x_0(i)-x_n(i))^2}{\sum_{j=1}^{M}(x_0(j)-\overline{x_0})^2}} \times 100\% \qquad (1)$$

Where: xo is considered as the unknown signal while xn is the enrolled signal for subject n [25, 26].

A verification model was used in this work for personal identification, which is preferred with a large number of subjects to get the system decision in a very short time. PRDth was calculated during the enrolment process for each subject. PRDn was calculated for each imposter (from the third run of tasks). A decision is obtained based on the stored database using the person's ID as illustrated in the proposed method in Fig. 5.

TABLE II. DATA SET DESCRIPTION USED IN THIS EXPERIMENT

Dataset Properties	Information
NUMBER OF SUBJECTS	109
NUMBER OF CHANNELS	64
SAMPLING FREQUENCY	160 HZ
SIGNAL TIME LENGTH	60 s
SYSTEM	BCI2000

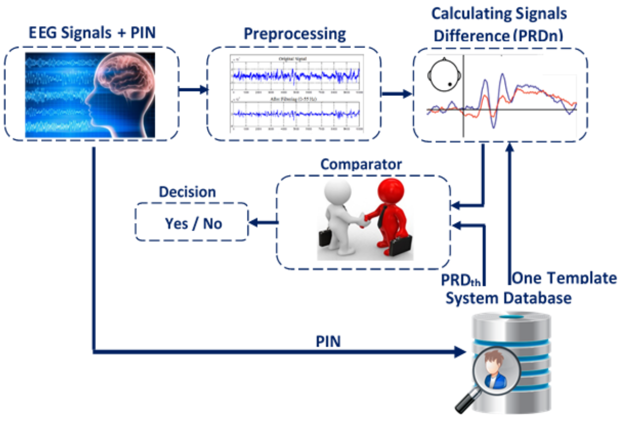


Fig. 5. The Proposed System Diagram.

IV. Experiment and Results

In this study, all 64 channels were used and tested for authorized and unauthorized subjects using the four tasks together and with each task alone as mentioned before. The system was tested five times, namely, system-based all four tasks together, system-based task 1, system-based task 2, system-based task 3, and system-based task 4. The system's performance is evaluated using the following measurements:

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fn+Fp} \times 100 \qquad (2)$$

$$FRR = \frac{Fn}{Tp+Fn} \times 100 \qquad (3)$$

$$FAR = \frac{Fp}{Tn+Fp} \times 100 \qquad (4)$$

$$Recall = \frac{Tp}{Tp+F_N} \times 100 \qquad (5)$$

$$Percision = \frac{Tp}{Tp+Fp} \times 100 \qquad (6)$$

$$F-score = \frac{2 \times Percision \times Recall}{Percision + Recall} \qquad (7)$$

Where: Tp is true positive (Accepted knowns), Tn is true negative (rejected unknowns), Fn is the false negative (rejected knowns) and Fp is the false positive (accepted unknowns) [27,28].

The obtained results will be presented in detail in the following sections, starting with task 2 which gives the highest system performance.

A. The Proposed System based on Task 2

The results of the system performance based on task 2 for all 64 channels individually are shown in Table III. The results in this table are arranged in descending order according to the system accuracy. The obtained results show that the channels located in the parietal (yellow) and occipital (pink) lobes give higher system accuracy than channels located in the frontal (blue) and temporal (green) lobes. These selected colors are derived from Fig. 1. This experiment is performed using single-channel mode because the target is trying to find the best one for human identification.

Also, z electrodes located in the vertical center give higher system accuracy (in the parietal and occipital lobes). Four electrodes give 100% system accuracy-based task 2 (Pz, P6, CP3, and CP1). FRR reaches 0% using 12 electrodes placed in the parietal and occipital lobes and FAR reaches 0% using 10 electrodes placed in the same lobes. Recall gives 1 using 12 electrodes while precision and f-score are given 1 using 10 electrodes distributed in the parietal and occipital lobes. The first four electrodes (Pz, P6, CP3, and CP1) in the table are given optimum system performance.

B. The Proposed System based on the Four Tasks together

The system accuracy based on the four tasks together for all 64 channels reaches 99.08 % using 8 channels individually and FRR reaches 0 % using 9 channels placed in the parietal and occipital lobes while FAR gives 0 % using 5 electrodes in the same lobes. Recall gives 1 using 9 channels, Precision reaches to 1 using 5 channels while f score starts at 99.5 %.

TABLE III. SYSTEM PERFORMANCE USING TASK 2

Elect.	Acc. %	FRR%	FAR%	Recall%	Per. %	f score %
Pz	100	0	0	100	100	100
P6	100	0	0	100	100	100
CP3	100	0	0	100	100	100
CP1	100	0	0	100	100	100
CP2	99.08	1	0	99	100	99.5
P3	99.08	0	11.11	100	99.01	99.5
P2	99.08	1	0	99	100	99.5
P4	99.08	1	0	99	100	99.5
PO3	99.08	0	11.11	100	99.01	99.5
Poz	99.08	0	11.11	100	99.01	99.5
Oz	99.08	0	11.11	100	99.01	99.5
C1	99.08	0	11.11	100	99.01	99.5
C2	99.08	1	0	99	100	99.5
P5	99.08	0	11.11	100	99.01	99.5
P1	98.17	2	0	98	100	98.99
O1	98.17	1	11.11	99	99	99
CPZ	98.17	0	22.22	100	98.04	99.01
PO7	98.17	1	11.11	99	99	99
P8	98.17	2	0	98	100	98.99
PO4	98.17	1	11.11	99	99	99
C3	98.17	1	11.11	99	99	99
O2	97.25	2	11.11	98	98.99	98.49
CP4	97.25	2	11.11	98	98.99	98.49
Cz	97.25	0	33.33	100	97.09	98.52
CP5	96.33	1	33.33	99	97.06	98.02
C4	96.33	1	33.33	99	97.06	98.02
CP6	96.33	1	33.33	99	97.06	98.02
P7	96.33	2	22.22	98	98	98
PO8	94.5	5	11.11	95	98.96	96.94
Iz	94.5	2	44.44	98	96.08	97.03
TP8	93.58	6	11.11	94	98.95	96.41
FC2	93.58	1	66.67	99	94.29	96.59
FC4	93.58	3	44.44	97	96.04	96.52
FCz	92.66	4	44.44	96	96	96
FC3	91.74	4	55.56	96	95.05	95.52
C6	90.83	6	44.44	94	95.92	94.95
C5	89.91	2	100	98	91.59	94.69
Fz	89.91	4	77.78	96	93.2	94.58
F4	89.91	4	77.78	96	93.2	94.58
T9	88.99	8	44.44	92	95.83	93.88
TP7	88.99	7	55.56	93	94.9	93.94
FC1	88.99	3	100	97	91.51	94.17
FC6	88.99	5	77.78	95	93.14	94.06
F2	88.99	5	77.78	95	93.14	94.06
FT7	88.07	7	66.67	93	93.94	93.47
T7	88.07	7	66.67	93	93.94	93.47
F3	87.16	7	77.78	93	93	93
F8	87.16	6	88.89	94	92.16	93.07
F6	87.16	6	88.89	94	92.16	93.07
F5	86.24	8	77.78	92	92.93	92.46
FT8	86.24	7	88.89	93	92.08	92.54
FC5	85.32	7	100	93	91.18	92.08
Afz	85.32	9	77.78	91	92.86	91.92
F1	85.32	7	100	93	91.18	92.08
AF4	85.32	8	88.89	92	92	92
F7	84.4	9	88.89	91	91.92	91.46
AF8	84.4	9	88.89	91	91.92	91.46
T8	84.4	13	44.44	87	95.6	91.1
AF3	82.57	12	77.78	88	92.63	90.26
T10	82.57	15	44.44	85	95.51	89.95
Fpz	80.73	14	77.78	86	92.47	89.12
Fp2	80.73	14	77.78	86	92.47	89.12
Fp1	78.9	16	77.78	84	92.31	87.96
AF7	78.9	14	100	86	90.53	88.21

C. *The Proposed System based on Task 1*

The same observations are obtained again by repeating the experiment based on task 1 individually, but the system accuracy decreases and starts at 96.33 %. FAR gives 0 % using 3 channels. FRR is higher than that in the previous experiment. Recall starts from 98 %, precision gives 1 using 3 channels and f-score starts from 96 %, so the system performance based on the four tasks together is better than that based on task 1 alone.

D. *The Proposed System based on Task 3*

It was observed from the system performance-based task 3 for all 64 channels that O2 channel gives the highest system accuracy using the proposed system-based task 3 alone with 0 % FRR and 11.11% FAR. FRR is accepted using all electrodes, especially in electrodes placed in the parietal and occipital lobes than others, while FAR increases gradually and starts at 11.11%. Recall gives 1 using one channel, precision reaches to 99 % and 99.5 % f-score.

E. *The Proposed System based on Task 4*

From the system performance-based task 4 for all 64 channels individually, 9 electrodes give the same system accuracy of 98.17 % with 0 % FAR (100% Specificity) and 2% rejected authorized subjects only and overall, the FRR is accepted based on task4 while FAR increases in frontal and temporal electrodes. Precision reaches to 1 using 12 channels, recall reaches to 98 % and f-score reaches to 98.9 %.

F. The Proposed System Performance Measures

Some measures are calculated for all 64 channels using the four tasks together and with each task individually. Fig. 6 shows the average system accuracy.

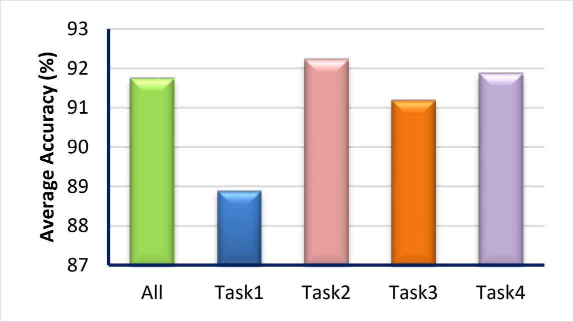


Fig. 6. The Average System Accuracy.

From the above figure, it was observed that the proposed system-based task 2 has the higher system accuracy, followed by task 4. Task 2 and task 4 are imaginary tasks, which means that using the imaginary tasks is preferred for human authentication than non-imaginary tasks. Fig. 7 shows the average system False Reject Rate (FRR).

FRR measures the identification percentage for authorized persons who are incorrectly rejected. The proposed system-based task 3 has the best FRR is 3.8 %, followed by using the four tasks together is about 4% and then task 2, which is 4.45 % as shown in Fig. 7.

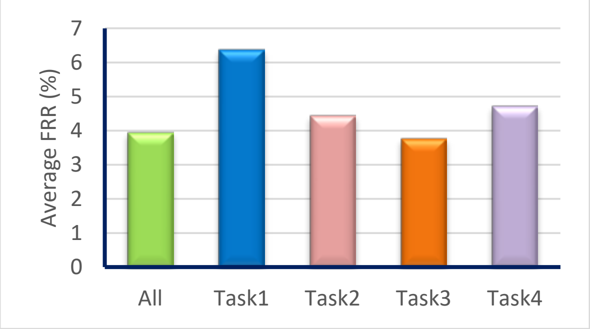


Fig. 7. The Average System FRR.

Fig. 8 shows the average system False Acceptance Rate (FAR). FAR measures the identification percentage for unauthorized persons who are incorrectly accepted.

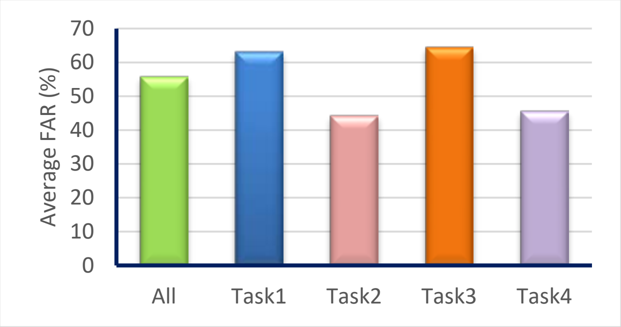


Fig. 8. The Average System FAR.

The system FAR based on task 2 is about 44.4 %, which is lower than others, followed by that based on task 4, which is about 45.7%. More studies will be tried to improve FAR for the proposed system.

Fig. 9 shows the average system Recall, it quantifies the number of authorized persons who are incorrectly rejected.

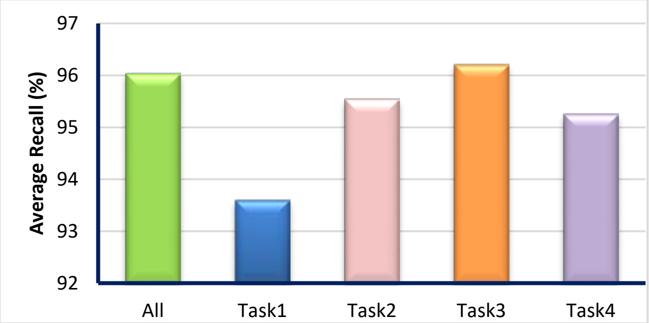


Fig. 9. The Average System Recall.

The average system recall for task 3 is about 96.2 % followed by that based on all tasks together 96 %, then task 2, which is equal to 95.5%. Fig. 10 shows the average system precision. It quantifies the number of authorized persons who are correctly accepted.

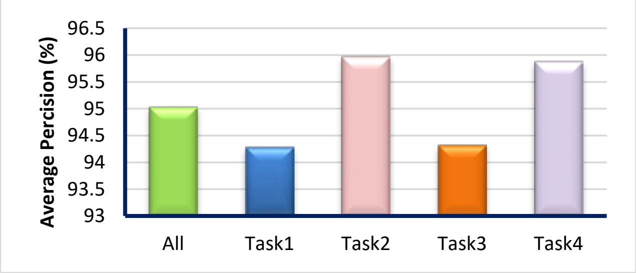


Fig. 10. The Average System Precision.

The average system precision based on task 2 is 95.98% followed by that based on task4 which is 95.89% (approximately the same for imaginary tasks). The average system precision for all tasks together comes after the imaginary tasks, while that based on the non-imaginary tasks gives 94 %.

Fig. 11 shows the average system F-score. It is a single value that gives the balance between precision and recall for the proposed system.

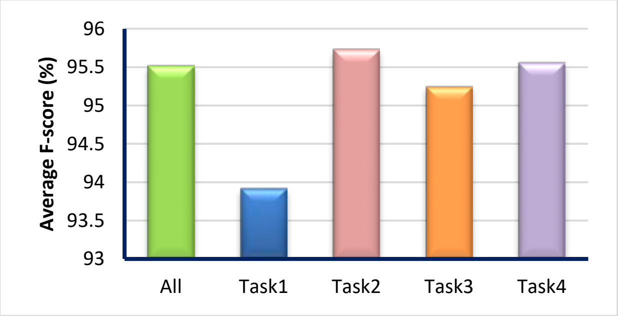


Fig. 11. The Average System F-score.

The average system F-score based on task 2 is 95.7 %, which is the best score obtained by the proposed system, followed by that based on task 4, which equals 95.5%.

From this experiment, it was observed that using EEG signals collected from some electrodes placed in the parietal and occipital lobes gives a good biometric system performance for authorized and unauthorized subjects. While using electrodes placed in frontal and temporal lobes gives accepted performance for authorized subjects only. Also, not all tasks are preferred for human identification based on EEG signal. Only imaginary tasks give better system performance than non-imaginary tasks.

V. Conclusion

Biometric system-based physiological traits were more secure and difficult to mimic or penetrate. Also, using verification mode is preferred with a large number of subjects to save time. EEG signals were used for personal identity. Nowadays, it is easy to send EEG or ECG signals by network using simple sensors. So, using EEG may become a powerful tool for human identification in the next few days. Trying to find a single electrode with high efficiency is very important for using EEG as a biometric trait to become easy to use and fast to make a decision. This work is trying to find this electrode by using existing data. This data presents EEG signals obtained by performing four tasks (two of them imaginary tasks and the others are non-imaginary tasks). This work-based on PRD, which is a simple measurement used to measure the difference between two signals. The proposed system was repeated for five experiments. The first one is based on using the four tasks together to make the system decisions, the second is based on task 1 only, the third is based on task 2, the fourth is based on task 3 while the fifth is based on task 4. Tasks 1 and 3 are non-imaginary tasks, while tasks 2 and 4 are imaginary.

The observations from the obtained results show that the proposed system-based task 2 is better than using other tasks individually and the four tasks together, the system performance-based task 4 is better than that based on task 1 and task 3 individually, which means that using the imaginary tasks is more suitable than using non-imaginary tasks for human authentication. Also, using electrodes placed in the parietal and occipital lobes is better than those placed in other lobes, and electrodes near the vertical brain center give higher system performance. Pz, P6, CP3, and CP1 electrodes give 100% system accuracy-based task 2 individually. It was observed that the proposed system sensitivity (recognize authorized subjects) is good and acceptable using all electrodes and increases as mentioned before in the parietal and occipital lobes electrodes, while the proposed system specificity (recognize unauthorized subjects) is acceptable in some electrodes placed in parietal and occipital lobes only, so more studies needed to improve the proposed system specificity. The proposed system gives instantaneous decision because its construction is based on verification mode, which makes it very suitable for large numbers.

References

[1] Marwa A. Elshahed,” Personal identity verification based ECG biometric using non-fiducial features”, International Journal of Electrical and Computer Engineering (IJECE), ISSN: 2088-8708, DOI: 10.11591/ijece.v10i3.pp3007-3013,2020.

[2] Mohamad El-Abed and Christophe Charrier, “Evaluation of Biometric Systems”, New Trends and Developments in Biometrics, pp. 149 - 169, ff10.5772/52084ff. ffhal-00990617f, 2012.

[3] A.S. Raju and V. Udayashankara,” Performance Evaluation of Multimodal Biometrics System”, International Journal of Pure and Applied Mathematics, Volume 118 No. 5, 367-382,2018.

[4] Abdullah Alhayajneh, Alessandro N. Baccarini, Gary M. Weiss, Thaier Hayajneh and Aydin Farajidavar,” Biometric Authentication and Verification for Medical Cyber Physical Systems”,MDPI, Electronics 7(12), 436, 2018.

[5] K. P. Thomas and A. P. Vinod, “Utilizing individual alpha frequency and delta band power in EEG based biometric recognition,” in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 004787–00479, 2016.

[6] Q. Gui, Z. Jin, and W. Xu, “Exploring EEG-based biometrics for user identification and authentication,” 2014 IEEE Signal Process. Med. Biol. Symp. IEEE SPMB 2014 - Proc., 2015.

[7] Masahiro Hakodaa, Hirokazu Miurab , Noriyuki Matsudab , Fumitaka Uchiob and Hirokazu Takib , “Measurement of Brain Activity on Force Adjustment Skill Acquisition by using EEG”, International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September Marseille, France.2017.

[8] https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/electroencephalogram.

[9] A. Elakkiya, S.Ramkumar, G. Emayavaramban and P.Buvaneswari,” A Survey of Biometrics Person Identification System Using EEG Brain Signal”, International Journal of Pure and Applied Mathematics, Volume 119, No. 15 2018, 3471-3475, ISSN: 1314-3395,2018.

[10] S. Siuly, Y. Li, Y. Zhang, EEG signal analysis and classification, (New York, NY: Springer Berlin Heidelberg), ISBN 9783319476520, 2017.

[11] Zuzana Koudelková1 and Martin Strmiska,” Introduction to the identification of brain waves based on their frequency”, MATEC Web of Conferences 210, 05012 ,2018.

[12] Jiawei Zhang,” Secrets of the Brain: An Introduction to the Brain Anatomical Structure and Biological Function”, IFM LAB TUTORIAL SERIES # 4,2019.

[13] https://www.md-health.com/Lobes-Of-The-Brain.html.

[14] Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel , Christoph Baumgartner, Bin He and Sándor Beniczky,” The standardized EEG electrode array of the IFCN”, Clinical Neurophysiology journal,2017.

[15] International 10-20 system manual, Trans Cranial Technologies ldt, 2012.

[16] https://archive.physionet.org/physiobank/database/eegmmidb/.

[17] Chesada Kaewwit, Chidchanok Lursinsap and Peraphon Sophatsathit,” High Accuracy EEG Biometrics Identification using ICA And AR Model”, Journal of ICT, 16, pp: 354–373, (Dec) 2017.

[18] R. Suppiah and A. P. Vinod, "Biometric identification using single channel EEG during relaxed resting state," in IET Biometrics, vol. 7, no. 4, pp. 342-348, 7 2018.

[19] Ramaswamy Palaniappan, Jenish Gosalia, Kenneth Revett and Andrews Samraj,” PIN Generation Using Single Channel EEG Biometric”, Springer, ACC 2011, Part IV, CCIS 193, pp. 378–385, 2011.

[20] Isao Nakanishi, Sadanao Baba and Chisei Miyamoto,“EEG Based Biometric Authentication Using New Spectral Features”,Proc. of 2009 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2009), Dec 2009.

[21] Sarineh Keshishzadeh, Ali Fallah and Saeid Rashidi,” Electroencephalogram Based Biometrics: A Fractional Fourier Transform Approach”, ICBEA '18, ACM ISBN 978-1-4503-6394-5/18/05,2018.

[22] JISSY J, RESHMA V K,” Brain Fastener as Eeg Biometric Authentifier”, International Research Journal of Engineering and Technology (IRJET), Volume: 07, Issue: 06 , June 2020.

[23] 23 Meriem Romaissa Boubakeur and Guoyin Wang,” Self-Relative Evaluation Framework for EEG-Based Biometric Systems”, MDPI, Sensors 2021, 21, 2097, 2021.

[24] http://www.bci2000.org

[25] Chan, Adrian D.C.; Hamdy, Mohyledin M.; Badre, Armin and Badee, Vesal,” wavelet distance measure for person identification using electrocardiograms”, IEEE Transactions on Instrumentation and Measurement · February 2008.

[26] Ruqaiya Khanam and Syed Naseem Ahmad,” Selection of Wavelets for Evaluating SNR, PRD and CR of ECG Signal”, International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.

[27] Kareem Hatam Nahavandi,” Calculating Sensitivity, Specificity and Predictive Values for MedicalDiagnostic Tests”, Gene Cell Tissue, doi: 10.5812/gct.80270,2018.

[28] Mahsa Zeynali and Hadi Seyedarabi,” EEG-based single-channel authentication systems with optimum electrode placement for different mental activities”, biomedical journal 4 2 (2 0 1 9) 2 6 1 -2 6 7, 2019.

Evaluation of Machine Learning Algorithms for Intrusion Detection System in WSN

Mohammed S. Alsahli[1], Marwah M. Almasri[2], Mousa Al-Akhras[3], Abdulaziz I. Al-Issa[4], Mohammed Alawairdhi[5]
College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, KSA[1, 2, 3, 4, 5]
College of Computing and Informatics, Saudi Electronic University; King Abdullah II School of Information Technology
The University of Jordan, Riyadh 11673, KSA; Amman 11942, Jordan[3]

***Abstract*—Technology has revolutionized into connecting "things" together with the rebirth of the global network called Internet of Things (IoT). This is achieved through Wireless Sensor Network (WSN) which introduces new security challenges for Information Technology (IT) scientists and researchers. This paper addresses the security issues in WSN by establishing potential automated solutions for identifying associated risks. It also evaluates the effectiveness of various machine learning algorithms on two types of datasets, mainly, KDD99 and WSN datasets. The aim is to analyze and protect WSN networks in combination with Firewalls, Deep Packet Inspection (DPI), and Intrusion Prevention Systems (IPS) all specialized for the overall protection of WSN networks. Multiple testing options were investigated such as cross validation and percentage split. Based on the finding, the most accurate algorithm and the least time processing were suggested for both datasets.**

Keywords—Internet of Things (IoT); Wireless Sensor Network (WSN); Information Technology (IT); Denial of Service (DoS); Artificial Intelligence (AI); Machine Learning (ML)

I. Introduction

With the rapid expansion of technology, new threats and security issues arise, which become a hot area for research. Wireless Sensor Network (WSN) is composed of distributed wireless sensor nodes that collect raw data from the surrounding environment. Each Sensor node is equipped with a radio transceiver, a small microcontroller, and a power source [1]. These nodes are very small and have limited processing capabilities. They are designed based on low-cost and low-energy consumption that provide limited processing power and limited communication as represented in Fig. 1. Due to the sensors' limitation in memory, processing power, and energy consumption, there are several potential security challenges inherently exist and should be properly addressed. The primary challenge is to protect the WSN without the availability of massive processing power and energy. Traditional security measures such as encryption is difficult to be implemented at the senor's level due to its limited processing capabilities.

With the increased and sophisticated attack types on networks and applications, it is difficult to protect them against such attacks manually or by common Off-The-Shelf software such as firewalls, antivirus, Intrusion Detection System (IDS) or Intrusion Prevention System IPS). This makes artificial intelligence (AI) and machine learning (ML) algorithms popular and ultimately essential in such scenarios. AI in general and ML in specific can be used to protect WSN by identifying and classifying potential attacks by learning previously detected patterns of attacks.

Machine learning is becoming more popular in recent years. It enables machines or computers to work and react similar to what humans do. These systems improve with experience by learning the expected behavior. AI can be applied in many applications such as natural language processing and generation, speech recognition, virtual agent, machine learning, deep learning, biometrics, robotic process automation, text analytics and Neuro-Linguistic Programming (NLP), as well as in many domains such as healthcare, business, education, autonomous vehicles, robotics, government, and public safety and security. Moreover, AI becomes very useful in predictive analysis and plays a fundamental role in the software field and content creation.

This paper investigates different datasets with different machine learning algorithms, namely Naïve Bayes, improved Naïve Bayes, IBK, and Random Forest algorithms in multiple scenarios. The purpose is to identify the best method to mitigate the risks, threats, and security vulnerabilities associated with WSN networks.

The rest of this paper is organized as follows. Section II discusses related work. Section III presents the underlying concepts and proposed methodology. Section IV shows the experimental results. Section V discusses and analyzes the findings. Finally, section VI concludes the paper.

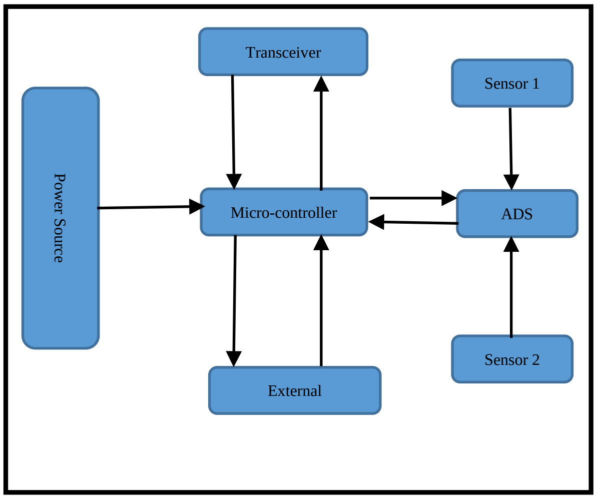


Fig. 1. WSN Mechanism.

II. Related Work

This section presents some researches about various attacks in WSN. In [2], authors have addressed Denial of Service (DoS) cyber-attacks on Wireless Sensor Networks (WSN) and how to mitigate these attacks. The researchers used specialized datasets for WSN constructed for classifying the types of attacks for their research. Four DoS attacks were considered: Flooding attack, Blackhole attack, Scheduling attack, and Gray-hole attack. The main purpose was to help WSN manufacturers to create and develop a system that detects and protects against DoS attacks in WSN. They have also discussed the challenges of protecting these networks due WSN limitations such as low processing, low power, and limited storages. They emphasized on the importance of mitigating and protecting against new and unprecedented attacks [2].

Moreover, the authors in [3] have focused on the classification's accuracy improvement of the Naïve Bayes algorithm, by finding more accurate probability estimation. This helps in solving the lack of the training data. Their approach was applied during the training phase without increasing the classification time. The first phase was building the classical Naïve Bayes classifier then fine-tune it in the second phase. Each training instance was classified, and if it is misclassified, it will contribute in fine tuning the probability value. Therefore, it will be correctly classified in the next round. Based on the findings, results showed an improved classification accuracy of many datasets.

Many researches have defined Wireless Sensor Network (WSN). It is typically composed of sensor nodes. These nodes gather data about the environment and send it back to the sink or the base station node. These data can be in different formats such as thermal, acoustic, optical, weather, pressure, chemical, and much more. It is extremely challenging task to develop an algorithm that is suitable for many applications scenarios in a diverse WSN environment; especially, considering data reliability and aggregation, localization, clustering, fault detection, and security [4].

Furthermore, the authors have highlighted the importance of utilizing ML in WSN for the following reasons [4]:

1) Using ML techniques could help in observing dynamic environments.

2) In some cases, WSN gathers new data in out-of-reach or threatening locations.

3) Accurate models are hard to be obtained in WSN since they are usually applied in sophisticated environments

4) Using ML techniques could be beneficial in extracting essential correlations.

The authors in [5] have emphasized the growing number of services that are providing facilities to humans which make using WSN valuable in many applications such as security systems, fire safety, various military applications, monitoring environmental conditions, and monitoring health condition. However, these WSNs encounter some weaknesses because of the nodes' exposure to various security attacks due to their limitations in power, processing, memory storage, bandwidth, data transmission via other nodes and multiple hops, its distributed nature, and self-organization. These attacks occurs at different levels of the OSI models. Therefore, it is important to build a security defense and monitoring system to protect against these attacks [5].

Similarly, the authors in [6] have discussed WSNs and their crucial role in different applications and usage; the vulnerabilities of the WSN due to their constrained resources. How DoS attack can be carried out at different layers of the network architecture. The authors focused specifically on the network layer because of the diversity of the attack at this layer. The authors reviewed many studies that use machine learning techniques pertaining to the network layer DoS attacks in WSN [6].

IDS and their important role in protecting against malicious attacks that affect the performance of the network have been addressed in [7]. The authors described Mobile Ad hoc networks (MANETs), WSN, and Internet of Things (IoT). The significance of the IDS and the need to protect such networks. Their proposed an IDS that has two stages. One that collects data using sniffers to generate correctly classified instances and in the second stage, a super node process data from different IDSs to differentiate benign from malicious nodes [7].

III. Underlying Concept and Methodology

This section presents the dataset types as well as the used machine learning techniques.

A. Datasets

A dataset is a collection of records that is gathered in a controlled lab environment. In this paper, two different datasets were used. The first dataset is called "KDDCup99 Dataset" which was derived from the DARPA 1998 dataset [8], [9]. It was selected and used to detect network breaches from a network security perspective. A network breach is the abuse of data and information to bypass the security rules and established regulations.

The authors in [10], have explained that the discovery of this interruption is a set of strategies and related activities that enable the progression of perceived methods for the identification of security classification. This dataset was provided by the archive, which was for a data mining competition held in aligning with KDD-99.

The author in [11] indicates that the features were to create a model that detects the bad connections or attacks as well as normal connections. The complete listing of the features defined for the connection records is listed in Table I.

TABLE I. DESCRIPTION OF KDDCUP99 DATASET FEATURES

	Feature Name	*Description*
1.	Duration	Number of seconds of the connection
2.	protocol_type	Type of the protocol, e.g., TCP, UDP, etc.
3.	Service	Network service on the destination, e.g., http, telnet, etc.
4.	Flag	Normal or error status of the connection
5.	src_bytes	Number of data bytes from source to destination
6.	dst_bytes	Number of data bytes from destination to source
7.	Land	1-connection is from/to the same host/port; 0-otherwise
8.	wrong_fragment	Number of ‘wrong’ fragments
9.	Urgent	Number of urgent packets
10.	Hot	The count of access to system directories, creation and execution of programs
11.	num_failed_logins	Number of failed login attempts
12.	logged_in	1 - successfully logged in; 0 otherwise
13.	num_compromised	Number of “compromised” conditions
14.	root_shell	1 - root shell is obtained; 0 otherwise
15.	su_attempted	1 – ‘su root’ command attempted; 0 – otherwise
16.	num_root	number of ‘root’ accesses
17.	num_file_creations	Number of file creation operations
18.	num_shells	Number of shell prompts
19.	num_access_files	Number of write, delete, and create operations on access control files
20.	num_outbound_cm ds	Number of outbound Commands in a ftp session
21.	is_hot_login	1 - the login belongs to the ‘hot’ list (e.g., root, adm, etc.) ; 0 – otherwise
22.	is_guest_login	1 - the login is a ‘guest’ login (e.g., guest, anonymous, etc.) ; 0 – otherwise
23.	Count	Number of connections to the same host as the current connection in the past 2 seconds
24.	srv_count	Number of connections to the same service as the current connection in the past 2 seconds
25.	serror_rate	% of connections that have ‘SYN’ errors to the same host
26.	srv_serror_rate	% of connections that have ‘SYN’ errors to the same service
27.	rerror_rate	% of connections that have ‘REJ’ errors to the same host
28.	srv_rerror_rate	% of connections that have ‘REJ’ errors to the same service
29.	same_srv_rate	% of connections to the same service and to the same host
30.	diff_srv_rate	% of connections to different services and to the same host
31.	srv_diff_host_rate	% of connections to the same service and to different hosts
32.	dst_host_count	Number of connections to the same host to the destination host as the current connection in the past 2 seconds
33.	dst_host_srv_count	Number of connections from the same service to the destination host as the current connection in the past 2 seconds
34.	dst_host_same_srv_rate	% of connections from the same service to the destination host
35.	dst_host_diff_srv_rate	% of connections from the different services to the destination host
36.	dst_host_same_src_port_rate	% of connections from the port services to the destination host
37.	dst_host_srv_diff_ host_rate	% of connections from the different hosts from the same service to destination host
38.	dst_host_serror_rate	% of connections that have ‘SYN’ errors to same host to the destination host
39.	dst_host_srv_ serror_rate	% of connections that have ‘SYN’ errors from same service to the destination host
40.	dst_host_rerror_rate	% of connections that have ‘REJ’ errors from the same host to the destination host
41.	dst_host_srv_ rerror_rate	% of connections that have ‘REJ’ errors from the same service to the destination host

The second used dataset is called “WSN Dataset” [12], which is specialized for WSN. It is used to detect different types of DoS attacks as well as normal behavior. The dataset is collected with different features and divided into different classes such as Blackhole, Grayhole, Scheduling, Flooding, and Normal. Low Energy Aware Cluster Hierarchy (LEACH) is the routing protocol that is used to collect the dataset that contains hundreds of thousands of records in WSN. It is designed to keep energy consumption low which is very important to maintain and improve the lifetime of WSN [13]. The problem or the limitation of LEACH is that it is only suitable for a small size WSN [13]. It assumes that all sensors can communicate with each other and with the sink (base station) as shown in Fig. 2. Table II represents the different WSN dataset attributes.

B. *Machine Learning Techniques*

Machine learning techniques are broadly categorized as unsupervised and supervised learning, which are for clustering, and classification/regression, respectively, as depicted in Fig. 3. Classification is a problem-solving technique for analyzing datasets or data models using algorithms such as Naïve and IBK. Regression is commonly used as a statistical tool to predict potential outcomes. The following subsections demonstrate various machine learning algorithms that were implemented on the above mentioned datasets.

1) Naïve Bayes: Naïve Bayes (NB) is a machine learning algorithm for AI software and computers. NB is based on mathematical calculation of probabilities that uses datasets (raw data or simple facts) to learn a concept. NB is used in a wide range of real applications and automated decision-making processes. A Naïve Bayes classifier is an algorithm that uses Bayes theorem features to classify objects. A NB is also known as simple Bayes or an independent Bayes. These classifiers use regular (or Naïve) independence intervals between the attributes of a data point.

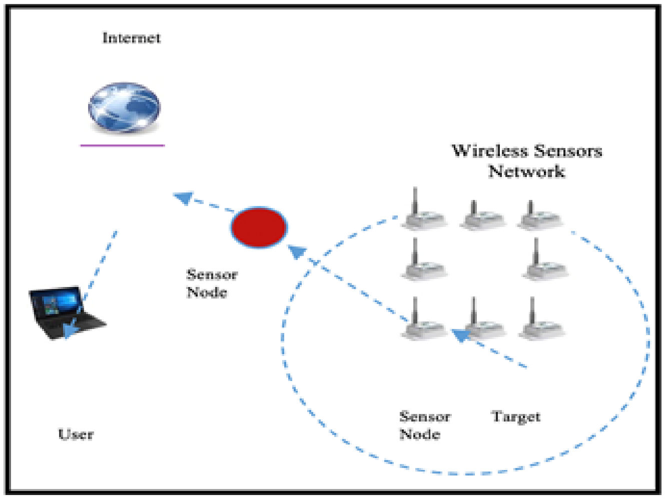


Fig. 2. WSN Network.

TABLE II. WSN-DS DATASET ATTRIBUTES [12]

#	*Attribute Name*	*Attribute Description*
1	Node ID	A unique ID to distinguish the sensor node in any round and at any stage
2	Time	The current simulation time of the node
3	Is CH	A flag to distinguish whether the node is CH or not
4	Who CH	The ID of the CH in the current round
5	Distance to CH	The distance between the node and its CH
6	Energy Consumption	The amount of energy consumed in the previous round
7	ADV_CH send	The number of advertise CH’s broadcast messages sent to the nodes
8	ADV_CH receives	The number of advertise CH messages received from CHs
9	Join_REQ send	The number of join request messages sent by the nodes to the CH
10	Join_REQ receives	The number of join request messages received by the CH from the nodes
11	ADV_SCH send	The number of advertise TDMA schedule broadcast messages sent to the nodes
12	ADV_SCH receives	The number of TDMA schedule messages received from CHs
13	Rank	The order of this node within the TDMA schedule
14	Data sent	The number of data packets sent from a sensor to its CH
15	Data Received	The number of data packets received from CH
16	Data sent to BS	The number of data packets sent to the BS
17	Distance CH to BS	The distance between the CH and the BS
18	Send Code	The cluster sending code
19	Attack Type	Type of the node. It is a class of five possible values, namely, Blackhole, Grayhole, Flooding, and Scheduling, in addition to normal, if the node is not an attacker

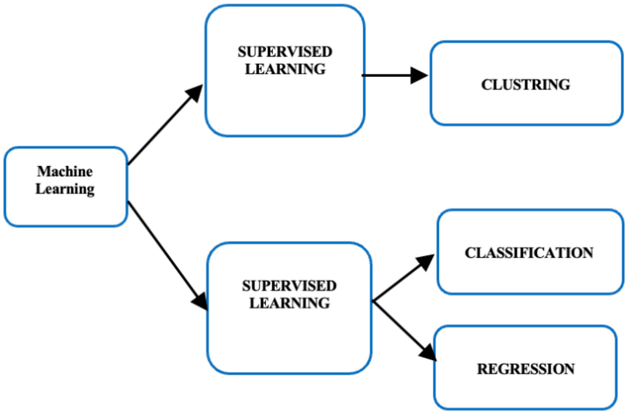


Fig. 3. Machine Learning.

The most common and widespread use of these Bayes algorithm is the use of spam filters or text and medical analysis. As these classifiers are easy to implement, they are most commonly used for machine learning. As stated by [14], Naïve Bayes classification uses probability theory to classify the data and makes use of Bayes theorem in its algorithm. The main feature of this classifier is that there can be an adjustment of the probability of an event as new data is introduced. It also assumes all the attributes that are in consideration are independent of each other. A Naïve Bayes classifier is not a single algorithm, but instead, it is a combination of specific machine learning algorithms in which statistical independence methods are used. A Naïve Bayes classifier makes a proper decision rule classification as long as the required class is more probable than any other present class. This fact is deemed accurate, as there is a slight inaccuracy in the probability estimation most of the times [3].

2) Fine Tune Naïve Bayes (FTNB): With respect to the Naïve Bayes classification, the tuning of parameters is limited, and it is recommended to improve the quality of the pre-processing and feature selection processes. The classifier performance and prediction can be improved by tuning and adjusting the classifier parameters, applying classifier combination techniques, or by monitoring the data fed to the classifier- either adding more data, refining existing one, or improving them [3].

3) Data Parsing (pre-processing): According to [15], the data is a string of raw text presented for each data point. A series of processes and steps convert this data into a structured vector such that the offset shows one feature and the value in the offset is correspondent to the frequency. Stemming, synonym finding and use of neutral words in the raw data text are one of the ways to improve the data parsing or the data processing methods.

a) Selection of Features: According to [16], the use cases for a Naïve Bayes classification like spam filtering are observed and utilized by showing how they fail or quickly can be improved. For assumption, an above average spam filter has a feature like a word frequency in all caps and words in titles or the occurrence of exclamation symbol in the title. The best feature for improvement is the use of long words or a group of more than a single word.

4) IBK algorithm: Instance Base Learner (IBK) algorithm is used in distance measure and classifying instances based on K-nearest neighbors to make predictions [17]. The computation in the test phase is very high and takes a long time, especially for a huge number or instances in the dataset. The default value of neighbors is 1. Sometimes called 1-NN [18].

5) Random forest algorithm: Random Forest or random decision forest algorithm is used for classification and regression of an ensemble of the collection of datasets. In WEKA program, Random Forest can only do the classification part, not the regression task. It operates by building a great number of decision trees in the training phase and perform the classification task. In WEKA, there is no output of the mean prediction or regression of each tree. Random Forest classification mean mapping input data in the dataset or instances to a category. This is also called categorization of the instances. The algorithm that does the classification, especially in the concrete implementation, is called the classifier [19].

IV. Experiments and Result

This section discusses and demonstrates the experiments conducted and their results. Both datasets have been classified using the above-mentioned machine learning algorithms (section III-B) using Cross-validation and percentage split techniques. Cross validation is a standard analysis tool used to verify the validity of the data mining model. It works by dividing the dataset into a number of folds or pieces and hold each fold in turn for testing and training all of the other pieces in the system. In dividing the dataset into layers or folds, it ensures that each layer or fold had the correct portion of class values [20]. Additionally, Percentage split determines the percentage used for training the system [20]. For our experiments, 66% was used for training and 34% was used for testing. The following subsections demonstrate the results obtained by each algorithm conducted on both datasets using cross-validation and percentage split techniques.

A. Naïve Bayes (NB) Algorithm

1) Cross-validation technique: Table III shows the results of running NB algorithm on both datasets (KDDCUP99 and WSN-DS) using cross-validation technique. Table IV demonstrates the weighted average accuracy using cross-validation technique in terms of several factors such as:

- True Positive Rate (TP): the rate that the system or an algorithm correctly classifies an instance as a positive class.
- True Negative Rate (TN): the rate that the system or an algorithm correctly classifies an instance as a negative class.
- False Positive Rate (FP): the rate that the system or an algorithm falsely (wrongly) classifies an instance as a positive class/.
- False Negative Rate (FN): the rate that the system or an algorithm falsely (wrongly) classifies an instance as a negative class.
- Precision: the ratio of correctly classified instances as positive to the instances that are classified by the algorithm as positive.
- Recall: the ratio of correctly classified instances as positive to the positive instances (whether classified correctly or not).
- Receiver Operating Characteristics (ROC): is a technique used as graph or curve to represent or visualize the performance of the classifiers. It is widely used in machine learning, data mining, and decision making. Also, it is used as a method of comparing diagnostic tests.

2) Percentage split technique: In this experiment, 66% of the data was used for training and 34% for testing. Table V shows the results of running Naïve Bayes (NB) algorithm on both datasets (KDDCUP99 and WSN-DS) using the percentage split technique. In addition, Table VI demonstrates the weighted accuracy average using the percentage split technique in terms of TP, TN, Precision, and ROC.

B. IBK Algorithm

1) Cross-validation technique: Table VII shows the results of running IBK algorithm on both datasets (KDDCUP99 and WSN-DS) using the cross- validation technique. Table VIII demonstrates the weighted accuracy average using the cross-validation technique.

2) Percentage split technique: In this experiment, 66% of the data was used for training and 34% for testing. Table IX shows the results of running IBK algorithm on both datasets (KDDCUP99 and WSN-DS) using the percentage split technique. Table X demonstrates the weighted accuracy average using the percentage split technique in terms of several factors.

C. Random Forest Algorithm

1) Cross-validation technique: Table XI shows the results of running the Random Forest algorithm on both datasets (KDDCUP99 and WSN-DS) using the cross-validation technique. Table XII demonstrates the weighted accuracy average using the cross-validation technique.

2) Percentage split technique: In this experiment, 66% of the data was used for training and 34% for testing. Table XIII shows the results of running IBK algorithm on both datasets (KDDCUP99 and WSN-DS) using the percentage split technique. Table XIV demonstrates the weighted accuracy average using the percentage split technique in terms of several factors.

TABLE III. THE RESULTS OF NAÏVE BAYES (NB) ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

	Dataset		
	KDDCUP99		*WSN-DS*
Correctly Classified Instances	459019	92.9151 %	459019
Incorrectly Classified Instances	35001	7.0849 %	35001
Kappa statistic	0.8828		0.8828
Mean absolute error	0.0061		0.0061
Root mean squared error	0.0765		0.0765
Relative absolute error	11.955 %		11.955 %
Root relative squared error	47.6941 %		47.6941 %
Total Number of Instances	494020		494020

TABLE IV. THE WEIGHTED ACCURACY AVERAGE OF NAÏVE BAYES (NB) ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

Weighted Avg. of	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
KDDCup99	0.929	0.000	0.989	0.929	0.951	0.948	1.000	0.991
WSN Dataset	0.954	0.012	0.966	0.954	0.957	0.847	0.980	0.971

TABLE V. THE RESULTS OF NAÏVE BAYES (NB) ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

	Dataset		
	WSN		*KDDCUP99*
Correctly Classified Instances	121606	95.4634%	121606
Incorrectly Classified Instances	5779	4.5366 %	5779
Kappa statistic	0.7678		0.7678
Mean absolute error	0.0182		0.0182
Root mean squared error	0.1324		0.1324
Relative absolute error	26.2165 %		26.2165 %
Root relative squared error	71.0237 %		71.0237 %
Total Number of Instances	127385		127385

TABLE VI. THE WEIGHTED ACCURACY AVERAGE OF NAÏVE BAYES (NB) ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

Weighted Avg. of	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
KDDCup99	0.930	0.000	NA	0.930	NA	NA	1.000	0.991
WSN Dataset	0.955	0.011	0.967	0.955	0.958	0.851	0.981	0.972

TABLE VII. THE RESULTS OF IBK ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

	Dataset		
	KDDCUP99		*WSN*
Correctly Classified Instances	493796	99.9547 %	493796
Incorrectly Classified Instances	224	0.0453 %	224
Kappa statistic	0.9992		0.9992
Mean absolute error	0		0
Root mean squared error	0.0063		0.0063
Relative absolute error	0.0791 %		0.0791 %
Root relative squared error	3.9104 %		3.9104 %
Total Number of Instances	494020		494020

TABLE VIII. THE WEIGHTED ACCURACY AVERAGE OF IBK ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

Weighted Avg. of	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
KDDCup99	1.000	0.000	1.000	1.000	1.000	0.999	1.000	0.999
WSN Dataset	0.994	0.025	0.994	0.994	0.994	0.970	0.985	0.992

TABLE IX. THE RESULTS OF IBK ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

	Dataset		
	KDDCUP99		**WSN**
Correctly Classified Instances	167869	99.9417 %	167869
Incorrectly Classified Instances	98	0.0583 %	98
Kappa statistic	0.999		0.999
Mean absolute error	0.0001		0.0001
Root mean squared error	0.0071		0.0071
Relative absolute error	0.1024 %		0.1024 %
Root relative squared error	4.4419 %		4.4419 %
Total Number of Instances	167967		167967

TABLE X. THE WEIGHTED ACCURACY AVERAGE OF IBK ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

Weighted Avg. of	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
KDDCup99	1.000	0.000	1.000	1.000	1.000	0.999	1.000	0.999
WSN Dataset	0.994	0.025	0.994	0.994	0.994	0.970	0.985	0.992

TABLE XI. THE RESULTS OF RANDOM FOREST ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

	Dataset		
	KDDCUP99		*WSN*
Correctly Classified Instances	493915	99.9787 %	167869
Incorrectly Classified Instances	105	0.0213 %	98
Kappa statistic	0.9996		0.999
Mean absolute error	0.0001		0.0001
Root mean squared error	0.004		0.0071
Relative absolute error	0.1064 %		0.1024 %
Root relative squared error	2.5242 %		4.4419 %
Total Number of Instances	494020		167967

TABLE XII. THE WEIGHTED ACCURACY AVERAGE OF RANDOM FOREST ALGORITHM USING THE CROSS-VALIDATION TECHNIQUE

Weighted Avg. of	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
KDDCup99	1.000	0.000	NA	1.000	NA	NA	1.000	1.000
WSN Dataset	0.997	0.016	0.997	0.997	0.997	0.985	0.997	0.999

TABLE XIII. THE RESULTS OF RANDOM FOREST ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

	Dataset		
	KDDCUP99		*WSN*
Correctly Classified Instances	167915	99.969 %	167915
Incorrectly Classified Instances	52	0.031 %	52
Kappa statistic	0.9995		0.9995
Mean absolute error	0.0001		0.0001
Root mean squared error	0.0046		0.0046
Relative absolute error	0.1225 %		0.1225 %
Root relative squared error	2.8772 %		2.8772 %
Total Number of Instances	167967		167967

TABLE XIV. THE WEIGHTED ACCURACY AVERAGE OF RANDOM FOREST ALGORITHM USING THE PERCENTAGE SPLIT TECHNIQUE

Weighted Avg. of	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
KDDCup99	1.000	0.000	NA	1.000	NA	NA	1.000	1.000
WSN Dataset	0.997	0.015	0.997	0.997	0.997	0.985	0.997	0.999

V. DISCUSSION AND PERFORMANCE EVALUATION

In this section, all results will be discussed and analyzed. Overall performance evaluation will be presented as well. Naïve Bayes algorithm, using the cross-validation technique, has classified most of the instances correctly on both datasets. TP in KDDCup99 is about 92.9% and in WSN-DS is 95.3%. The errors or incorrectly classified instances were 7.08 and 4.064, respectively. Therefore, NAÏVE algorithm is more accurate with WSN dataset than KDDCup99 dataset. Moreover, the weighted accuracy average of both datasets is very similar. Using the percentage split technique with the former algorithm on both datasets showed more accurate results as compared with the cross-validation.

Moreover, IBK algorithm was run on both datasets using cross-validation. Both processes took no time at all, less than one second. As can be seen from the results of the correctly classified instances, both datasets were very close even though the number of instances in each dataset are not the same. The TP in KDDCUP99-DS is about (100%) and in WSN-DS is (99.4%). The errors or incorrectly classified instances were (0.552%) in WSN-DS and (0.0453%) in KDDCup99-DS. Whereas the correctly classified instances in WSN-DS is (99.4%) and in KDDCup99-DS is (99.9%) which is an excellent accuracy in both datasets, almost (100%). This is also reflecting on the weighted average of both datasets against the IBK algorithm. The numbers are very similar, almost the same (100%).

With the percentage split, using IBK algorithm was very accurate with KDDCup99 and WSN datasets. The errors or incorrectly classified instances were (0.584%) in WSN-DS and (0.058%) in KDDCup99-DS. And the correctly classified instances in WSN-DS is (99.4%) and in KDDCup99-DS is (99.9%) which is an excellent accuracy in both datasets, almost (100%). To sum up, IBK using percentage split test algorithm is very accurate with KDDCup99 dataset and with WSN dataset compared with the cross validation.

Furthermore, the Random Forest algorithm has been run on both datasets using the cross validation and percentage split options. With the cross validation, the TP in KDDCUP99-DS is about (100%) and in WSN-DS is (99.7%). The errors or incorrectly classified instances were (0.2779%) in WSN and (0.0213%) in KDDCup99. Also, the correctly classified instances in WSN-DS is (99.7%) and in KDDCup99-DS is (99.9%) which is an excellent accuracy in both datasets, almost (100%). For the percentage split, both datasets took few seconds to process (6.24 and 8.45 respectively). The TP in KDDCup99-DS is about (100%) and in WSN-DS is (99.7%). The errors or incorrectly classified instances were (0.2724%) in WSN-DS and (0.031%) in KDDCup99-DS. The correctly classified instances in WSN-DS is (99.7%) and in KDDCup99 is (99.9%) which is an excellent accuracy in both datasets, almost (100%). It can be concluded that Random Forest using percentage split test algorithm is very accurate with KDDCup99 dataset and with WSN dataset.

As an overall performance evaluation among all algorithms and test options for KDDCup99 dataset, the NAÏVE Bayes algorithm with cross-validation test option is the least accurate (92.92%), meaning it has the least correctly classified instances. On the other hand, the Random Forest algorithm with cross-validation test option (99.98%) was the most accurate. Similarly, for WSN dataset, the NAÏVE Bayes algorithm with cross-validation test option is the least accurate results (95.35%), meaning it has the least correctly classified instances and the Random Forest algorithm with cross-validation test option is the most accurate one (99.73%).

Moreover, the accuracy and processing time were recorded for both datasets using all test options as shown in Fig. 4, 5, 6 and 7. The least time taken was using the IBK algorithm using percentage split test option on WSN dataset (0.05) seconds, then with the KDDCup99 dataset algorithm using percentage split test option (0.08) seconds. As for accuracy measurement, the Random Forest algorithm is the most accurate algorithm in both datasets with all test options. The highest accuracy was registered using cross validation on KDDCup99 dataset (99.9787 %), then on WSN dataset (99.7276 %) using the percentage split test option as shown in Fig. 4 and Fig. 6, respectively.

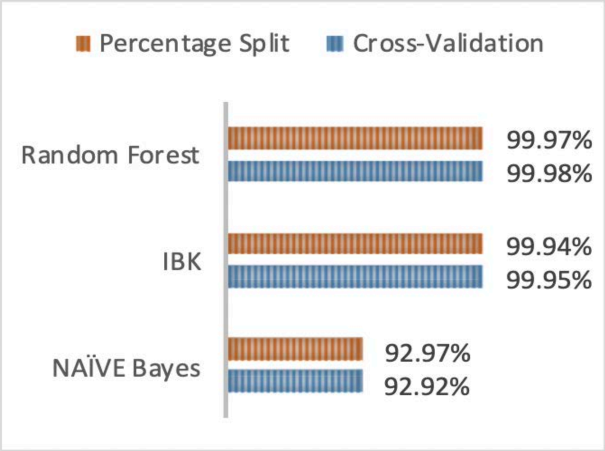


Fig. 4. Comparison of Accuracy on KDDCup99 Dataset.

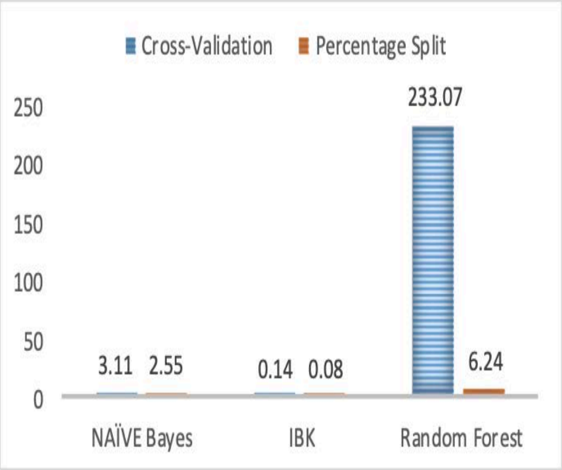


Fig. 5. Comparison of the Processing Time on KDDCup99 Dataset.

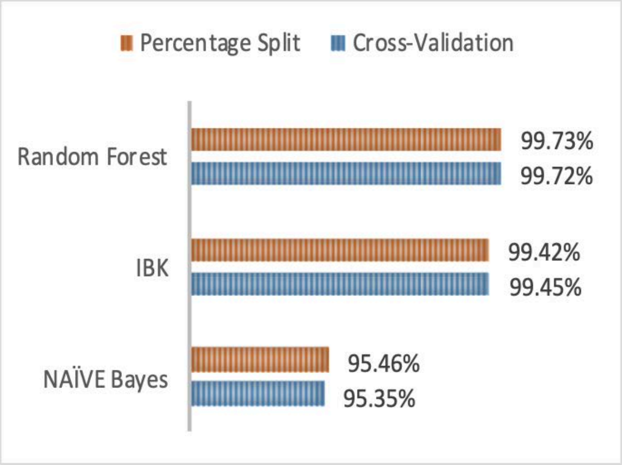


Fig. 6. Comparison of Accuracy on WSN Dataset.

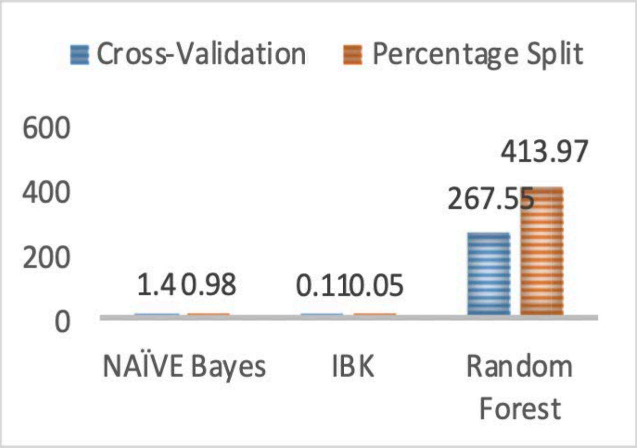


Fig. 7. Comparison of the Processing Time on WSN Dataset.

VI. Conclusion

Due to the importance of protecting WSN against rogue entities of hackers and intruders, taking into considerations all constraints such as limited power, storage, and processing capabilities, a model/dataset needs to be trained to mitigate new or modified attack types in networks.

This paper has analyzed and compared different machine learning algorithms against two datasets (WSN and KDD99) using WEKA tool. The purpose was to further assist in analyzing and protecting WSN networks in combination with Firewalls, Deep Packet Inspection (DPI), and Intrusion Prevention Systems (IPS) that are specialized in protecting WSN networks. Multiple testing options were investigated such as cross validation and percentage split. Based on the finding, the most accurate algorithm and the least time consuming were suggested for both datasets. Future research is needed to create more datasets to characterize various types of attacks in the wireless sensor networks.

References

[1] P. Kurer, D. M, and H. S. Guruprasad, “Energy Aware Dynamic Clustering and Hierarchical Route based on LEACH for WSN,” International Journal of Computer Networking Wireless and Mobile Communications, vol. 3, no. 3, pp. 79-86, 2013.

[2] A. I. Al-issa, M. Al-Akhras, M. ALsahli, and M. Alawairdhi, “*Using Machine Learning to Detect DoS Attacks in Wireless Sensor Networks,”* Paper presented at the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), April 2019.

[3] K. Hindi, “Fine tuning the Naïve Bayesian learning algorithm,” *AI Communications, vol.* 27, no. 2, pp. 133-141, 2014. doi: 10.3233/AIC-130588.

[4] M. A. Alsheikh, S. Lin, D. Niyato, and H. Tan, “Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications,” *IEEE Communications Surveys & Tutorials, vol. 16*, no. 4, pp. 1996-2018, 2014. doi: 10.1109/COMST.2014.2320099.

[5] B. Ashwini, S. Abhale, and S. Manivannan, “Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network,” *Optical Memory and Neural Network*s, vol. 29, no. 3, pp. 244-256, 2020. Available: https://link.springer.com/article/10.3103/S1060992X20030029#citeas.

[6] S. Gunduz, B., Arslan, and M. Demirci, “*A Review of Machine Learning Solutions to Denial-of-Services Attacks in Wireless Sensor Networks*,” 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015. Available: https://ieeexplore.ieee.org/document/7424301/authors#authors.

[7] A. Amouri, V. T. Alaparthy, and S. D. Morgera, “A Machine Learning Based Intrusion Detection System for Mobile Internet of Things,” *Sensors (Basel)*, vol. 20, no. 2, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31947567/.

[8] A. M. Al Tobi and I. Duncan, “KDD 1999 generation faults: a review and analysis,” *Journal of Cyber Security Technology*, vol. 2, no. 3-4, pp. 164-200., 2018. doi: 10.1080/23742917.2018.1518061.

[9] E. Kabir, J. Hu, H. Wang, and G. Zhuo, “A novel statistical technique for intrusion detection systems,” *Future Generation Computer Systems, vol.* 79, no. 1, pp. 303-318, 2018. doi: https://doi.org/10.1016/j.future.2017.01.029.

[10] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani,, “A detailed analysis of the KDD CUP 99 data set,” Paper presented at the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[11] C. Elkan, “Results of the KDD'99 classifier learning,” *SIGKDD Explor. Newsl., vol. 1*, no. 2, pp. 63–64, 2000. doi: 10.1145/846183.846199.

[12] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, “WSN-DS: A Dataset for Intrusion Detection Systems in Wireless Sensor Networks.,” *Journal of Sensors*, 2016. doi: 10.1155/2016/4731953.

[13] I. Almomani and B. Al-Kasasbeh, “*Performance analysis of LEACH protocol under Denial of Service attacks,”* Paper presented at the 2015 6th International Conference on Information and Communication Systems (ICICS), 2015.

[14] S. L. Ting, W. H. Ip, and A. Tsang, “Is Naïve Bayes a Good Classifier for Document Classification?” *International Journal of Software Engineering and its Applications*, *vol.* 5, no. 3, pp. 37-46., 2011.

[15] D. Meretakis and B. Wüthrich, “*Extending naïve Bayes classifiers using long itemsets,”* Paper presented at the KDD '99, 1999.

[16] I. Androutsopoulos, J. Koutsias, K. Chandrinos, and C. Spyropoulos, “*An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages*,” Proceedings of the 23rd annual international ACM SIGIR conference on Research and development, 2000. doi: 10.1145/345508.345569.

[17] K. El-Hindi and M. Al-Akhras, “Smoothing Decision Boundaries to Avoid Overfitting in Neural Network Training,” *Neural Network World, vol. 21*, no. 4, pp. 311-325, 2011.

[18] I. I. Baskin, G. Marcou, D. Horvath, and A. Varnek, “*Classification Models,” Tutorials in Chemoinformatics*, Varnek, A. (Ed.), 2017.

[19] F. Syeda, M. A. B. Mirza, A. Baig, and M. Pawar, “Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis,” *IOSR Journal of Computer Engineering, vol.* 10, no. 6, pp. 1-6, 2013. doi: 10.9790/0661-1060106.

[20] T. Borovicka, M. Jirina, and P. Kordík, “Selecting Representative Data Sets,” *Advances in data mining knowledge discovery and applications*, pp. 43-70, 2012.

Improving Packet Delivery Ratio in Wireless Sensor Network with Multi Factor Strategies

Venkateswara Rao M[1]
Research Scholar
Department of Computer Science and Engineering
KoneruLakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Srinivas Malladi[2]
Professor
Department of Computer Science and Engineering
KoneruLakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

***Abstract*—In the design of wireless sensor network (WSN), packet delivery ratio is an import parameter to be maximized. In existing schemes, a secure zone-based routing protocol was implemented for life time improvement in WSNs. In multi - hop communication, a new routing criterion was formulated for packet transmission. Security against message tampering, dropping and flooding attacks was incorporated in the routing metric. The approach skipped risky zones as a whole from routing and chooses alternative path to route a packet in secured manner with less energy consumption. Though energy conservation and attack resilience are achieved, congestion in WSN is increased and because of it packet delivery ratio is diminished. To address this problem, we propose a solution to improve the packet delivery ratio with a multi factor strategies involving routing, differentiation of flows, flow-based congestion control with retransmission and redundant packet coding. Detailed analysis and simulations are undertaken to evaluate the efficiency of the contemplated work compared to the existing solutions.**

Keywords*—*Multi factor strategies; novel routing metric; packet coding; packet delivery ratio

I. Introduction

Wireless sensor network is the collection of sensor nodes which has the capability of wireless data communication. Sensor nodes sense the environment variables depending on the application requirements and send them to sink either directly or through multi hop transmission.

A. *Motivation*

WSN is being increasingly used in many applications like Military ,precision farming, industrial safety, smart home etc. which needs reliable and high speed data transmission . It is required to design WSN applications by considering the requirements like more packet delivery ratio, life time and security.

B. *Earlier Research Work*

In earlier works [1], a zone-based sensor network is designed with the consideration of life time and security. The sensor network is secured against message tampering, dropping and flooding attacks. WSN Life time is increased by optimizing the power usage and packet routing. The network is split to zones and each zone is assigned a score. This score is calculated depending on past security lapses and present residual energy in that zone. A zone with higher score is preferred for routing. In zone based attack resilient routing the network life span and attack resilient capability are improved. But data packets are lost due to higher congestion in WSN. The solution in this work minimizes the packet delivery ratio owing to a significant raise in total numbers hops and node overhead in the network.

The major focus of earlier solutions [1,5,6] is finding the optimal route for data transmission .The identified gaps in the previous works are lower packet delivery ratio, less network life time ,lower throughput. These gaps are addressed in the proposed multi factored approach.

C. *Proposed Work*

Multi factored Secure Routing approach is proposed to improve packet delivery ratio in WSN by addressing the draw backs related to energy efficient attack resilient routing.

In the proposed Multifactor approach, packet delivery ratio is improved by selecting energy efficient route, Differentiation of flows in the network and flow based congestion control and Redundancy & Retransmission management,

The entire WSN is divided in to zones. Energy efficient path from source to sink is selected depending on the preference score calculated at each neighboring zone. The preference score (PS) is measured depending on security score, energy score and reliability score available in the zone. The zone heads are sorted depending on the distance from sink node. Among the neighbouring head nodes, the node whose PS value above threshold $T_1$is selected to send data packet.

In the differentiation of flows Packet flow is split into two categories of high and low priority in the network. At each routing hop, the flow of packets is controlled differently for the packets based on high and low priorities. Packets with high priority are transmitted and low priority packets are buffered during network congestion period.

The low priority packets are buffered to reduce network congestion. This allows for reliable transmission of packets. When a node receives a Packet Forward Success (PFS) message, it calculates the Round Trip Time (RTT). When RTT is above the configured threshold, data flow rate is reduced, when RTT is below the configured threshold, data flow rate is increased.

In Redundancy and Retransmission management data content sent from source is divided into n packets using Reed Solomon erasure coding approach. Even if some packets are lost during transmission, the whole message is reconstructed at the destination without retransmission.

Packet delivery ratio is increased by introducing a weighted routing metric with the consideration of various factors. The goal is to increase the packet delivery ratio applying multiple factors like flow management, retransmission control and weighted routing metric to zone based sensor network.

In the following, Section II presents the related review for proposed solution. Section III elaborates the proposed Multi Factored Strategy, Section IV presents salient features in proposed solution, Results related to proposed schemes are given in Section V. Lastly, section VI explores the conclusions and future scope of the proposed solution.

II. Related Work

An opportunistic routing scheme is presented in [2] to increase the packet delivery ratio in WSN. Relay nodes are selected based on minimization of retransmission. Due to the reduction of retransmission, the overall throughput of the network is increased and thereby reducing the packet loss. Security is not considered for end to end delivery .An opportunistic routing scheme similar to [2] is proposed in [3] with the aim of increasing the reliability and reducing the latency. The best relay satisfying the QoS criteria is selected based on reliability and time guarantee. Less network life time is the limitation for this work. Authors in [4] proposed a distributed multi path algorithm with the goal of higher packet reliability. The path selection is adapted to network changes and failures. With the availability of multiple base stations and routing on multiple paths to these stations, fidelity of packet is improved in this work. Nevertheless, the major limitations are increased network overhead and reduced throughput. An optimal energy efficient routing protocol is formulated in [7].improved Packet delivery ratio and Network Life time is addressed in this solution. A stateless position-dependent routing technique is formulated in [8].Stateless routing is realized using greedy perimeter routing. Geographic routing along with greedy data transmission is able to increase the packet delivery time. Detection of malicious nodes in sensor network is explored in [9, 10]. Network life time is increased by enhanced energy efficient clustering of sensor network in [11]. LEACH algorithm is improved to select the best cluster head to enhance the lifespan of the network. Throughput is less in this work. Ant colony-based routing algorithm is proposed in [12] with the goal of higher packet delivery ratio. The ant colony algorithm uses the fitness function optimized based on communication distance, transmission direction and current residual energy. Packet delivery ratio is increased using opportunistic multipath routing in [13]. Similar to it, in [14] A multi path routing algorithm is formulated to increase packet reliability by forwarding on multiple paths. The solution also uses aggregation to diminish the number of efficient transmissions in WSN.Packet prioritization and optimized back-off MAC protocol is proposed in [15]. Due to reduction of collision in the optimized back-off MAC protocol, packet reliability is improved. The optimized back off MAC works by assigning back off time depending on the packet priorities.A virtual node concept is introduced in [16] to enhance the QoS in WSN based on clustering technique. The scheduling is realized using TDMA in this network. High priority is assigned to certain designated nodes. During higher traffic load, the packets from these designed nodes are given more importance in routing so that latency for those packets is reduced. A stateless opportunistic routing protocol is defined in [17] with the objective of enhancing the packet reliability. Forwarding area for packet is adapted dynamically depending on the density of the sensor nodes, so that reliability is increased. Increased overhead is not addressed. Compressive sensing is used to reduce the packet overhead in [18]. Compressive sensing improves the network utilization and packet delivery ratio. Compressive sensing also minimizes the power usage in WSN.Packet reliability is increased using energy aware Quality of Service (QoS) transmission protocol depending on various parameters in [19]. Energy consumption is reduced by optimal positioning of nodes in [20]. Packet low delivery ratio is limitation in this work. Clustering based energy minimization strategies are explored in [23, 28, 21]. Attack resilient routing in wireless sensor network in discussed in [22, 24, 25, 26, 27].

III. Proposed Multi Factored Strategy

The architectural representation of proposed multi factored approach is depicted in Fig. 1. The architecture represents four important concepts to achieve more packet delivery ratio and life span.

- Selection of energy efficient routing path.
- Redundancy and Retransmission management.
- Differentiation of flows in the network.
- Flow based congestion control.

A. Packet Routing

The Whole area of WSN is split into zones of N×N size. With one hop routing, any node in the zone will connect to any other node. Node close to centre of the zone that can observe all packet transmissions is selected as zone head. The information of nodes within each zone is present in the sink. Three distinct scores, security score, energy score and reliability score are attained in each zone. The value of the scores varies from 0 to 10. The zone with high score is recommended for routing when compared with zone having lower score. The zone scores are calculated based on fuzzy function using the following input variables.

- Packet traversing count (PTC).
- Packet traversing failure count (PTFC).
- Tampered incidence count (TIC).
- No tampered count (NTC).

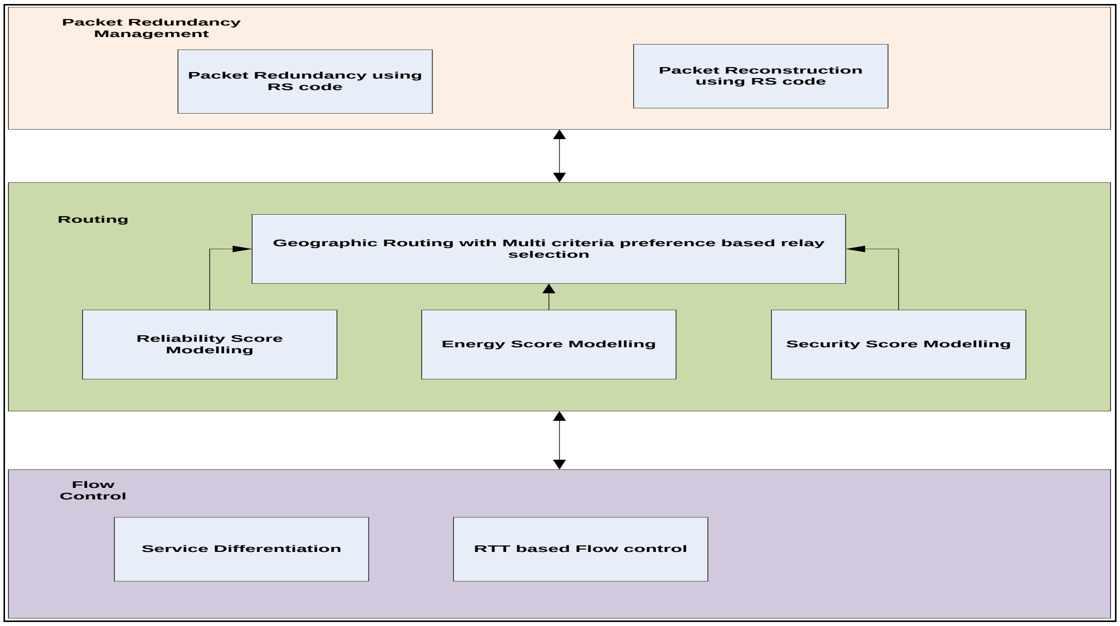

Packet Redundancy Management
Packet Redundancy using RS code
Packet Reconstruction using RS code
Routing
Geographic Routing with Multi criteria preference based relay selection
Reliability Score Modelling
Energy Score Modelling
Security Score Modelling
Flow Control
Service Differentiation
RTT based Flow control

Fig. 1. Architecture of Secure Routing based on Multi Factored Approach.

These scores are detailed in [1] and the counters are kept at sink node. But these counters are kept at zone head in the proposed solution. The fuzzy function for calculating the security score (SS) of the zone is shown in (1).

$$SS = \mu_1 * Q(PTC) + \mu_2 * Q(PTFC) + \mu_3 * Q(TIC) + \mu_4 * Q(NTC) \quad (1)$$

Where

$Q(x)$: The fuzzification kernel of input x.

The de-fuzzification score is measured by using center of gravity formula as shown in (2).

$$Score = \frac{\int \mu_{\bar{D}r}(x).xdx}{\int \mu_{\bar{D}r}(x).dx} \quad (2)$$

Where x = {PTC, PTFC, TIC, NTC}.

The path security score SS_p (3) is measured as.

$$SS_P = \frac{\sum_{i=1}^{N} SS_i}{N} \quad (3)$$

Where SS_i is the zone security score. There are N zones in the path.

The energy score (ES) of a zone is measured as in (4).

$$ES = \frac{10*(E - TPC * E_c)}{E} \quad (4)$$

Where E is the node's initial energy and E_c is the energy utilized for packet transmission at node. TPC is the total packets transmitted.

The path energy score ES_P is calculated using (5).

$$ES_p = \prod_{\min of\ all\ N} ES_i \quad (5)$$

Where ES_i is the energy score of the path.

Rather than the revised AODV based routing addressed in [1] ,the packets are transmitted using a geographical routing through preference score-dependent path selection. Past packet forward statistics are used to calculate the reliability score for the zone. Packet forward success message is sent by each node with successful forward for packet to immediate hop. Whenever this message is received, packet forward success counter (PFS) is incremented. Periodically, the reliability score (6) is measured as.

$$R_t = \alpha \times R_{t-1} + (1 - \alpha)\frac{PFS}{TPC} \quad (6)$$

With $R_0 = 0$ and α is constant.

Initially the data packet is sent from the source node to the zone head. The head of zone sends a packet containing hello message to other neighbouring heads. The heads of zones receiving 'HELLO' message calculate the preference score and send back a 'HELLO_RES' as response to 'HELLO'. The preference score (PS) is measured (7) as the sum of weighted values of security score, energy score and reliability score.

$$PS = w_1 * SS_p + w_2 * ES_p + w_3 * R_t \quad (7)$$

With $w_1 + w_2 + w_3 = 1$ and $w_3 > w_2 > w_1$

The zone accepting 'HELLO_RES' , selects the zone head with highest PS score as relay hop and route the DATA packet to that zone head.

After the reception of HELLO_RES by zone head, the K nearest neighbour heads are sorted depending on the distance from sink node. Among K neighbouring head nodes, the node whose PS value above threshold $T_1$is selected to send data packet. Data packet is forwarded to all K neighbour nodes, if none of the neighbour has PS values aboveT_1. This process is repeated at every hop until data packet received by the sink node.

Before applying fuzzy function on the input variables PTC, PTFC, TIC, NTC , these variables must converted from numerical to categorical values of Low (L), Medium (M) and High (H) using the transformation functions . Transformation function for the input variable PTC is shown in the Fig. 2.

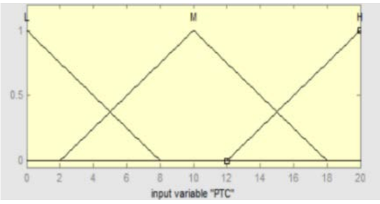


Fig. 2. PTC Transformation Function.

Transformation function for the input variable PTFC is shown in the Fig. 3.

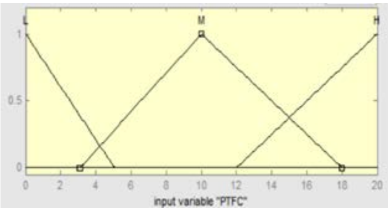


Fig. 3. PTFC Transformation Function.

Transformation function for the input variable TIC is shown in the Fig. 4.

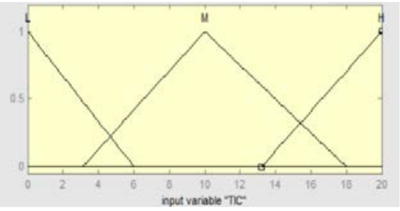


Fig. 4. TIC Transformation Function.

Transformation function for the input variable NTC is shown in the Fig. 5.

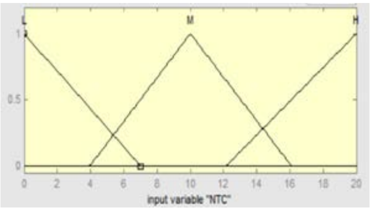


Fig. 5. NTC Transformation Function.

The output variable of preference score is converted from numerical to categorical value using the transfer function given in Fig. 6.

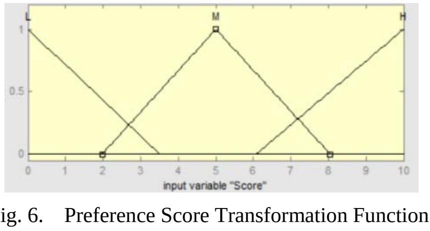


Fig. 6. Preference Score Transformation Function.

The input variables PTC, PTFC, TIC and NTC are mapped with output variable of preference score using fuzzy rule base.

B. Flow Differentiation and Congestion Control

Packet flow is split into two categories of high and low priority in the network. At each routing hop, the flow of packets is controlled differently for the packets based on high and low priorities. Packets with high priority are transmitted and low priority packets are buffered during network congestion period. Due to buffering of low priority packets, network congestion reduces. This allows for reliable transmission of packets. The round trip time (RTT) is calculated, when a node receives a packet forward success response .Based on probability delay, RTT (8) is measured as below.

$$RTT = \begin{cases} \sum_{i=0}^{\infty} f_i(a).f_i(b) \ , x = 0 \\ \sum_{i=0}^{\infty} f_i(a).f_{2x+i}(b) + \sum_{i=0}^{\infty} f_i(b).f_{2x+i}(a), x > 0 \end{cases} \quad (8)$$

Where: a is forward path.

b is backward path.

f is the probability mass function..

Adaptive flow control for the packets is shown in Fig. 7.

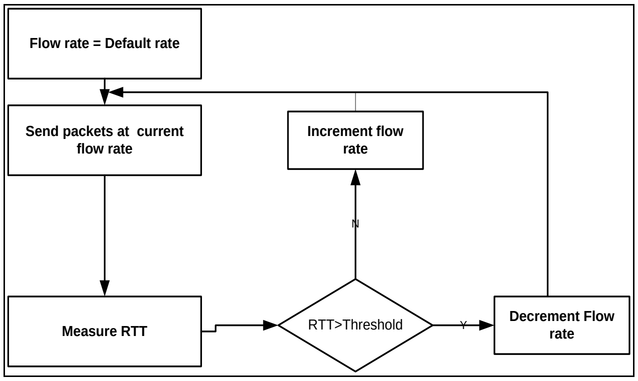


Fig. 7. Flow of Packets.

The flow rate is initially set to the default value and data flow control at each node begins with this value [Fig. 7]. Based on the reception of packet forward success message, round trip delay is calculated. Data Flow rate is decreased when RTT is above configured threshold and increased when RTT is below configured threshold.

C. Redundant Packet Coding

The message content sent from source is divided into n packets using erasure coding (n,k) .Only k of n packets are sufficient for reconstructing the entire message content. The source node sends n packets of the message and if k out of n packets is received at sink, the message can be reconstructed without any necessity for retransmission. This work uses Reed Solomon erasure coding. Use of Reed Solomon code has following advantages.

- The reed Solomon transformed contents are encrypted and secure.
- Retransmission can be avoided to reconstruct the entire message with lesser number of packets.

IV. Salient Features in Proposed Solution

The proposed multi factor strategy has following salient features:

- Packet delivery ratio and network life span are less in earlier works [1],[5],[6] which are addressed in the proposed solution using multi factored approach.
- Risky areas in the network are quantified using scores and routing is adapted to skip those risky areas.
- Packet flow is differentiated and flow is managed to reduce the congestion in the network. Flow management is dynamic to congestion in the network.
- Retransmission is avoided in the network due to erasure coding.
- Attack resiliency is assured with proactive routing in secure paths.

V. Results

The proposed multi factored strategy is simulated in NS2 platform. The simulation parameters used for testing the proposed solution is given in TABLE I.

The performance of the proposed work approach is simulated and evaluated by comparing with the following solutions.

- Secure Energy Efficient Routing is contemplated in [1].
- Optimal Energy Efficient Routing is formulated in [5].
- Secure Localized Routing is defined and implemented in [6].

Performance evaluation parameters considered for comparison are: packet delivery ratio, packet delay, node overhead, network life time and throughput.

Packet delivery ratio (9) measures the ratio of successfully received packets at sink to the total number of packets sent by the sources. It is calculated as

$$PDR = \frac{number\ of\ packets\ recived\ at\ sink}{total\ number\ of\ packets\ sent} \tag{9}$$

The results of packet delivery ratio against number of nodes are given in Table II and Fig. 8.

The packet delivery ratio of the proposed solution is 6.95% more Compared to [1], 14.6% more compared to [5] and 16.51% more compared to [6].

The results of average delay against number of nodes are given in Table III and Fig. 9.

Compared to [1], average delay is lower by 15.38% in the proposed solution, 24.13% lower compared to [5], 37.14% lower compared to [6]. The delay is lower in the proposed solution due to reduction in number of hops in the geographic routing strategy adopted in the proposed work.

The result of network overhead against number of nodes is shown in Table IV and Fig. 10.

TABLE I. Simulation Parameters

Criterion	**Value(s)**
No. of Nodes	50 - 250
Communication range(m)	100
Simulation expanse(m^2)	1000 × 1000
Allocation of Priority (%)	20
Disposition of sensor Node	Random
Time of Simulation	30
Queue size of Interface	50
Medium Access Control(MAC)	802.11

TABLE II. Comparison of Packet Delivery Ratio

No.of Nodes	**Proposed Multi Factored Approach**	**Secure Energy Efficient Routing[1]**	**Optimal Energy Efficient Routing[5]**	**Secure Localized Routing[6]**
50	93	87	82	81
100	92	86	81.5	79.5
150	91	85.2	79.22	78
200	90.5	84	78	76.9
250	89	83.2	76.8	75.1

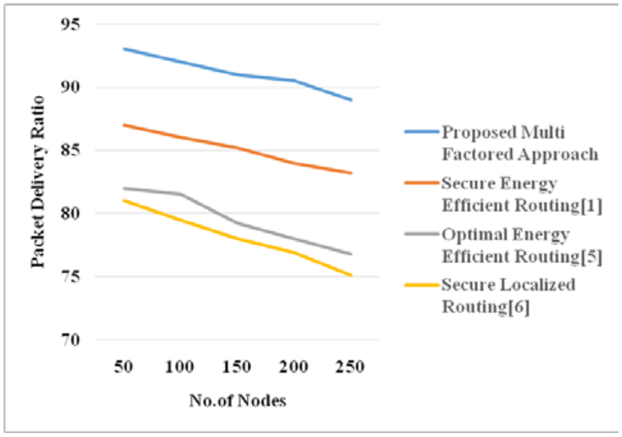


Fig. 8. Packet Delivery Ratio.

TABLE III. Comparison of Delay

No.of Nodes	**Proposed Multi Factored Approach**	**Secure Energy Efficient Routing[1]**	**Optimal Energy Efficient Routing[5]**	**Secure Localized Routing[6]**
50	15	18	21	24
100	14	16	18	22
150	13	15	17	21
200	12	15	16	20
250	12	14	15	18

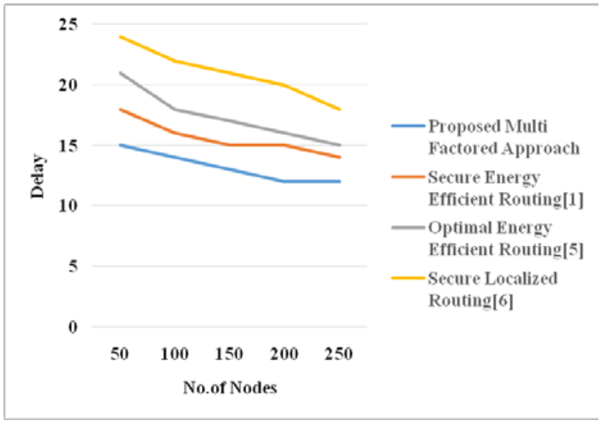


Fig. 9. Delay Comparison.

TABLE IV. ANALYSIS OF OVERHEAD

No.of Nodes	Proposed Multi Factored Approach	Secure Energy Efficient Routing[1]	Optimal Energy Efficient Routing[5]	Secure Localized Routing[6]
50	72	80	82	85
100	85	92	93	97
150	89	97	100	106
200	94	104	108	116
250	97	116	118	127

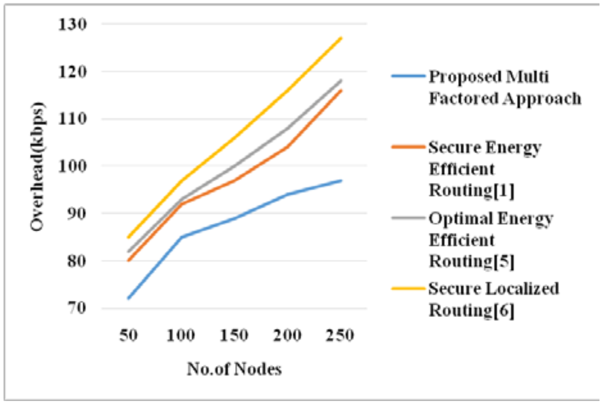


Fig. 10. Analysis of Overhead.

Compared to [1], overhead is lower by 10.63% in proposed solution, 12.77% lower than [5] and 17.7% lower than [6]. Adaptive multi path propagation and geographic routing has reduced the network overhead in proposed solution.

The result of life time against number of nodes is given in Fig.11 and Table V.

Compared to [1], life time is higher by 19.4% in the proposed solution, 72.7% more compared to [5] and 102.1% more compared to [6]. Due to the minimization of number of hops and retransmissions, energy consumption is reduced and this has increased the life span in the proposed work.

The results of network throughput for different rate of packet generation are given in Table VI and Fig. 12.

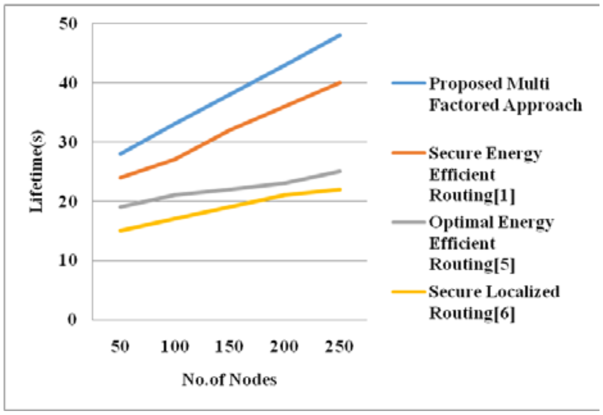


Fig. 11. Life Time Comparison.

TABLE V. LIFE TIME COMPARISON

No.of Nodes	Proposed Multi Factored Approach	Secure Energy Efficient Routing[1]	Optimal Energy Efficient Routing[5]	Secure Localized Routing[6]
50	28	24	19	15
100	33	27	21	17
150	38	32	22	19
200	43	36	23	21
250	48	40	25	22

TABLE VI. THROUGHPUT COMPARISON

Total Sending Rate (kbps)	Proposed Multi Factored Approach	Secure Energy Efficient Routing[1]	Optimal Energy Efficient Routing[5]	Secure Localized Routing[6]
100	92	84	83	81
150	140	136	129	125
200	181	173	167	160
250	224	208	201	195
300	267	251	242	238

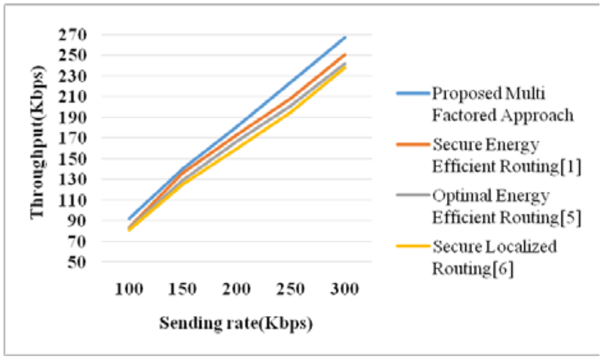


Fig. 12. Throughput Comparison.

The proposed solution has higher throughout in terms of 6.1% more than [1], 9.97% more than [5] and 13.14% more than [6].

The split ratio between high and low priority packet is set to 70:30 and throughput is measured for varied rate of packets from source. The comparison is given in Table VII and Fig. 13.

TABLE VII. COMPARISON OF THROUGHPUT ON 70:30 SPLIT

Total Sending rate (kbps)	Proposed Multi Factored Approach	Secure Energy Efficient Routing[1]	Optimal Energy Efficient Routing[5]	Secure Localized Routing[6]
100	65	52	50	47
150	100	83	82	79
200	130	100	96	90
250	160	120	117	110
300	195	160	157	147

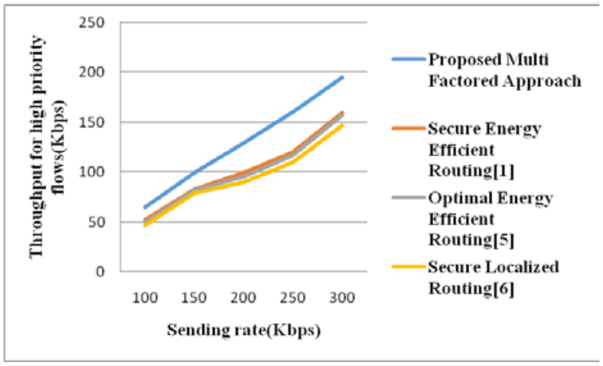


Fig. 13. Throughput Comparison on 70:30 Split.

The proposed solution has higher throughput in terms of 9.97% greater than [1], 29.48% greater compared to [5] and 47.42% greater compared to [6].

VI. CONCLUSION AND FUTURE SCOPE

A multi factored strategy with the objective of providing increased packet delivery ratio in zone-based sensor network is proposed in this work. Geographic routing is adapted with relay selection based on a preference score. Adaptive multi path propagation, flow differentiation and congestion control, redundancy coding are the multi factored strategies proposed in this work. Redundancy in the coding has reduced the number of retransmissions in the network. Flow differentiation and flow control has reduced the congestion in the WSN. Preference score is calculated based on energy availability, security and reliability of the nodes and use of it in routing has increased the packet reliability. Due to multi factored strategy, the packet delivery ratio improved by 6.95% in the proposed work, life time improved by 19.4% and delay reduced by 15.38% compared to the existing works. These results are very much useful for further research in the areas of Military applications, Environmental monitoring etc. In the further enhancement, the proposed solution can be improved to reduce the average energy consumption of nodes within the zone in order to improve packet delivery ratio.

REFERENCES

[1] Venkateswara Rao M and Srinivas Malladi, “Secure Energy Efficient Attack Resilient Routing Technique for Zone based Wireless Sensor Network” International Journal of Advanced Computer Science and Applications (IJACSA), 11(12), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111267.

[2] Hasnain, Muhammad, Mazhar Hussain Malik, and Mehmet Emin Aydin. "An adaptive opportunistic routing scheme for reliable data delivery in WSNs." Proceedings of the 2nd International Conference on Future Networks and Distributed Systems. 2018.

[3] Chen, Aiguo, et al. "RTGOR: Reliability and Timeliness Guaranteed Opportunistic Routing in wireless sensor networks." EURASIP Journal on Wireless Communications and Networking 2018.1 (2018): 86.

[4] Velasquez-Villada, Carlos, and YezidDonoso. "Multipath routing network management protocol for resilient and energy efficient wireless sensor networks." Procedia Computer Science 17 (2013): 387-394.

[5] Khamayseh, Yaser M., Shadi A. Aljawarneh, and Alaa Ebrahim Asaad. "Ensuring survivability against Black Hole Attacks in MANETS for preserving energy efficiency." Sustainable Computing: Informatics and Systems 18 (2018): 90-100.

[6] Poongodi, Thangamuthu, and M. Karthikeyan. "Localized secure routing architecture against cooperative black hole attack in mobile ad hoc networks." Wireless Personal Communications 90.2 (2016): 1039-1050.

[7] Prasad, A. Y., and Balakrishna Rayanki. "A generic algorithmic protocol approaches to improve network life time and energy efficient using combined genetic algorithm with simulated annealing in MANET." International Journal of Intelligent Unmanned Systems (2019).

[8] Yan Sun, Junpeng Guo, "Speed Up Greedy Perimeter Stateless Routing Protocol for Wireless Sensor Networks (SU-GPSR)", IEEE,2017.

[9] Vamshi krishna, H., & Swain, G. “Identification and avoidance of malicious nodes by using certificate revocation method.” International Journal of EngineeringandTechnology(UAE), 7(4.7 Special Issue 7) (2018)., 152-156.

[10] Kalaipriyan, T., et al. "Monkey King Algorithm for Solving Minimum Energy Broadcast in Wireless Sensor Network." Advances and Applications in Mathematical Sciences 17.1 (2017): 129-145.

[11] Prasad, A. Y., and R. Balakrishna. "Implementation of optimal solution for network lifetime and energy consumption metrics using improved energy efficient LEACH protocol in MANET." Telkomnika 17.4 (2019): 1758-1766.

[12] Sun, Yongjun, Wenxin Dong, and Yahuan Chen. "An improved routing algorithm based on ant colony optimization in wireless sensor networks." IEEE communications Letters 21.6 (2017): 1317-1320.

[13] Kim, Sangdae, et al. "Opportunistic Multipath Routing in Long-Hop Wireless Sensor Networks." Sensors 19.19 (2019): 4072.

[14] Dong, Mianxiong, Kaoru Ota, and Anfeng Liu. "RMER: Reliable and energy-efficient data collection for large-scale wireless sensor networks." IEEE Internet of Things Journal 3.4 (2016): 511-519.

[15] Onwuegbuzie, Innocent Uzougbo, et al. "Optimized backoff scheme for prioritized data in wireless sensor networks: A class of service approach." PloS one 15.8 (2020): e0237154.

[16] Almobaideen, Wesam, Mohammad Qatawneh, and OriebAbuAlghanam. "Virtual node schedule for supporting QoS in wireless sensor network." 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE, 2019.

[17] Ghoreyshi, Seyed Mohammad, Alireza Shahrabi, and TuleenBoutaleb. "A stateless opportunistic routing protocol for underwater sensor networks." Wireless Communications and Mobile Computing 2018 (2018).

[18] Li, Yimei, and Yao Liang. "Compressed sensing in multi-hop large-scale wireless sensor networks based on routing topology tomography." IEEE Access 6 (2018): 27637-27650.

[19] Monowar, Muhammad Mostafa. "An Energy-aware Multi-constrained Localized QoS Routing for Industrial Wireless Sensor Networks." Adhoc& Sensor Wireless Networks 36 (2017).

[20] Yan, Ziwei, et al. "Energy-efficient node positioning in optical wireless sensor networks." Optik 178 (2019): 461-466.

[21] Karthikeyan, T., V. Brindha, and P. Manimegalai. "Investigation on Maximizing Packet Delivery Rate in WSN Using Cluster Approach." Wireless Personal Communications 103.4 (2018): 3025-3039.

[22] Chowdary, Krishna, and K. V. V. Satyanarayana. "Malicious Node Detection and Reconstruction of Network in Sensor Actor Network." Journal of Theoretical & Applied Information Technology 95.3 (2017).

[23] Mallikarjuna Rao, Y., M. V. Subramanyam, and K. Satya Prasad. "Cluster-based mobility management algorithms for wireless mesh networks." International Journal of Communication Systems 31.11 (2018): e3595.

[24] Kadiravan, G., Pothula Sujatha, and J. Amudhavel. "A State of Art Approaches on Energy Efficient Routing Protocols in Mobile Wireless Sensor Networks." IIOAB Journal 8.2 (2017): 234-238. www. https://www.iioab.org/.

[25] Chowdary ,K., & Satyanarayana, K. V. V. “A novel secured data transmission and authentication technique against malicious attacks in WSNs.” Journal of Advanced Research in Dynamical and Control Systems, (Special Issue -18) (2017), 161-173.

[26] Arage Chetan, S., & Satyanarayana, K. V. V. “Novel routing protocol for secure data transmission in wireless ad hoc networks.” International Journal of Innovative Technology and Exploring Engineering, 8(4S2) (2019)., 101-108.

[27] Manjunath, B. E., and P. V. Rao. "Trends of Recent Secure Communication System and its Effectiveness in Wireless Sensor Network." Journal of Innovation in Electronics and Communication Engineering 6.2 (2016): 46-52.

[28] Gummadi, Annapurna, and K. Raghava Rao. "EECLA: Clustering And Localization Techniques To Improve Energy Efficient Routing In Wireless Sensor Networks." Journal of Theoretical & Applied Information Technology 96.1 (2018).

UML Sequence Diagram: An Alternative Model

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

***Abstract*—The UML sequence diagram is the second most common UML diagram that represents how objects interact and exchange messages over time. Sequence diagrams show how events or activities in a use case are mapped into operations of object classes in the class diagram. The general acceptance of sequence diagrams can be attributed to their relatively intuitive nature and ability to describe partial behaviors (as opposed to such diagrams as state charts). However, studies have shown that over 80% of graduating students were unable to create a software design or even a partial design, and many students had no idea how sequence diagrams were constrained by other models. Many students exhibited difficulties in identifying valid interacting objects and constructing messages with appropriate arguments. Additionally, according to authorities, even though many different semantics have been proposed for sequence diagrams (e.g., translations to state machines), there exists no suitable semantic basis refinement of required sequence diagram behavior because direct style semantics do not precisely capture required sequence diagram behaviors; translations to other formalisms disregard essential features of sequence diagrams such as guard conditions and critical regions. This paper proposes an alternative to sequence diagrams, a generalized model that provides further understanding of sequence diagrams to assimilate them into a new modeling language called thinging machine (TM). The sequence diagram is extended horizontally by removing the superficial vertical-only dimensional limitation of expansion to preserve the logical chronology of events. TM diagramming is spread nonlinearly in terms of actions. Events and their chronology are constructed on a second plane of description that is superimposed on the initial static description. The result is a more refined representation that would simplify the modeling process. This is demonstrated through remodeling sequence diagram cases from the literature.**

Keywords*—Requirements elicitation; conceptual modeling; static model; events model; behavioral model

I. Introduction

In object-oriented analysis, we identify classes by examining usage scenarios, where classes are determined through nouns or noun phrases [1, 2]. This is followed by analyzing classes with the intent of encapsulation (bundling data and methods) while still keeping data and operations separate. Then, the analysis moves to the task of specifying operations, which define the *behavior* of objects, including the communication that occurs between objects by passing messages to one another. In this phase, requirements might demand examination of how an application behaves as a consequence of external events. A behavioral model indicates how software will respond to external events. According to [2], in general, an *event* occurs whenever the system and actor exchange information; in this case, the event is the fact that information has been exchanged. The creation of the behavioral model necessitates the following steps [2]:

1) Evaluate all use cases to fully understand the sequence of interaction within the system

2) Identify events that drive the interaction sequence and understand how these events relate to specific objects.

3) Create a sequence for each use case.

B. Sequence Diagrams

The UML behavioral representation, the sequence diagram (SD), is the second most common UML diagram that represents how objects interact and exchange messages over time [3, 4]. SDs have been used informally for several decades [5]. The first standardization of SDs came in 1992, and since then, there have been several dialects and variations. SDs show how messages are sent between objects or other instances to perform a task. They are used during the detailed design phase, in which precise interprocess communication must be established according to formal protocols. When testing is performed, the behavior of the system can be described as SDs [5].

Shen [6] observed that the general acceptance of SDs can be attributed to their relatively intuitive nature and ability to describe partial behaviors (as opposed to such diagrams as state charts). SDs plays an important role in helping users understand system operation and visualize the interactions among a system's objects. According to [7], "the reason of their success compared to other formalisms like state machines is that they are easy to use and understand." The SD bridges the UML use case model and the object classes specified in the structural model. In UML, diagrams must be tightly integrated to avoid inconsistencies; however, such tight integration is infeasible and often impractical [8]. UML 2 extended SDs with such features as recent enrichment and new symbols that allowed programmers to indicate additional procedural details [3, 6].

SDs are used for different purposes, such as showing the flows of method calls inside a program or giving a partial specification of interactions in a distributed system [9]. SDs are utilized to automate test case generation [10]. SDs show "how events or activities in a use case are mapped into operations of object classes in the class diagram. Events are basic behavioral constructs of SDs that can be combined to form larger behavioral constructs called fragments" [11].

C. Problems

Many different semantics have been proposed for SDs for various purposes (e.g., in terms of translations to state machines) [9, 11]. In short, there does not exist a suitable semantics-based refinement of required SD behavior because direct style semantics do not precisely capture required SD behaviors, and translations to other formalisms disregard essential features of SDs such as guard conditions and critical regions [11]. It is not easy to select suitable semantics: the various formal semantics for SDs handle even the most basic diagrams quite differently [9]. Thus, there is a need for a refinement of SD behaviors that clarifies the issue of the relationship of the static and dynamic features.

Modeling complex interaction behaviors relies on a good understanding of SDs [11]. However, SDs often pose the greatest difficulties among novices learning modeling [11, 12]. Modeling SDs overwhelms some learners, as it involves a large number of interacting items that must be handled concurrently [13].

Over 80% of graduating students are unable to create a software design or even a partial design [14]. According to [14], “many students had no idea how SDs were constrained by other models. Many exhibited difficulties in identifying valid interacting objects and constructing messages with appropriate arguments. Though students understood the role of objects, messages and arguments individually, they were daunted when considering all constraints imposed by other models, concurrently.”

D. Generalizing SDs

This paper proposes an alternative generalized model to SDs that provides further understanding of SDs, assimilating them into a new modeling language called a thinging machine (TM). The SD is extended horizontally, removing the superficial dimensional limitation of vertical-only expansion to preserve the logical chronology of events. TM diagramming is spread nonlinearly in terms of actions. Events and their chronology are constructed on a second plane of description that is superimposed on the initial static description. The result is a more refined representation that simplifies the modeling process. This is demonstrated by remodeling SD cases from the literature.

For the sake of having a self-contained paper, the next section introduces TM as our main tool in scrutinizing SDs. Sections 3 and 4 present case studies that contrast SDs with the proposed TM modeling.

II. TM Modeling

The TM model articulates the ontology of the world in terms of an entity that is simultaneously a *thing* and a *machine*, called a *thimac* [15-25]. A thimac is like a double-sided coin. One side of the coin exhibits the characterizations assumed by the thimac, whereas on the other side, operational processes emerge that provide dynamics. A thing is subjected to doing, and a machine does.

Thimacs are a source for generic constructs that can be applied in conceptual modeling to describe structure and behavior as a world of systems (thimacs). The generic actions in the machine (see Fig. 1) can be described as follows:

Arrive: A thing moves to a machine.

Accept: A thing enters the machine. For simplification, we assume that all arriving things are accepted; hence, we can combine the *arrive* and *accept* stages into one stage: the receive stage.

Release: A thing is ready for transfer outside the machine.

Process: A thing is changed, but no new thing results.

Create: A new thing is born in the machine.

Transfer: A thing is input into or output from a machine.

Additionally, the TM model includes storage and triggering (denoted by a dashed arrow in this study’s figures), which initiates a flow from one machine to another. Multiple machines can interact with each other through movement of things or triggering. Triggering is a transformation from one series of movements to another.

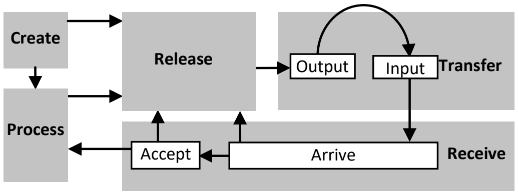


Fig. 1. Thinging Machine.

III. ATM Example

Consider the simple SD for withdrawing cash from an ATM shown in Fig. 2 [26].

A. Static TM Model

Fig. 3 shows the corresponding TM model. The figure illustrates the following:

1) The user inserts his/her card (1), which flows (2) to the ATM to be processed (3). The ATM extracts the card’s number (4), which is sent to the bank system (5).

Note: In the original SD, the message from the ATM to the bank is “verify card”. Here, the message does not state what is sent.

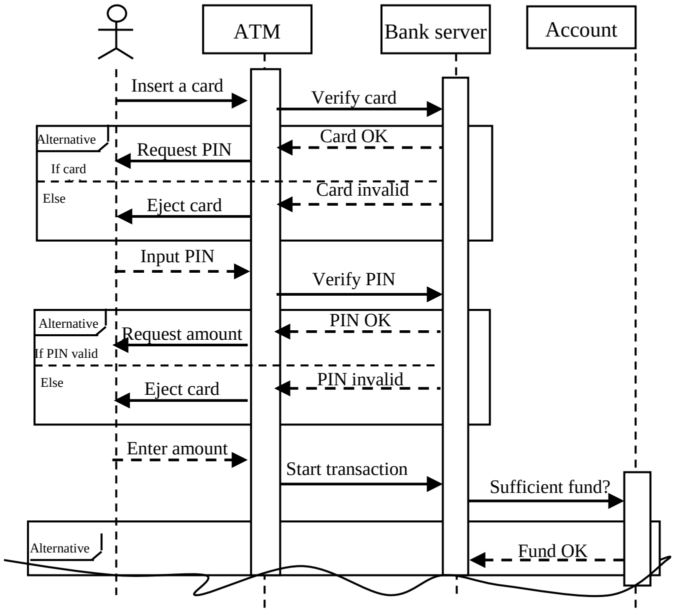


Fig. 2. Sample Sequence Diagram. (Redrawn, Incomplete from [26]).

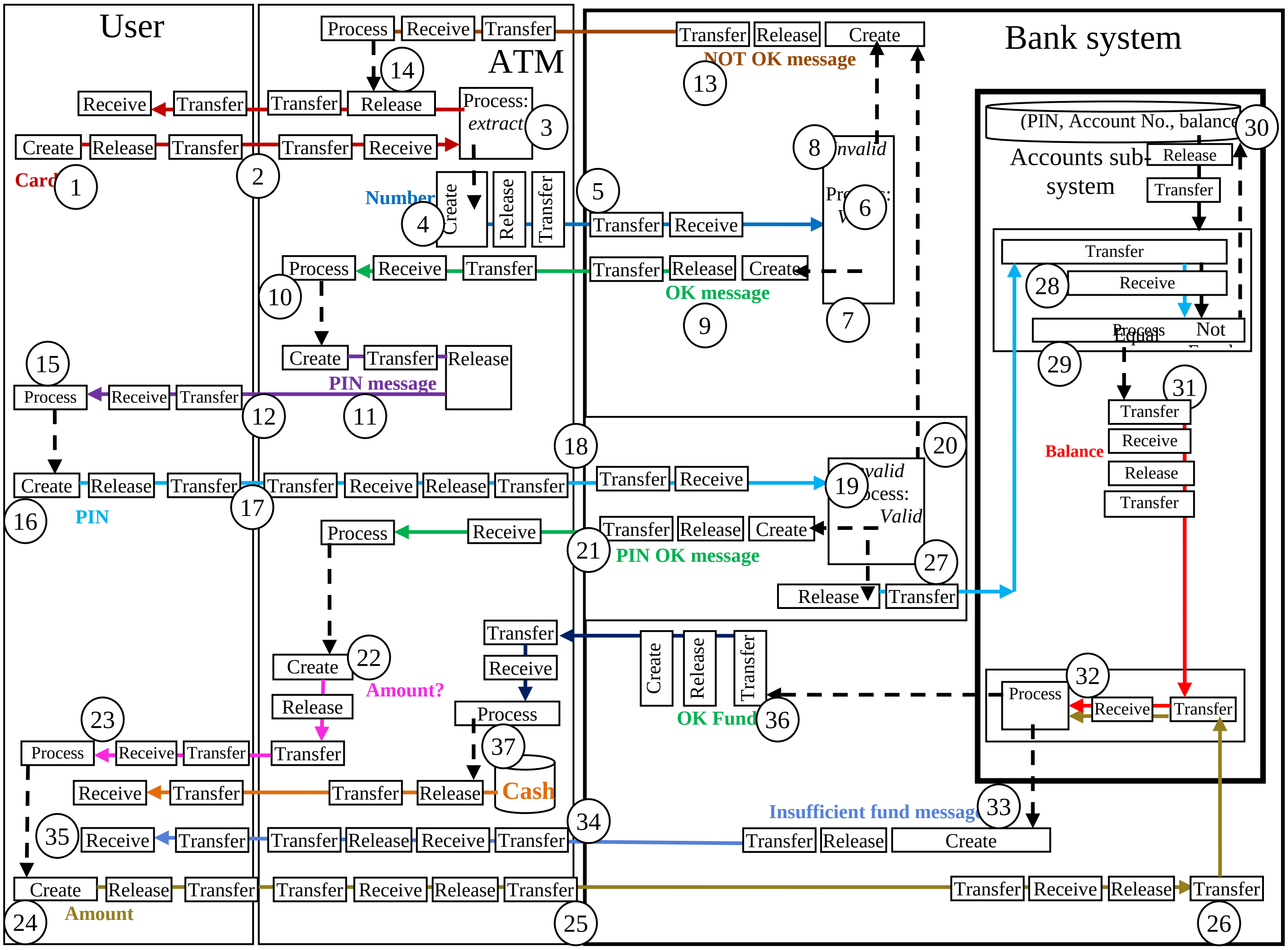


Fig. 3. Static Model of the ATM.

What is sent can be misunderstood as the physical card, which is called a "message". The correct understanding is that the ATM sends the embedded card *number* to be verified. This ambiguity about what is sent can be eliminated by explicitly specifying the process of extracting the number from the physical card. In a TM, the arrow from the user to the ATM is a physical card, and the arrow from the ATM to the bank is a number; these two arrows never cross.

2) The card number is processed (6) in the bank system (not in the bank itself) as valid (7) or invalid (8). If the card number is valid, then an OK message (9) is sent to the ATM. The ATM processes the message (10) and triggers the construction of a request for a PIN (11) that flows to the user (12). If the card number is invalid, then the bank system constructs a "not OK" message (13), which flows to the ATM to trigger the ejection (14) of the card.

Note: This TM part corresponds to *Alternatives* in the SD. The flows of different kinds of things (card, card number, OK/not OK, PIN messages) are separated by triggering.

3) The user processes (reads; 15) the message to trigger the creation of the PIN (16) that flows to the ATM (17), which sends it to the bank system (18). In the bank system, the PIN is processed (19). If the PIN is invalid (20), a not-OK message (13) is sent to the ATM, causing the ejection of the card (14). If the PIN is valid (21), an OK message (22) is sent to the ATM. The ATM constructs a message asking for the amount, which is sent to the user (23).

Note: The above description corresponds to the second *Alternative* in the SD. The TM model does not repeat the process of ejecting the card when a PIN is invalid (13 and 14).

4) The user enters the amount (24), which is received by the ATM and sent to the bank system (25), which in turn sends it to the account subsystem (26).

Note: The account system is a subsystem of the bank system. It is important to know that the bank system stored the previously given, say, PIN because the account subsystem needs such information to allocate the amount and decide whether the account is sufficient. This is performed when the PIN is valid (27). The SD skips over these parts of the scenario, leaving the model either incomplete or disconnected.

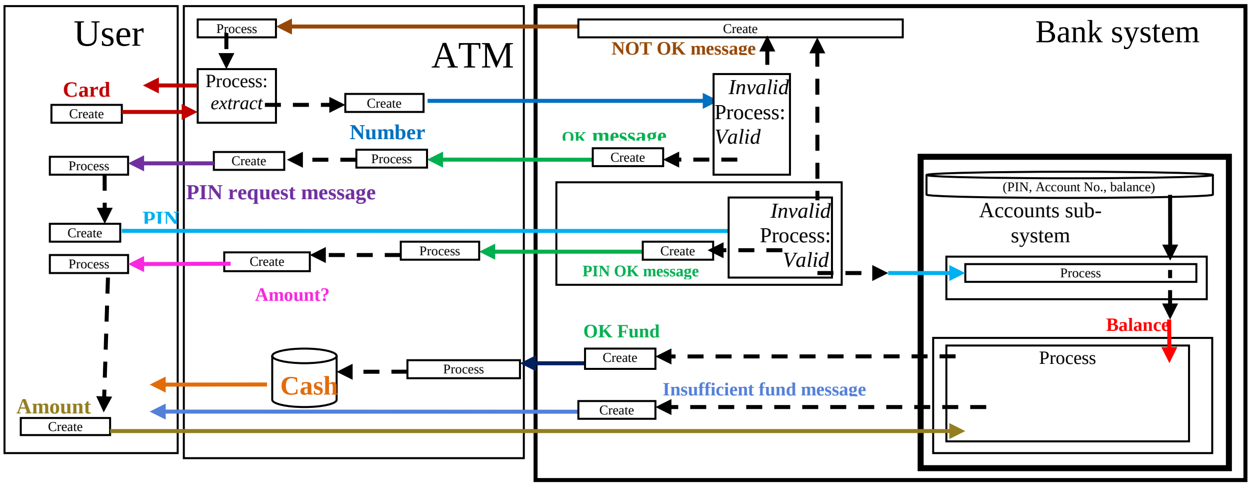


Fig. 4. Simplification of Static ATM Model.

5) The account subsystem receives the PIN (28) and searches for the corresponding account or amount in its database (i.e., PIN, account number, balance). Accordingly, the corresponding balance is extracted (31) and compared (32) with the requested amount. If the funds (balance) are insufficient, then an insufficient funds message (33) is sent to the ATM, which in turn sends it to the user (34 and 35). If funds are sufficient, then an "OK fund" message (36) is sent to the ATM. The ATM releases the corresponding amount of cash to the user (37).

B. *Simplification (if Needed)*

Of course, sequencers will complain that the TM model is more complex than their SD. There is no argument against the model being simple, however—the model must include as many important situations as is practical. Simplicity should not reduce reasonable completeness. In the ATM example, it is reasonable and important to expect that what is sent to the bank would be the number embedded in the card. It is also reasonable and important to indicate that the amount message (sufficient/insufficient) sent by the account system involves searching for a balance in the account database. Such omissions are common in SDs. The additions supplied by the TM model to the ATM problem reflect more or less relevant expansions to the conception of the problem. Fig. 3 can be simplified by assuming that the arrow direction indicates the direction of flow; thus, the transfer, release, and receive steps can be eliminated, resulting in Fig. 4.

C. *Time and Behavior*

Additionally, time in SDs is said to be represented along the vertical dimension. However, this confuses time with logical order. Suppose that the numbers 1, 2, and 3 are listed vertically in order. This does not mean that 1 *happens* at time 1, 2 happens at time 2, and 3 happens at time 3. The relationship among 1, 2, and 3 is the logical relationship "less than" or "greater than", and there are no time-related events involved. To make the relationships among 1, 2, and 3 into events, we must introduce the idea of "existence" or "presence" ("creation" in a TM). Suppose that for a certain system, at time zero, 3 is created (born in the context of the system), and then 1 is created. Then, we can have the events (time zero, 3) and (time zero + 1, 1). In this case, we can say that 3 happened before 1, regardless of the logical relationship between 1 and 3 (e.g., less than). Of course, with this understanding, we can simplify the events and write them as *3* then *1* (without mentioning times explicitly).

Another issue is that the SD forces non-logical relationships. For example, the card and the PIN can be entered in any order or concurrently (e.g., to speed up the transaction). Because these are independent and multiple methods of verification, their order and execution are logically immaterial. Such a situation can be observed clearly when websites ask for an account and password simultaneously: when one of them is wrong, the system does not tell you which one. Incorporating this multi-verification case requires extending the SD. Still, the basic conceptual problems in the SD are claiming that the vertical order is time, failing to define what time is, and failing to define what an event is.

A model exists in time as much as space, and it must situate itself in the temporal. The TM model fuses space and time into a single dynamic model of events. TM modeling involves two planes of modeling: staticity and dynamics. The static model involves spatiality (containerization) and actionality (generic actions). Spatiality involves recognizing the thimac areas that partition the model, taking into account the connections (flows and triggering) between these areas.

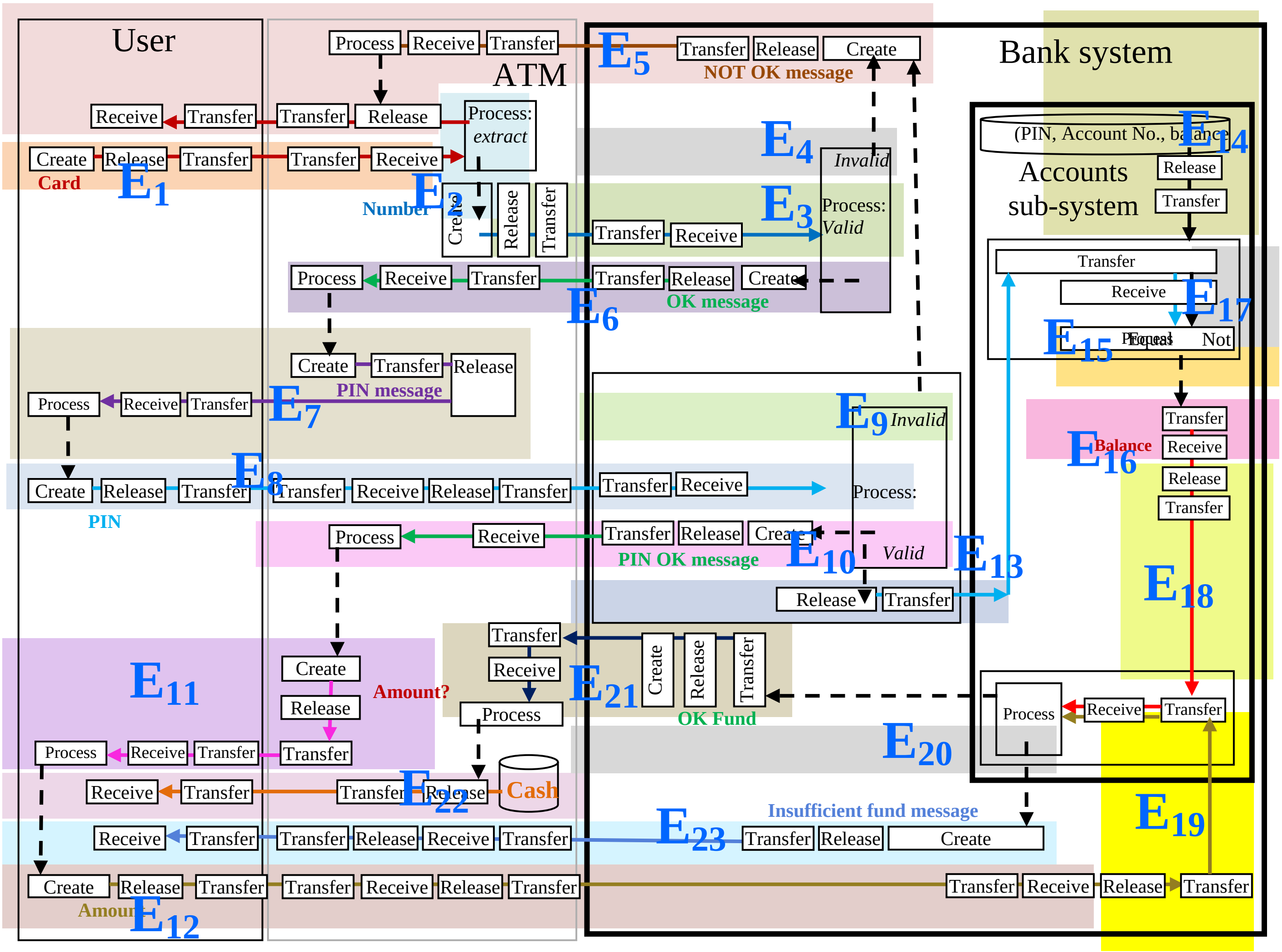


Fig. 5. Dynamic ATM Model, Regions of Events.

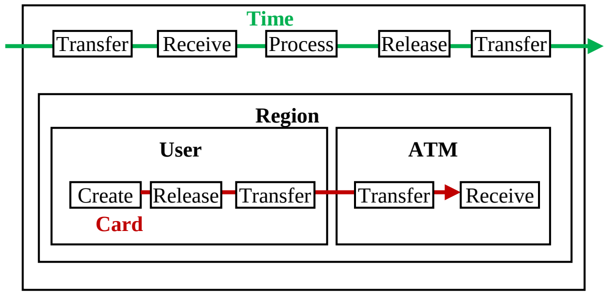


Fig. 6. The Event the user Inserts a Card that is received by the ATM.

A union of TM spatiality/actionality with time defines Event 5. The event blends such a spatiality/actionality thimac with time. Actuality here means the five generic actions. In the ATM example, Fig. 5 shows the machine of the event *the user inserts a card that is received by the ATM*. Accordingly, the static model (Fig. 3) is divided into parts; each represents the region of an event, as shown in Fig. 6.

The SD does not include the notion of time because an event is made up of the time + region of the event, which includes the boundary and actions. We can specify the events in the ATM model as follows.

Event 1 (E_1): The user inserts a card that flows to the ATM.

Event 2 (E_2): The ATM processes the card and extracts its number.

Event 3 (E_3): The card number is sent to the bank, where it is processed.

Event 4 (E_4): The card number is invalid.

Event 5 (E_5): A not-OK message is sent to the ATM, and the ATM ejects the card.

Event 6 (E_6): The card number is valid; hence, an OK card number message is sent to the ATM.

Event 7 (E_7): The ATM requests the PIN.

Event 8 (E_8): The user inputs the PIN, which flows to the ATM, which sends it to the bank.

Event 9 (E_9): The PIN is invalid.

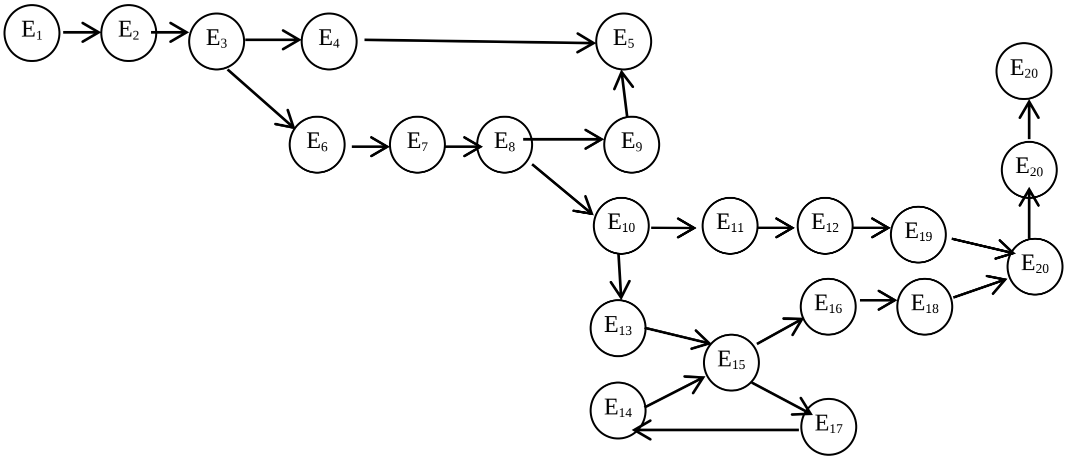


Fig. 7. Behavioral Model of the ATM.

Event 10 (E_{10}): The PIN is valid; hence, an OK PIN message is sent to the ATM.

Event 11 (E_{11}): The ATM requests the amount.

Event 12 (E_{12}): The user inputs the amount, which flows to the ATM, which sends it to the bank.

Event 13 (E_{13}): The bank system sends the PIN to the database system.

Event 14 (E_{14}): The database system retrieves a PIN record.

Event 15 (E_{15}): The database system compares the user's PIN with the record's PIN.

Event 16 (E_{16}): The two PINs match; hence, the corresponding balance is extracted.

Event 17 (E_{17}): The two PINs do not match.

Event 18 (E_{18}): The balance flows to be compared with the amount.

Event 19 (E_{19}): The amount flows to be compared with the balance.

Event 20 (E_{20}): The balance and amount are compared.

Event 21 ($E2_{1}$): The amount of funds is OK; hence, a message is sent to the ATM.

Event 22 (E_{22}): The ATM disburses cash.

Event 23 (E_{23}): Funds are insufficient; hence, a message is sent to the ATM that, in turn, sends it to the user.

Fig. 7 shows the resultant behavioral model of the ATM.

D. *Contrasting the SD with the TM Model*

Consider contrasting the two representations:

- The SD with its objects, messages, communication arrows, events, vertical chronology of events, fragments, activities (e.g., insert, verify, eject, request, input, start, enter, etc.) … (see Fig. 2)
- The TM model with its things, machines, five actions, two types of arrows, events, behavior (see Fig. 3, 6, and 7).

It seems that the TM model is more systematic (coherent whole), with a clear separation of the static and dynamic (e.g., events) aspects of the system. Such a claim will be substantiated further with the modeling case in the next section.

IV. ROLE OF THE SD IN UML

The SD is a UML interaction diagram that *bridges* the user requirements specified in the use case model and the object classes specified in the structural model [27]. Syn [13] discussed the example shown in Fig. 8 to demonstrate the systems analysis process, by which the class diagram is derived from use cases. According to [13], the figure "shows the important role of the sequence diagram. The sequence diagram bridges the requirements specified in the use case model with operations of object classes in the structural model."

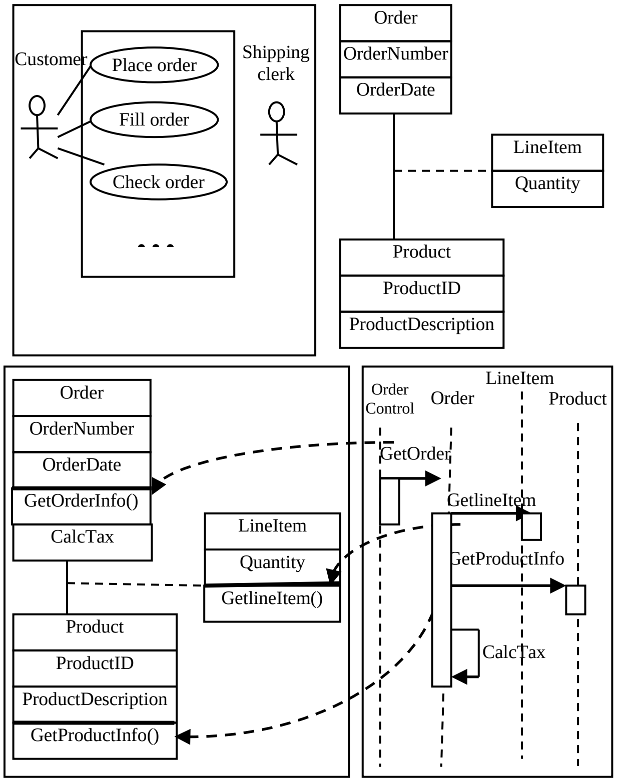


Fig. 8. Sample UML Systems Analysis Process. (Redrawn, Incomplete from [13]).

A. *Static Model*

Fig. 8 reflects how complicated the integration of the UML diagrams is and the role of the SD in such a process. By contrast, a TM has a single diagram that makes it easy to develop an integrated and consistent specification. Fig. 9 shows the static TM model of Syn's [13] ordering system built according to our understanding of Syn's description with minimum additions required to fill the gaps in the given details.

- In Fig. 9, there are two spheres (machines): the customer (Circle 1) and the company (2). The customer creates an order (3). Note that the order structure (4) is similar to a UML class diagram without the methods.
- The order then flows to the company (5), where it is received (6) and processed (7).
- This processing extracts the product number (8) from the order and sends this number to a module (9) that extracts information about that product.
- Accordingly, a process starts for the record of that product in the product database (10). In such a process, a record is retrieved from the product file (11), processed (12), and the product number is extracted (13). Hence, the two product numbers (from the order and from the system file) are compared (14).
  - If they are not the same (15), a new record is retrieved from the database (16).
  - If they are the same, the product price (17) and description (18) are extracted from the product record.
- From the price (17) and the quantity (19 – extracted from the order), the total price is calculated (20). Tax (21) is also calculated. It is possible to develop the invoice here, but because [13] did not mention such an item, we ignore it.
- Finally, the order (22) and the product description (23) are sent to the inventory warehouse to retrieve the actual product (24) and deliver it to the customer (25).

B. *Dynamic Model*

We continue developing the behavior of Syn's [13] ordering system. First, decomposability is applied to form events. Then, the chronology of events is identified to specify behavior. Fig. 10 shows the decomposition of the static model into regions of events. For simplicity's sake, the time flow is not shown. Accordingly, we have the following events in Syn's [13] ordering system.

Event 1 (E_1): A customer creates an order.

Event 2 (E_2): The order flows to the company.

Event 3 (E_3): The order is processed to extract the product number.

Event 4 (E_4): The product number flows to be processed to retrieve the corresponding record from the database.

Event 5 (E_5): A record is retrieved from the product database.

Event 6 (E_6): The product number is extracted from the retrieved record.

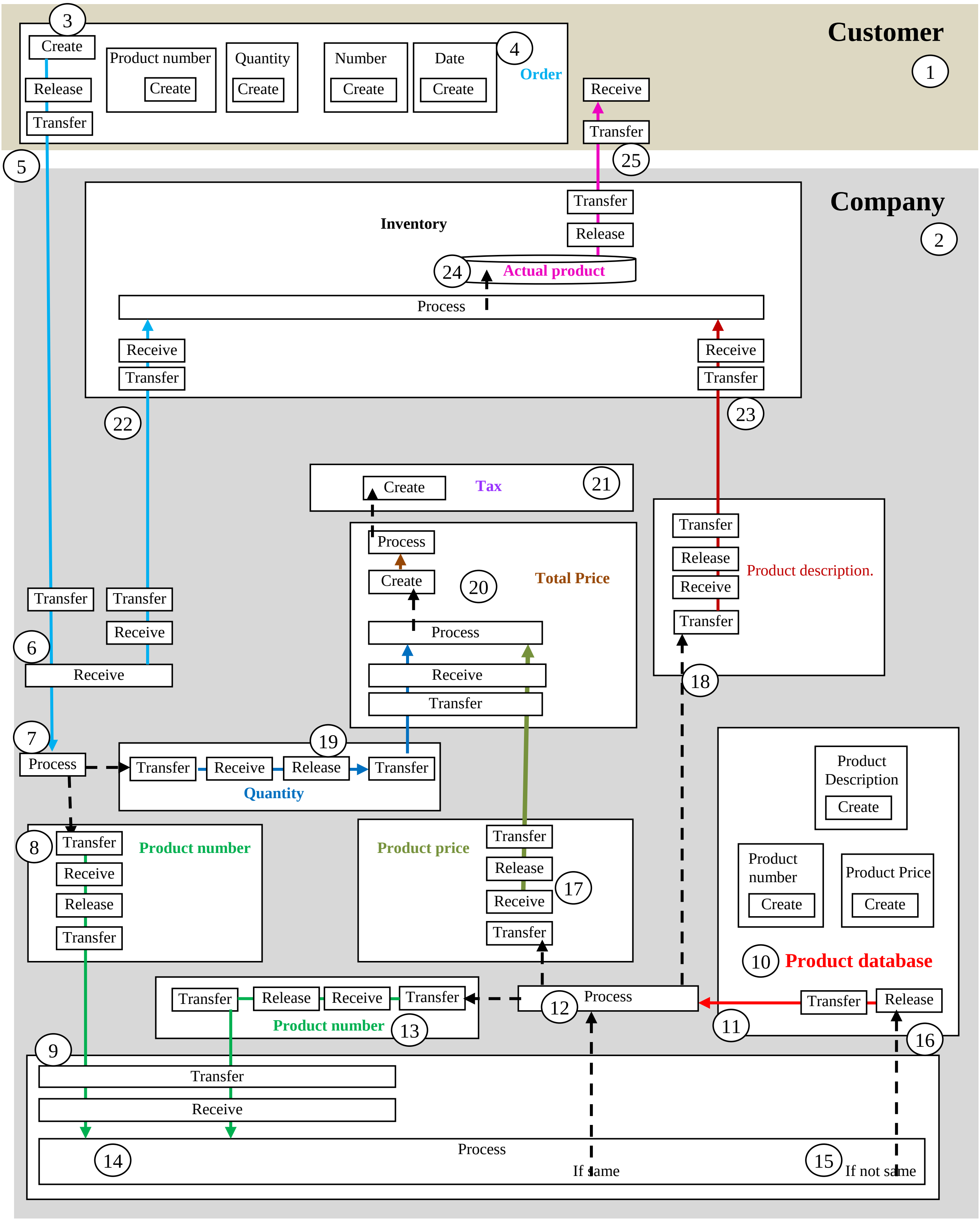


Fig. 9. Static TM Model of Syn's [13] Ordering System.

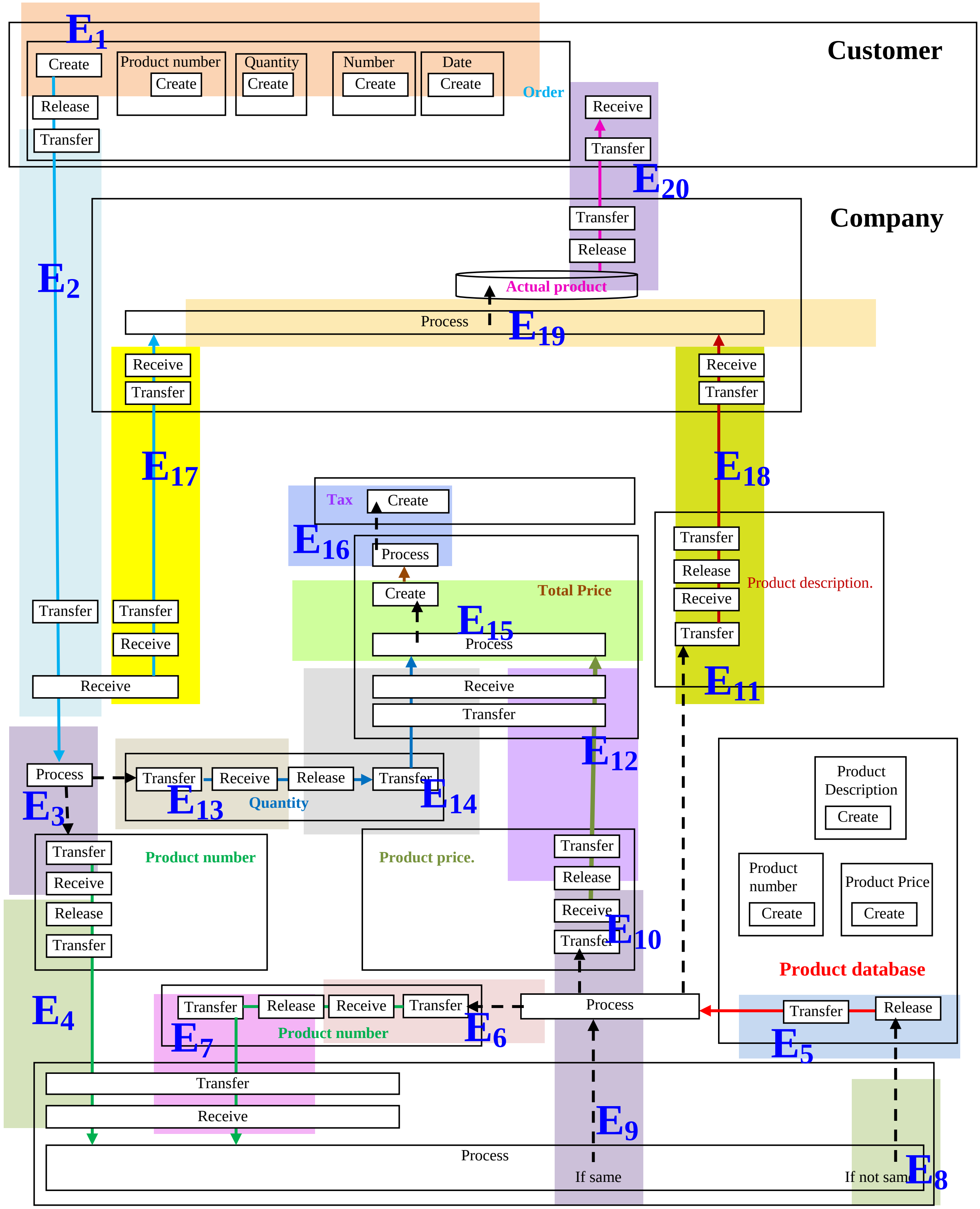


Fig. 10. Decomposition of the Static Model.

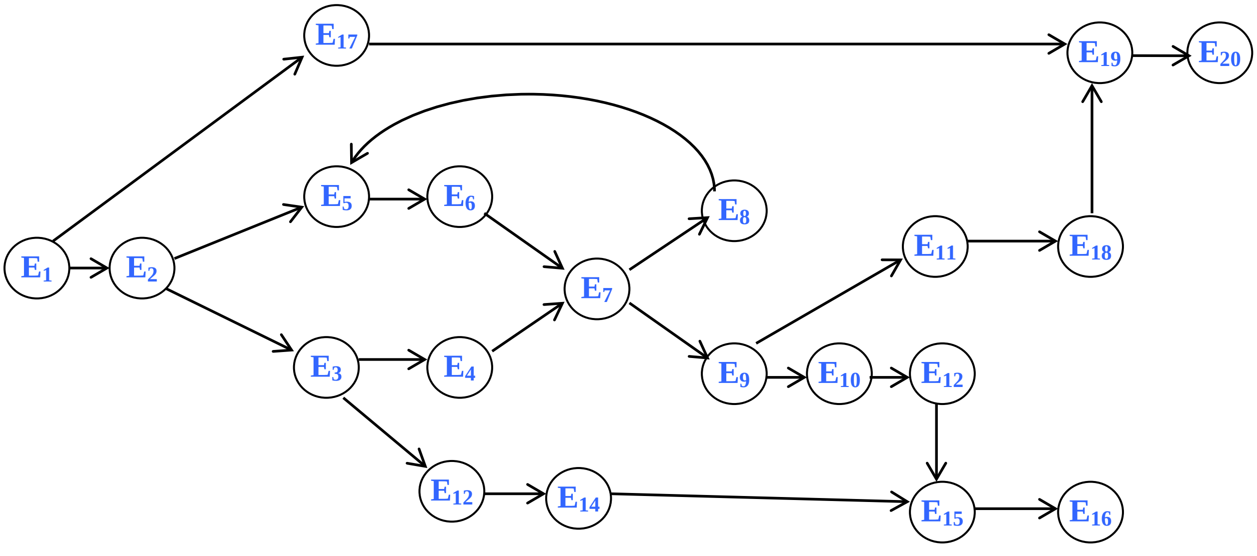


Fig. 11. Behavioral Model of the Ordering System.

Event 7 (E_7): The product number of the retrieved record flows to be compared with the order product number.

Event 8 (E_8): The product number in the order is not the same as the product number in the retrieved record.

Event 9 (E_9): The product number in the order is the same as the product number in the retrieved record.

Event 10 (E_{10}): The product price is extracted.

Event 11 (E_{11}): The product description is extracted.

Event 12 (E_{12}): The product price flows to a procedure that calculates the total price.

Event 13 (E_{13}): The quantity in the order is extracted from the order.

Event 14 (E_{14}): The quantity flows to the procedure that calculates the total price.

Event 15 (E_{15}): The total price is calculated.

Event 16 (E_{16}): The tax is calculated based on the total price.

Event 17 (E_{17}): The order flows to the inventory system.

Event 18 (E_{18}): The product description flows to the inventory system.

Event 19 (E_{19}): The order and the product description are used to retrieve the actual product.

Event 20 (E_{20}): The actual product is sent to the customer.

Fig. 11 shows the behavioral model of the ordering system.

V. Conclusion

In this paper, we proposed an alternative generalized model to the UML SD. This was motivated by difficulties mentioned in the literature regarding developing a suitable semantics-based refinement of required SD behavior. Additionally, studies have shown that graduating students were unable to create software designs, and many students exhibited difficulties in identifying valid interacting objects and constructing messages with appropriate arguments.

The proposed modeling methodology, called TM, extends the modeling process horizontally by removing the superficial dimensional limitation of vertical-only expansion to preserve the logical chronology of events. TM diagramming is spread nonlinearly in terms of actions. Events and their chronology are constructed on a second plane of description that is superimposed on the initial static description. The result is a more refined representation that simplifies the modeling process. We demonstrated this by remodeling SD cases from the literature. TM modeling can be applied to all systems that incorporate SDs. We claim that the results would be a clearer description with better semantics based on the notion of TM actions and events.

The TM model is more systematic (a coherent whole), with clearer separation of the static and dynamic aspects of the system than SD modeling. Our claim is substantiated by the two remodeled study cases above. Accordingly, it seems that adopting this new approach requires studying how to integrate the TM model into the UML diagramming apparatus. Such an issue is a research topic for future work.

References

[1] H. Koç, A. M. Erdoğan, Y. Barjakly, and S. Peker, “UML diagrams in software engineering research: A systematic literature review,” MDPI Proc. 2021, vol. 74, 13. DOI: 10.3390/proceedings2021074013

[2] R. S. Pressman and B. R. Maxim, Software Engineering: A Practitioner’s Approach, 8th ed., New York: McGraw-Hill Education, 2015.

[3] H. Bersini, “UML for ABM,” J. Artif. Soc. & Soc. Simul., vol. 15, no. 1, 2012, 15(1) 9. DOI:10.18564/jasss.1897

[4] S. Alhazmi, C. Thevathayan, and M. Hamilton, “Learning UML sequence diagrams with a new constructivist pedagogical tool: SD4ED,” Proc. 52nd ACM Tech. Symp. Comput. Sci. Educ., pp. 893–899, March 2021. DOI: 10.1145/3408877.3432521

[5] Ø. Haugen, K. E. Husa, R. K. Runde, and K. Stølen, “Why timed sequence diagrams require three-event semantics,” in Scenarios: Models, Transformations and Tools. Lecture Notes in Computer Science, vol. 3466, S. Leue and T. J. Systä, Eds. Berlin: Springer, 2005, pp. 1–25. DOI: 10.1007/11495628_1pp 1-25

[6] H. Shen, “A formal framework for analyzing sequence diagram,” Ph.D. Dissertation, University of Texas at San Antonio, May 2013.

[7] S. Busard, C. Ponsard, and C. Pecheur, “Verification of scenario-based behavioural models using Capella and PyNuSMV,” Proc. 9th Int. Conf. Model-Driven Eng. Software Dev., pp. 337–343, 2021. DOI: 10.5220/0010346103370343

[8] R. Balzer, “Tolerating inconsistency,” Proc. 13th Int. Conf. Software Eng., pp. 158–165, May 1991.

[9] Z. Micskei and H. Waeselynck, “The many meanings of UML 2 sequence diagrams: A survey,” Software & Syst. Model., vol. 10, no. 4, pp. 489–514, 2011.

[10] Z. A. Hamza and M. Hammad, “Analyzing UML use cases to generate test sequences,” Int. J. Comput. Digital Syst., vol. 10, no. 1, pp. 127–134, January 2021. DOI: 10.12785/ijcds/100112

[11] L. Lu and D.-K. Kim, “Required behavior of sequence diagrams: Semantics and refinement,” 16th IEEE Int. Conf. Eng. Complex Comput. Syst., pp. 127–136, April 2011, Las Vegas, United States. DOI: 10.1109/ICECCS.2011.20

[12] V. Y. Sien, “An investigation of difficulties experienced by students developing unified modelling language class and sequence diagrams,” Comput. Sci. Educ., vol. 21, no. 4, pp. 317–342, 2011. https://doi.org/10.1080/08993408.2011.630127

[13] T. Syn, “Improving novice analyst performance in modeling the sequence diagram in systems analysis: A cognitive complexity approach,” Ph.D. Dissertation, Florida International University, Miami, Florida, 2009.

[14] S. Alhazmi, C. Thevathayan, and M. Hamilton, “Interactive pedagogical agents for learning sequence diagrams,” in Artificial Intelligence in Education: AIED 2020 Lecture Notes in Computer Science, vol. 12164, I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham: Springer, 2020, pp 10–14. DOI: 10.1007/978-3-030-52240-7_2

[15] S. Al-Fedaghi, “Diagrammatic formalism for complex systems: More than one way to eventize a railcar system,” Int. J. Comput. Sci. Network Secur., vol. 21, no. 2, pp. 130–141, 2021. DOI: 10.22937/IJCSNS.2021.21.2.15

[16] S. Al-Fedaghi, “UML modeling to TM modeling and back,” Int. J. Comput. Sci. Network Secur., vol. 21, no. 1, pp. 84–96, 2021. DOI: 10.22937/IJCSNS.2021.21.1.13

[17] S. Al-Fedaghi, “Advancing behavior engineering: Toward integrated events modeling,” Int. J. Comput. Sci. Network Secur., vol. 20, no. 12, pp. 95–107, 2020. DOI: 10.22937/IJCSNS.2020.20.12.10

[18] S. Al-Fedaghi and M. BehBehani, “Thinging machine applied to information leakage,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, pp. 101–110, 2018. DOI: 10.14569/IJACSA.2018.090914

[19] S. Al-Fedaghi and A. Alrashed, “Threat risk modeling,” 2nd Int. Conf. Commun. Software & Networks, pp. 405–411, February 2010, Singapore. DOI: 10.1109/ICCSN.2010.29

[20] S. Al-Fedaghi, G. Fiedler, and B. Thalheim, “Privacy enhanced information systems,” 15th Eur.–Jpn. Conf. Inf. Model. & Knowl., in Front. Artif. Intell. Appl., vol. 136: Information Modeling and Knowledge Bases XVII, Y. Kiyoki et al., Eds. IOS Press, 2006, pp. 94–111.

[21] S. Al-Fedaghi, “Conceptual temporal modeling applied to databases,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 1, pp. 524–534, 2021. DOI: 10.14569/IJACSA.2021.0120161

[22] S. Al-Fedaghi, “Conceptual software engineering applied to movie scripts and stories,” J. Comput. Sci., vol. 16, no. 12, pp. 1718–1730, 2020. DOI: 10.3844/jcssp.2020.1718.1730

[23] S. Al-Fedaghi, “Modeling in systems engineering: Conceptual time representation,” Int. J. Comput. Sci. Network Secur., vol. 21, no. 3, pp. 153–164, March 2021. DOI: 10.22937/IJCSNS.2021.21.3.21

[24] S. Al-Fedaghi, Model Multiplicity (UML) Versus Model Singularity in System Requirements and Design, Int. J. Comput. Sci. Network Secur., VOL.21 No.4, pp. 103-114, April 2021. DOI: 10.22937/IJCSNS.2021.21.4.15

[25] S. Al-Fedaghi, “Conceptual Model with Built-in Process Mining,” Int. J. Adv. Comput. Sci. Appl., Vol. 12, No. 3, pp. 140–149, March 2021. DOI: 10.14569/IJACSA.2021.0120318

[26] Lucidchart, “How to make a UML sequence diagram,” YouTube Video presentation, August 27, 2018.

[27] J. F. George, D. Batra, J. S. Valacich, and J. A. Hoffer, Object-Oriented Systems Analysis and Design, 2nd ed., Upper Saddle River, NJ: Prentice Hall, 2007.

A Succinct Novel Searching Algorithm

Celine[1], Shinoj Robert[2], Maria Dominic[3]
Research Scholar
Department of Computer Science
Sacred Heart College, Tirupattur, India

***Abstract*—A searching algorithm was found to be effective in producing acutely needed results in the operation of data structures. Searching is being performed as a common operation unlike other operations in various formats of algorithms. The binary and linear search book a room in most of the searching techniques. Going with each technique has its inbuilt limitations and explorations. The versatile approach of different techniques which is in practice helps in bringing out the hybrid search techniques around it. For any tree representation, the sorted order is expected to achieve the best performance. This paper exhibits the new technique named the biform tree approach for producing the sorted order of elements and to perform efficient searching.**

***Keywords*—Time complexities; space complexities; searching algorithm; biform tree; pre-order traversal**

I. Introduction

Data structure performs as fundamental in the area of computing. The efficient search and sort are possible only if the data is organized into the headed process as structured delegacies.

The data organization plot is exemplified as a well-known data structure representation from Fig. 1 and clearly states the individual representation on data-based classification. It also projects the clear representation takes away to the immediate access and handling of data. Understanding data structure and algorithm is very difficult unless searching and sorting are not made and also brought into effect. Each desirable algorithm is chosen based on the data structure type [1]. All searching algorithms lead to efficient retrieval of a specific element from the listed aggregation of elements [2]. Until the desired result is found the search process continuous in all versatile techniques.

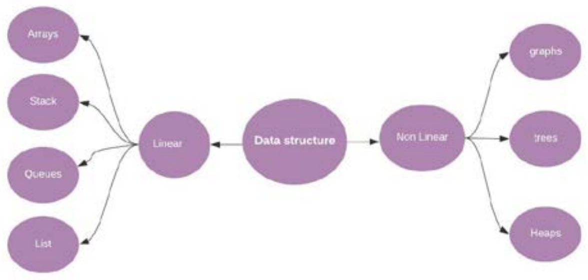


Fig. 1. Data Organization Representation.

II. Searching Techniques

A. Eccentrics of Searching

Searching is a process of accruing and discovering factors from the given list. The searching and sorting algorithms assist in arranging the elements in some order. In deliverance to the efficiency of algorithms including merge and sort technique, sorting is needed [3]. In general, searching is applied to alphabets, strings, and characters other than numbers.

The search algorithms projected in Fig. 2 are on the mind map view featuring significant divisions and classifications of search algorithms that are widely applied on Artificial Intelligence techniques-based algorithms. From the root of the search algorithm, the classification is divided into major divisions as clueless (uninformed) and communicated (informed). The first division clueless of searching explicit the exploration in each step. The goal is to split into each state of activity and explored if not. Less domain knowledge is expected in this type of search and it increases in time complexity. The operations are executed in a brute force method and the result of the current step advances to the next level of implementation. The brute force carries the selective information of traversing from a tree to a festinated step. The communicated search comes with information patterns in each step to find the solution which results quickly in the process. This pattern includes the domain knowledge that results in the heuristic way of approach within fair timing. The complex rich problems are focused and lead to a better solution in this way of approach [4]. The classifications above represent the different searching techniques by all possible means of representing the utmost classification needed.

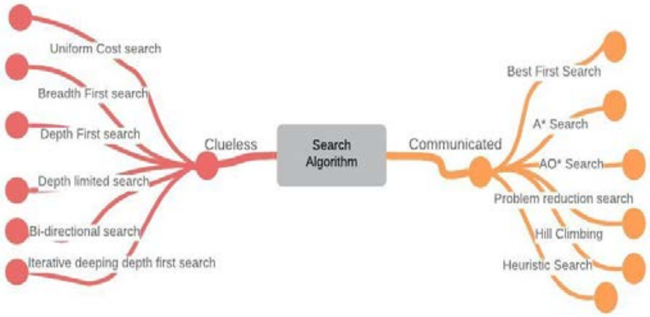


Fig. 2. Searching Types Classifications.

B. Handed-Down (Traditional) Searching Techniques

The search algorithm is one of the class algorithms among the existing classifications on constant, logarithmic, linear, quadratic, and exponential algorithms. The searching is to fill one more piece of the above classifications that are valuable. A searching algorithm is a kind of obvious statement where the word is predominantly mentioned mechanism in the web portal. A search is a way to find a group of items from an implicit or an explicit way of collection. Any searching technique is made easy along with the properties provided. The properties search to reach completeness including the time and space complexities. Every sorting technique comes with a prerequisite in which the effective searching was done looking over the data provided. The comparison is on existing traditional searching techniques assures the complexities they built-in. The searching is made reasonable and efficient through the possibilities as figured and mentioned in search type classifications.

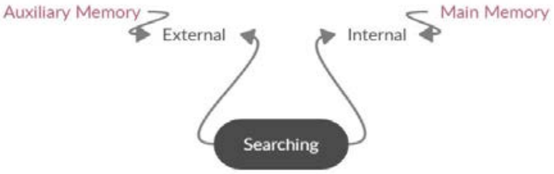


Fig. 3. Searching Proficiencies.

Fig. 3 projects the searching operations on two main classes as external searching and internal searching. The external searching is colligated with auxiliary memory occupies in the files hived away on disk storage. Internal searching is concerned with minimum data that resides on the data processor's main memory [5].

Search classification
Sorting
Searching

Fig. 4. Searching Classifications.

The searching technique has the base classification of Traditional sorting and searching as projected in Fig. 4. The comparative analysis of the different searching algorithms is possible from the base classification field.

III. Comparative Analysis

A. Comparison on Searching Performance

The performance and efficiency of each algorithm differ based on the data provided for each separate or repeated task. The methodology applied to assess the performance is time & space complexity which have a better modification of words over time and memory space in CPU [6]. Any search algorithms are calculated based on certain attributes of their complexities. Efficient searching fulfills the completeness of the searching algorithm. Table I projects the performance of the binary searching on sorted techniques that is brought through a comparative study in three strategies Performance, effectiveness & output.

TABLE I. A Comparative Study on Different Searching Algorithms over Searching

Algorithm	**Technique**	**Performance**		
		Best-face	***Worst-face***	***Fair-face***
Binary search	Divide &conquer	O(1)	O(log2 n)	O(log n)
Sequential Search	Linear search	O(1)	O(n)	O(n)
Hash Search	Hashing	O(1)	O(n)	O(1)
Tree search	Divide & conquer	O(1)	O(n)	O(n)
Interpolation search	Binary search	O(1)	O(n)	O(n)
Jump Search	Linear search	O(1)	O(n)	O(1)
Hybrid Search	Interpolation & Binary	O(1)	O(√n)	O(n)
Exponential search	Sorting	O(1)	O(log n)	O(log n)
Fibonacci search	Comparison-based	O(1)	O(log n)	O(log n)
DFS	Graph data structure	O(1)	O(|V| +|E|)	O(n+m)
BFS	Graph data structure	O(m)	O(b^m)	O(|V| +|E|)
Heuristic search	Greedy search	G(n)	O(bm)	O(bm)
Bi-directional	Graph search	O(bd)	O(bd/2)	O(bd/2)
Sequential Search	Linear search	O(1)	O(n)	O(n)

{N = 23, 10, 75, 05, 15, 82, 19, 07, 31,100}

Fig. 5. Set of N Random Elements.

The purpose of the comparison table is to equate the different searching techniques based on the individual effective performance. The running complexities include best, worst, and average (fair) cases as a comparative study. Beginning with the binary search each algorithm is classified with the technique built-in. The sequential algorithm derived from linear search finds for a particular element starting apiece, considered to be efficient with classified order. Hashing searching techniques have got their advantage over larger data volume of data sets. The results are implemented without any collision possible actions through open and close addressing methods such as family collisions [7]. Tree search addresses the issue of involving combinations with the basic idea of divide and conquer in better measure and conquer technique and the process continues till it pruned [8]. The interpolation search falls another page from the binary search techniques for the particular data set provided the best effort with the sorted factors. Jump search has a limitation over certain block representations with the intervals and has control as block search. Hybrid search is constructed over the sorted and unsorted distribution of arrays. It blends from the staple of binary and interpolation search for efficient search acquiring the advantage point over both algorithms [9]. The exponential search is selfsame to binary search in which the proportions are equal at the depth of the node, this projects the minor level enhanced by an increasing factor of 2 [10]. This exponents the children at each level in order.

From the classification on different searching algorithms, BST is one of the best-practiced techniques of linear data representation types. BST holds a special name in terms of representing in sorted order which pushes the process much effective & easier. It results in optimal performance for the end-user. The construction of a binary tree protrudes from the root node [11]. The tree has the best picture once it is portrayed in the hierarchy of parent-child with a single character as a parent node (root node).

IV. Proposed Methodology

The success story of each searching algorithm will result in finding the desired element. This paper proposes the construction of effective searching. This methodology was applied to reduce the time complexities of searching principles. This paper effort in the construction of an algorithm that is capable of causing searching efficiently. To achieve such efficiency, the given N random numbers will be converted into binary form and a binary tree for the binary form will be constructed. Performing the technique of pre-order traversal on a binary tree provides a sorted set of output.

A. Phase I: Binary Conversion

The given N numbers will find a sorted position so that any given number is made available in the tree.

The set of N numbers listed is shown as the example for constructing a binary tree. The step begins by involving binary conversion for the given 10 digits into binary form. For the given values projected in Fig. 5, and converted into binary representation in an unsorted order.

B. Phase II: Grouping

The binary conversion is practiced for the set of given N numbers. The presentment of numbers into binary format is projected in Table II. These binary format numbers are grouped based on the number of bits as depicted below representation Fig. 6.

C. Phase III: Biform Tree Approach

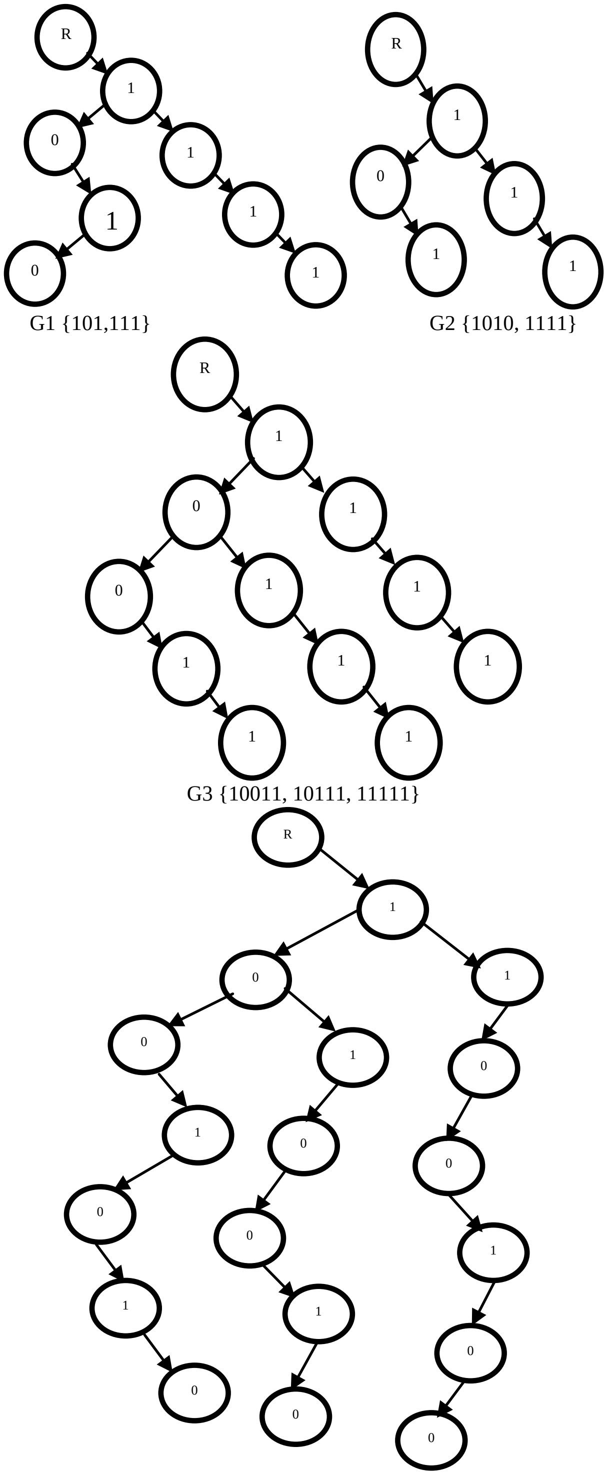
Construction of biform tree for all groups G1, G2, G3, and G4 are projected in the respective Fig. 7(a) to 7(d). The tree construction is initiated from the root node. If the binary number is 1, then it is skewed to the right-hand side of the root node otherwise to the left-hand side of the root node. This operation is continued for all the n bits.

TABLE II. Binary Representation on N Random Elements

N	23	10	75	05	15	82	19	07	31	100
	10111	1010	1001011	101	1111	1010010	10011	111	11111	1100100

G1 = {101, 111}
G2 = {1010. 1111}
G3 = {10111, 10011, 11111}
G4 = {1001011, 1010010, 1100100}

Fig. 6. Elements in Group-Wise.



G1 {101,111}

G2 {1010, 1111}

G3 {10011, 10111, 11111}

G4 {1001011, 1010010, 1100100}

Fig. 7. Biform Tree Representation for G1, G2, G3, and G4.

The effective searching for storing any sorted element is made easy in this way of tree representation. This tree construction helps in making the search faster and effective in paving an easy representation. The grouping of digits will help in looking for the needed elements by avoiding the time in search of the rest.

D. Phase IV: Tree Traversal

As a tree is a self-referential data structure, traversal can be defined by recursion. The traversal algorithm can be broadly classified into depth-first and breadth-first search as shown in Fig. 8. The depth-first algorithm is branched towards pre-order, in-order, and post-order traversal.

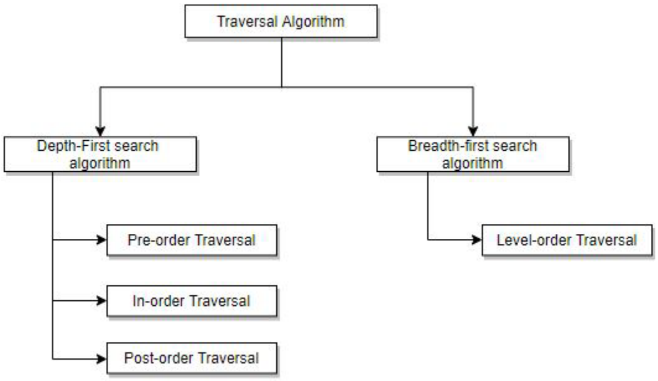
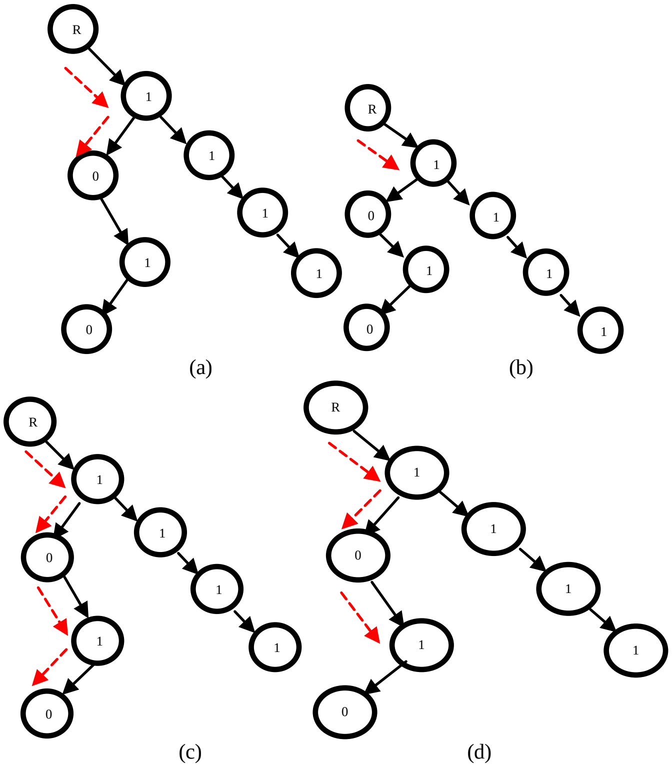


Fig. 8. Traversal Algorithm Classifications.

Classifying the traversal techniques from the depth-first search algorithm, this paper utilizes pre-order traversal to receive the numbers in sorted order.

E. Phase V: Searching Outgrowth

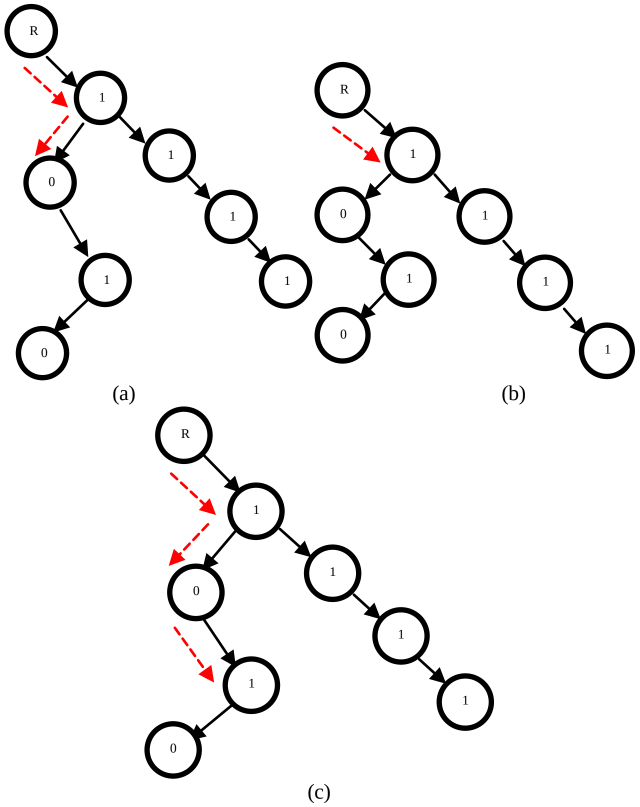
To arrange the elements in sorted order and search, pre-order traversal is performed on the tree in a depth-first manner as projected in Fig. 7. To search the particular element 1010, the list made available in the searching process is to be carried in G2 as follows.



(a) (b)

(c) (d)

Fig. 9. Searching Process of Each Node.



(a) (b)

(c)

Fig. 10. Searching Process of Each Node (a, b, and c) for Element Not Found.

The traversal begins from the root node R and moves towards the left child of its own. Since the unavailability in the left child for R, the tree travels through the right child as projected in Fig. 9(a). The travel continues through the left child of the current node as it is shown in Fig. 9(b) and the process is repeated in the following Fig. of 9(c) and 9(d). The traveling process is completed once it reaches the leaf node. Now arrange the visited node from the root node as 1010. Hence the element is found in G2.

The traversal begins from the root node R and moves towards the left child and proceeds for the right child as imaged in Fig. 10(a) by visiting the right child node labeled 1 and to the left child labeled 0. The travel continues for the right child labeled 1. Instead of the availability for the right child with the label 1, the left child labeled as 0 presence is found in the tree. The conclusion is derived from the unavailability of element 10111 in the G2. In the search made the availability of left child with label 0 is found but not the other child named right with the label of 1. Grouping in order digit representation which the binary conversion is made emphasizes on a searching part, reckon the given element to be searched and let's spot the random element as 82. The element state is to be found and the number of comparisons made here is spotted as second since it has the immediate search of prefix value. In case of the element 84, the search is ineffective resulting in not found status since the given random element is not found in any of the ordered group in particular to G4, instead if it is then the searching group falls under the G4 classification. The searching is made according to the value given as input. Here the given value is 84, in mechanical the search begins at G4 since the value ranged to 84 lies within that region. The immediate result of found or not found status on any random search is made possible by group classifications.

V. Programme Encryption of the Algorithm

The proposed algorithm is assorted into separate parts classifying sort and search functions. The algorithm produced is made effective in the C++ language. The algorithm for sorting and searching is represented easier as the given element pitch in.

VI. Experimental Result

Searching for a particular element is projected clearly under phase II, classifying bit and group-wise respectively. The bitwise operation of search is the best fit in the case of integer data type [12]. Since the main advantage over the proposed algorithm is found to be effective searching and the arrangement of given numbers is on bitwise. Searching is made easier based on group classification.

TABLE III. Algorithm Representation for Sorting

Sorting Representation	
Step 1	*Consider N random numbers*
Step 2	*Convert all N numbers into binary form*
Step 3	*Group the converted binary numbers based on the number of bits. Let it be m.* *G1(m)=3, G2(m)=4, G3(m)=5, G4(m)=7*

Step 4	*Step 4: Construct a binary tree for each Group G*
	for (i = 1 ; i < = m ; i ++) // where m ≥ 3
	{
	insert (Node node, key values) // Insert a new node in the tree
	{
	node ← input node object
	value ← Actual value of the node // values are either 0 or 1
	if (node = = null) then
	return Node (value);
	else if (value [i] = = 0) then
	node.left ← insert (node.left , value)
	Else
	node.right ← insert (node.right, value)
	return node;
	}
	}

Step 5	*Perform the pre-order traversal in the binary tree to receive the numbers in sorted order*
Step 6	*Repeat step 4 and step 5 for all Groups*
Step 7	*Stop the Process*

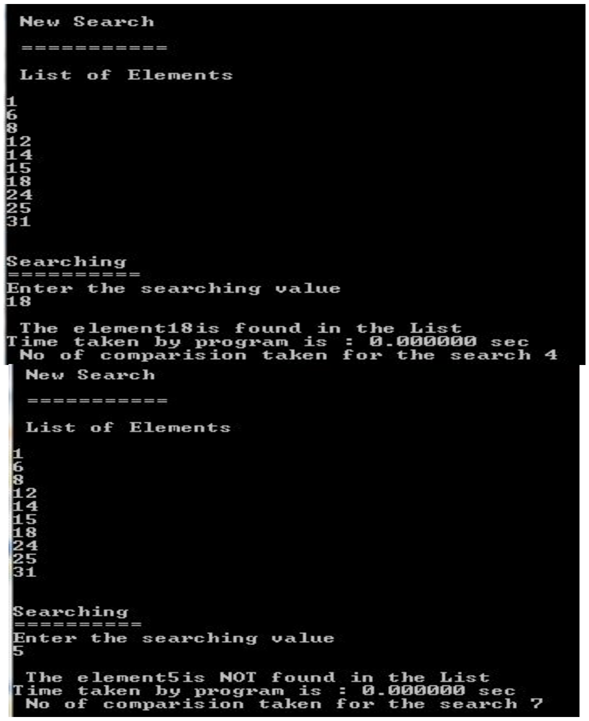
TABLE IV. Algorithm for Searching a Particular Element in the List

searching representation	
Step 1	*Start a process*
Step 2	*Get the input from the binary trees G_n, where all the numbers in sorted order and n are the number of Groups*
Step 3	*Enter the element to be searched.* *Let it be L.* *Check the element L belongs to which Group.* $\forall_i$ G_i *where 1<= i <=n* *if ($L \in G_n$)* *{* *Start from the root node R* *for (i = 1 ; i <=G_i(m) ; i ++)* *{* *if (L[i] == 0)* *Perform the searching in left sub-tree* *search (root → left)* *Else* *Perform the searching in right sub-tree* *search (root → right)* *}* *Else* *print element L is not available in N* *}*
Step 4	*Stop the process*

The algorithm projected in Tables III and IV on sort and search functionalities is applied to produce the experimented result on comparison over linear, binary, and the proposed search. In the next phases of comparisons, the results are proven over a tried-out study between linear, binary, and the proposed search. The prediction time, an element found and the comparison made are taken as measuring facts for the comparisons. Below Table V shows the comparison involves the searching technique.

TABLE V. Comparison between Linear, Binary, and Proposed Search

Elements (in random)	Algorithms		
	Binary	Linear	Proposed
	Prediction Time	Prediction Time	Prediction Time
1	0.00100	0.00200	0.00000
6	0.00100	0.00200	0.00000
8	0.00100	0.00100	0.000000
12	0.00100	0.00100	0.000000
14	0.00200	0.00100	0.000000
15	0.00100	0.00200	0.000000
18	0.00100	0.00100	0.000000
24	0.00100	0.00000	0.000000
25	0.00100	0.00100	0.000000
31	0.00100	0.00200	0.000000



```
 New Search
 ===========
 List of Elements
1
6
8
12
14
15
18
24
25
31

Searching
==========
Enter the searching value
18
 The element18is found in the List
Time taken by program is : 0.000000 sec
 No of comparision taken for the search 4
 New Search
 ===========
 List of Elements
1
6
8
12
14
15
18
24
25
31

Searching
==========
Enter the searching value
5
 The element5is NOT found in the List
Time taken by program is : 0.000000 sec
 No of comparision taken for the search 7
```

Fig. 11. The Experimented Result on the Element Found and Not Found in the Proposed Searching Technique.

Table V focus on emphasizing searching by making a comparison on linear, binary, and the proposed search numbers in an individual search mode. The experimented result is proven with the given number of 10 random numbers Consider the element 18 to be found in binary search, the prediction time taken by the binary search includes 0.00100 seconds where linear carries the same.

The biform approach tree technique is the searching time projected in Fig. 11 consists of 0.000 seconds and considers to be the quickest in delivering the search result compared with the two existing searches. In the case of an element not found represented in Fig. 10, the total time taken for traversal in search of the given element 5 is found to 0.000000 seconds. This includes the tree representation as stated in the above discussion on searching occurrences in the proposed methodology topic.

VII. Graphical Representation

The graphical representation for the effective searching over the above discussed different programming on searching techniques were applied on a set of linear, binary, and the proposed search.

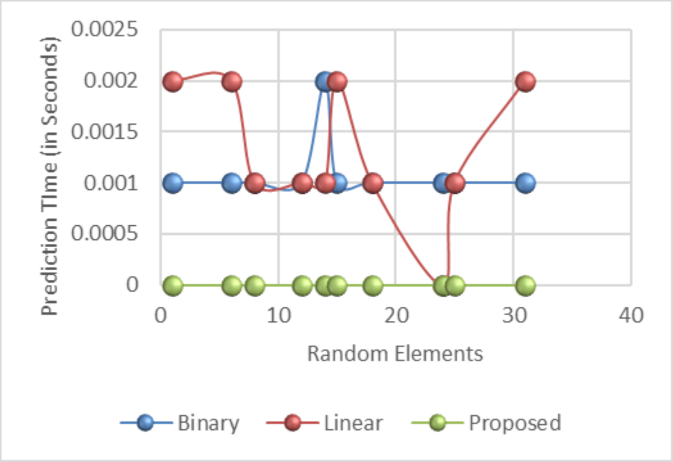


Fig. 12. Graphical Representation of Binary, Linear, and Biform Search.

A. *Graphical Representation of Searching an Element among for 10 Elements*

The graph representation is presented over linear, binary and the proposed search is exemplified as graphical representation in Fig. 12. The X-axis represents the random elements given in numbers and the Y-axis indicates the arrival of searches in seconds. The individual search includes linear, binary, and proposed are represented in separate line charts. The dots connected with blue represents the binary and others simultaneously. The graph withstands the elements up to 10 combinations in a random approach in the above graphical representation. The data projected as X-axis is made an approximate joint on Y-axis in which the total time is taken for each data and mapped into a line chart representation. Similarly, the remaining searches for the given set of random elements take place in the graphical representation.

VIII. Graphical Representation of Searching an Element among for 50 Elements

The graph projected in Fig. 13 explains the prediction time taken for each set of linear, binary, and proposed search techniques. This line chart is represented on 50 random sets of elements. The comparison time increases when the bit sizes get increased in each representation.

The maximum threshold for the binary, linear search is considered as 0.004 seconds while the minimum search happens and succeeds on the proposed search as 0.001 seconds when the given element bit sizes increase. While the elements are distributed in random the classification takes place based on group phases as discussed in the structured methodology. From the given program for efficient searching, the line chart is discussed to prove the efficient technique for searching is on a proposed searching method.

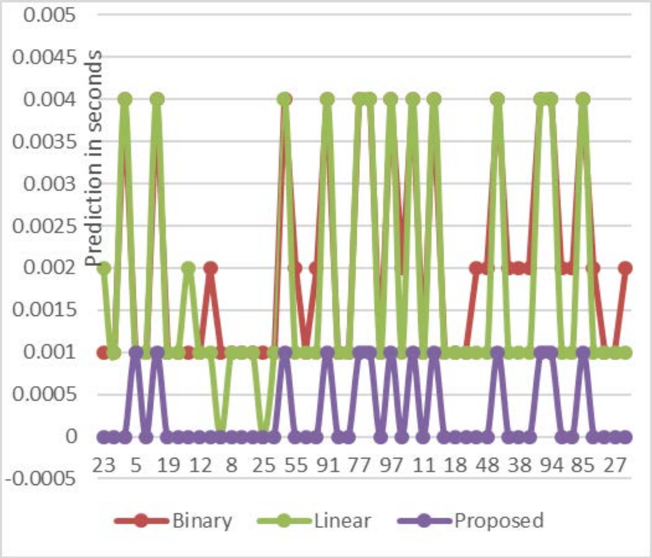


Fig. 13. Graphical Representation of Binary, Linear, and Biform Search Over 50 Elements.

IX. Graphical Representation of Searching an Element among for 100 Elements

The Graphical chart representation of Fig. 14 projects the status for the possible search of carrying a maximum of 100 random values in a given element. Fig. 14 illustrates how the searching takes place as efficient when more elements are into it.

As stated in the combination of more than 50 inputs the same criteria fit in for the case of 100 inputs. Here in Fig. 13 indicates the proposed search goes to the maximum prediction time of 0.001 seconds when the bit size increases in a large volume of size while the linear and binary goes to the extent of 0.004 seconds. From the given every efficient search in the graphical format of presentments the tree structure stood as a prerequisite for easier delegacies. Alike sorted data for binary search, tree structure for the biform tree approach act as a prerequisite inefficient search. All the above three graphical representation initialize and emphasizes on searching techniques that deliver the fast result on any values given in.

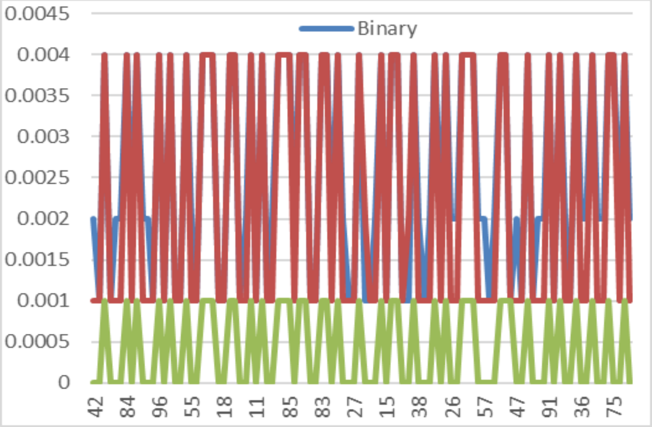


Fig. 14. Graphical Representation of Binary, Linear and Proposed Search over 100 Elements.

X. Practical Applications and Enhancements

The tree structure proposed and experimented in the graphical representation defined for searching any random element will be an effective and efficient way. The tree representation way of organizing the data was brought into effect by searching for any random value and it is highlighted in all the experimented results that search took place. Both the result of found and not found was done efficiently and it was proven in the sample searching on the experimented result page. The time complexity for the biform approach search for best face and fair face is O(log n) and the worst face is O(n).

The biform approach search compared with the existing (linear and binary) will utilize the tree structure that has a binary tree nature in organizing the data that act as an integral to lots of existing applications considering the binary tree node representation. The pointer includes left-right and parent deliberated in space consumption in the large representation of values.

The proposed search was applied in the renowned areas of acquiring knowledge from a knowledge base using tree representation, in the prediction grounded model for directly learn word representation. In the field of synonyms detection and word prediction where the words are evaluated for easy word representation. The tree structure is represented as a prerequisite for any search that is being done in the proposed method. For effective searching, a data structure in the form of a sorted order tree structure is a prerequisite like the existing searches including binary, linear, etc. The biform tree approach search was found to be effective in the finding of the learning content for the learner from the repository. Once the content is identified in the repository the corresponding knowledge graph has been created for the system to understand and learn about the learner to provide the needed learning content. The novel biform tree structure is a prerequisite for efficient searching and also for sorting the elements while doing the pre-order traversal over that tree.

XI. Conclusion

The different searching techniques make way to crystalize and clarify the problem in an effective way. It is an evolutionary method among the infinities in the improvement of time complexities. The biform tree approach inquiry on searching techniques introduces the binary representation in digits that makes the search easier & faster with well-defined generalized fit in by making a comparison table study over performance on the search and the new data structure has a built-in form of tree representation using the searching technique. The beauty of the biform tree approach is the same binary tree is utilized to produce the sorted order numbers and for the element search. The precursor for element searching is provided by the sorted binary tree in which the elements are in sorted order. Altogether searching was made effective over designing a data structure in the form of tree representation which makes the searches an efficient one with the biform approach search.

REFERENCES

[1] N. Sultana, S. Paira, S. Chandra and S.S Alam, "A brief study and analysis of different searching algorithms", International conference on electrical, computer and communication technologies, Vol 4, pp. 1-4, February 2017.

[2] A. Shoaib "A survey in different searching algorithms", International research journal of Engineering and Technology, vol. 7,p. 275, January 2020.

[3] Kamlesh Kumar Pandey and N. Pradhan, "A comparison and selection on the basic type of searching algorithm in data structure", International journal of computer science and mobile computing, Vol 3, pp. 751-758 July 2014.

[4] Choing R, Sultano J. H and W. J. Jap, " A comparative study on informed and uniformed search for intelligent travel planning in borneo island", International symposium of Information Technology, IEEE. Vol. 3, pp. 1-5, August 2008.

[5] Ana Bell, Eric Grimson, and John Guttag, "6. 0001 Introduction to computer science and programming using python, Massachusetts Institute of Technology: MIT, OpenCourseware, Fall 2016.

[6] K Roopa and J Reshma, "A comparative study on sorting and searching algorithms, International Journal of engineering and technology, Vol 5, p. 1416, January 2018.

[7] Dapeng Liu and Shaochum Xu, "An Empirical study on the performance of Hast table", Dapeng Liu et all, 13th International conference on Computer and Information science (ICIS), Vol 3, pp 60-68, January 2015.

[8] Henning Fernau and Daniel Raible, "Searching trees: an essay", International conference on Theory and applications of Models of computation, pp 59-70, May 2009.

[9] A. S. Mohammed, S. E. Amrahov and F. Celebi, "Efficient hybrid search algorithm on ordered dataset", computer engineering department, Turkey, August 2017.

[10] Xinguo Deng, Yangguang Yao, Jia Chen and Yufeng Lin "Combining breadth-first with depth-first search algorithms for VLSIwire routing", International conference on advance computer theory and engineering, pp. V6-486, 2010.

[11] Inayat Rehman, S. Khan and M. S.H. Khayal, "A survey on maintaining binary search tree in optimal shape" International conference on Information Management, and Engineering,pp. 365-369 , June 2009.

[12] K. Yordzhev, "The bitwise operations related to a fast sorting algorithm", International Journal of Advanced Computer science and applications, Vol. 4, pp. 103-107, November 2013.

AUTHORS' PROFILE

S. Celine is a part-time research scholar in the Department of Computer Science, Sacred Heart College, Tirupattur district, and working as an Assistant professor in the Department of Computer Science, Government of Arts College for Men, Krishnagiri, Tamil Nadu. Her area of research is in the field of e-Learning using Deep Learning.

Shinoj Robert is a part-time research scholar in the Department of Computer Science, Sacred Heart College, Tirupattur District, and working as an Assistant professor from 2014 in the Department of Computer Application, Don Bosco College, Yelagiri Hills, and His area of research in the field of Machine learning and E-learning.

Dr. M. Maria Dominic obtained his B.Sc., M.Sc., and M.Phil. and Ph.D. in Computer Science. He has been working in Sacred Heart College, from 1996 onwards in various capacities He has also worked in Multimedia University, Malaysia on a Contractual Basis. He has co-authored a book on OOP using C++ published by Pearson education. He has published more than 20 research articles in International Journals. He has 4 Ph.D. Research scholars working under him in the field of Artificial Intelligence especially in Machine learning and deep learning.

Earthquake Prediction using Hybrid Machine Learning Techniques

Mustafa Abdul Salam[1]
Artificial Intelligence Dept., Faculty of Computers and Artificial Intelligence, Benha, Egypt

Lobna Ibrahim[2]
Scientific Computing Dept., Faculty of Computers and Artificial Intelligence, Benha, Egypt

Diaa Salama Abdelminaam[3]
Information Systems Dept., Faculty of Computers and Artificial Intelligence, Benha, Egypt
Faculty of Computer Science, Misr International University, Cairo

***Abstract*—This research proposes two earthquake prediction models using seismic indicators and hybrid machine learning techniques in the region of southern California. Seven seismic indicators were mathematically and statistically calculated depending on pervious recorded seismic events in the earthquake catalogue of that region. These indicators are namely, time taken during the occurrence of n seismic events (T), average magnitude of n events (M_mean), magnitude deficit that is the difference between the observed magnitude and expected one (ΔM), the curve slope for n events using inverse power law of Gutenberg Richter (b), mean square deviation for n events using inverse power law of Gutenberg Richter (η), the square root of the released energy during T time (DE1/2) and average time between events (µ). Two hybrid machine learning models are proposed to predict the earthquake magnitude during fifteen days. The first model is FPA-ELM, which is a hybrid of the flower pollination algorithm (FPA) and the extreme learning machine (ELM). The second is FPA-LS-SVM, which is a hybrid of FPA and the least square support vector machine (LS-SVM). These two models' performance is compared and assessed using four assessment criteria: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Percent Mean Relative Error (PMRE). The simulation results showed that the FPA-LS-SVM model outperformed the FPA-ELM, LS-SVM, and ELM models in terms of prediction accuracy.**

Keywords*—Extreme learning machine; least square support vector machine; flower pollination algorithm; earthquake prediction

I. Introduction

Earthquake is movements and shaking inside the ground that produce energy in rocks. Like many natural disasters, Earthquake causes many damages, financial loss, and injuries [1]. Earthquakes happen daily in various regions around the world. The more susceptible regions to earth-shaking are japan [2], Indonesia, south California, turkey, Iran, and Taiwan [3]. People can feel the Earthquake if its magnitude is more than 2.5, but if the magnitude is less than 2.5, the Earthquake will not be felt. The magnitude of highly caused damaged earthquakes is more than 4.5 [4]. Sometimes earthquakes are responsible for huge numbers of deaths.

So, scientists work hard in this field to prevent these severe effects. The best effort is done to Alert people in time because a wrong alert will cause unnecessary losses. Certainly, people cannot stop the occurrence of earthquakes, but they can adopt protective measures and precautions to minimize the deleterious effects by predicting earthquake magnitude using machine learning techniques. There are a lot of methods that can be applied to predict earthquake magnitude using different sensors, devices, magnetic and electrical waves, or seismic indicators obtained from the processing of the historical data of earthquakes [1]. Really, there is no perfect model that results from 100%, but at least a trial is made to improve the accuracy as much as possible [3].

Artificial intelligence plays an essential role in predicting and classifying problems. A neural network is a very effective tool to solve complicated non-linear issues [1]. Technology and machine learning provide many robust mechanisms to study seismic data and indicators. Data mining and machine learning are highly successful instrument in the prediction domain, especially if massive data is required as weather forecasting, stock prediction, and so on [5]. Dataset plays an essential role in determining the purposed model's quality and performance, so we search a lot to choose the more exact dataset [4].

There are thousands of machine learning algorithms, but no one is always suitable for all issues because there are many factors that affect that, like the number of features (input indicators), number of records of the dataset, and type of problem (classification or regression), so we will apply some machine learning algorithms and compare their results with each other to determine the more suitable algorithm for this issue. The comparison will be applied here to compare our work with other researchers that applied different machine learning algorithms on the same dataset [6]. In this article, the magnitude will be predicted using historical records of south California. Depending on the seismic indicators as an input for machine learning algorithms, which are obtained from performing some statistical and mathematical equations on raw data (time and magnitude) [7]. Machine learning techniques can predict well if provided by a suitable dataset, which is divided into (70%for train and 30%for test) [8]. In this paper, ELM, LS-SVM, and FPA will be used. ELM (Extreme Learning Machine) is a straightforward algorithm because it depends on a feed-forward neural network. The data moves in one direction (forward direction) with only one hidden layer. ELM is a well-known and well-resulted algorithm in classification and regression domains. Unlike other ANN algorithms that face the overfitting problem and the long-running time problem, ELM overcomes these problems and can achieve high accuracy and high-speed results. ELM often uses a sigmoid function, which will be applied in this work.

SVM (Support Vector Machine) is a well-known deep learning tool for regression that depends on kernel methods giving high prediction results. SVM is a developed algorithm of machine learning to avoid ANN shortages. SVM gives optimal global solutions avoiding local minima problems [9].

LS-SVM (Least Square Support Vector Machine) is the edited version of the SVM algorithm. LS-SVM simplifies the SVM method, but it needs kernel parameters, which are important for regression problems. So we need an appropriate way for the optimal choice of LS-SVM parameters, So optimizing LS-SVM with FPA is done [9].

FPA (Flower Pollination Algorithm) was developed in 2012 by Yang, which uses pollination of flowers. FPA was applied for many non-linear problems giving high results [9].

ELM and LS- SVM are optimized by FPA (Flower Pollination Algorithm), enhancing the accuracy and minimizing errors. ELM and LS-SVM are considered supervised learning where they take seven seismic parameters (time is taken during the occurrence of n seismic indicators (T), average magnitude during n events (M_mean), magnitude deficit, which is the difference between the observed magnitude and expected one (∆ M), the curve slope for n events using inverse power law of Gutenberg Richter (b), mean square deviation for n events using inverse power law of Gutenberg Richter (η), the square root of the released energy during T time (DE1/2)), average (mean) time between events (μ) [10] as network inputs and produces the future magnitude as a network output to train well and be able to test after that. No algorithm can predict 100%, but some shortages of ELM algorithm will be removed by using optimization (swarm intelligence) algorithm FPA to get FPA _ELM (the result of optimizing ELM with Flower Pollination algorithm, which optimizes weights of input nodes and biases of a hidden layer of ELM). The optimization of SVM with FPA is applied to enhance the performance of LS-SVM and get the FPA-LS-SVM model that yields high results.

This work is divided into four parts; the first part is processing the raw data to find the seismic indicators. The second part is to input these indicators into machine learning algorithms to predict or output the expected future magnitude. The third part is to optimize the output using artificial intelligence techniques (swarm intelligence). The fourth part is comparing different results using different algorithms.

In the rest of this paper, the related work is showed in Section II, Applied Algorithms in Section III. Data and parameters calculations in Section IV, methodology in Section V, Results and discussion in Section VI, conclusion and future work in Section VII, and finally the references in machine learning and earthquake fields that is helpful in this work.

II. Related Work

Of Course, the earthquake catastrophe takes a large space of scientists' interest, so there are many kinds of research and scientists' efforts which are spent in this domain. Researchers worldwide do their best to predict where and when the Earthquake occurs depending on seismic indicators and other seismic electric signals using machine learning techniques and optimization algorithms.

Adeli et al. applied the probabilistic neural network on seismic indicators to predict the earthquake magnitude. This model predicts well for magnitude from 4.5 to 6 in the southern California region [11]. Moustra et al. who use the artificial neural network to predict earthquake occurrence using time series data and seismic electric signals in the Greece region, there are two case studies which finally led to when appropriate data presented to NN, it can predict accurately [12]. Hegazy et al. optimized ELM with FPA (flower pollination algorithm), which shows a better accuracy when applied to prediction task [13]. In this paper, Ma et al. proposed a genetic algorithm-based neural network giving GA_NN model is applied to six seismic indicators to find the relation between these indicators and the maximum earthquakes in china [14]. Maceda et al. here SVM is applied to earthquake problem. This paper decided that SVM is perfect for solving classification problems using a small-size training dataset [15]. Li et al. depended on the collected data from earthquake stations to be used in machine learning techniques to distinguish between Earthquake and non-earthquake [16]. Rajguru et al. optimized ANN (artificial neural network) with GA (genetic algorithm) to predict the source location and dimensions of the earthquakes giving good results although the complexity of the problem [17]. Wang et al. introduced a deep learning algorithm (LSTM) to find the relation among the earthquakes in different places that is useful in earthquake prediction [18]. Reyes et al. used an ANN model to predict earthquake magnitude in a limited interval or bounded by threshold during the following five days. Applying statistical tests and experiments showed the higher success rate of that method than other machine learning classifiers [19]. Martínez-Álvarez et al. used various seismic indicators as inputs for ANN in different seismic zones, and this showed that the seismic indicators are the best features for earthquake prediction [20]. Zhou et al. showed how efficiently the neural network could predict the earthquake magnitude using the LM_BP algorithm [21]. Morales-Esteban et al. used statistical methods and applied ANN on Earthquake in two seismic regions The Western Azores–Gibraltar fault and the Alborán Sea whether the magnitude is limited in an interval or determined by threshold [22]. Wang et al. proposed a RBF(Radial Basis Function) neural network to earthquake prediction, and the results showed that RBF is an effective tool for that non-linear problem [23]. Alarifi et al. used ANN for earthquake prediction in the Red Sea region, and the results showed the high ability of ANN forecasting than any other statistical models [24]. Asencio–Cortés et al. proposed a machine learning model to predict the earthquake magnitude after the next seven days using cloud infrastructure [25]. Tan et al. proposed support vector machines (SVM) and the results were good and added a new method for predicting earthquakes [26]. Florido et al. interested in processing earthquake data in

Chile. The data were labeled by applying to clustering algorithms. It was easy to identify Earthquakes with a magnitude larger than 4.4 [27]. Last et al. used many data mining methods and time series to predict the magnitude of the highest earthquake event in the following year depending on data from the previous records using a multi-objective info-fuzzy network algorithm [28]. Rafiei et al. was interested in giving early alarm weeks before maximum earthquake event using classification algorithm (CA) combining with mathematical optimization algorithm (OA) whose role is to find the location of the Earthquake with maximum magnitude [29]. Kerh et al. used a genetic algorithm combined with a neural network to evaluate the response of the Earthquake in Taiwan that produces high results comparing to neural network model only [30]. Mirrashid et al. used the system of adaptive neuro-fuzzy inference to predict the coming earthquakes with magnitude 5.5 or higher developed by three algorithms subtractive clustering (SC), grid partition (GP), and fuzzy C-means (FCM). The results showed that the ANFIS-FCM model gives the highest accuracy in predicting the magnitude of the Earthquake [31]. Asim et al. used many models of computational intelligence for earthquake prediction in northern Pakistan, and the feed-forward neural network showed the highest performance compared with other models [32]. Umar et al. combined the Logistic Regression (LR) with Frequency Ratio (FR) to overcome the limitations of each one individually. This model achieved high success and good earthquake prediction in Indonesia [33]. Asim et al. focused on computing the seismic indicators mathematically and then applying the tree classifiers on them to predict the earthquake magnitude in Hindu Kush, showing that the rotation forest gave a good prediction than random forest [34].

Machine learning (ML) methods such as artificial neural networks (ANNs), support vector machines (SVMs), extreme learning machine (ELM), are considered the most commonly used ML models in classification, regression. But these methods may suffer from local minima and overfitting problems due to using local optimization training algorithms such as gradient descent algorithm in ANN [35]. Swarm Intelligence algorithms such as particle swarm optimization (PSO), follower pollination algorithm (FPA), ant colony optimization (ACO), and artificial bee colony (ABC), can solve the problems or drawbacks of machine learning models such as ANN, SVM, and ELM methods [36,37]. Using swarm intelligence or meta-heuristic algorithms in optimizing and training classical machine learning models can enhance the accuracy and generalization ability of these methods [38-43].

III. Preliminaries

A. ELM (Extreme Learning Machine)

This model is a feed-forward neural network with only one hidden layer, a very rapid learning method with solutions for many problems caused by traditional neural network algorithms like overfitting, local minima, slowness, achieving better performance higher accuracy. ELM is a simple model that is performed through three steps.

- First step: ELM chooses the weights of input nodes and hidden biases randomly**.**
- Second step: ELM performs calculations to generate the output matrix of the hidden layer.
- Third step: calculations of the output weight.

ELM uses a single hidden layer with N hidden nodes and f(x) the activation function to learn distinct samples (M) $(\mathbf{x_i}, \mathbf{t_i})$, $\mathbf{x_i}=[\mathbf{x_{11}}, \mathbf{x_{12}} \ldots. \mathbf{x_{1k}}]^{\mathbf{T}} \in \mathbf{R^k}$.

$\mathbf{t_i} = [\mathbf{t_{11}}, \mathbf{t_{12}} \ldots. \mathbf{t_{1d}}]^{\mathbf{T}} \in \mathbf{R^{kd}}$ after that, the non-linear problem is turned to the linear problem:

$$H\beta=T \tag{1}$$

the hidden output matrix H is defined as follow:

$$H=\begin{bmatrix} \mathbf{f}(\mathbf{w_1} . \mathbf{x_1} + \mathbf{b_1}) & \cdots & \mathbf{f}(\mathbf{w_N} . \mathbf{x_1} + \mathbf{b_N}) \\ \vdots & \ddots & \vdots \\ \mathbf{f}(\mathbf{w_1} . \mathbf{x_M} + \mathbf{b_1}) & \cdots & \mathbf{f}(\mathbf{w_N} . \mathbf{x_M} + \mathbf{b_N}) \end{bmatrix} \tag{2}$$

Where $\mathbf{w_j} = [\mathbf{w_{j1}}, \mathbf{w_{j2}} \ldots. \mathbf{w_{jk}}]^{\mathbf{T}}$ is the weight vector between input nodes and hidden ones, **b** is the bias of hidden nodes where j= 1,2 ,..N.

$\mathbf{w_j} . \mathbf{x_i}$ where i=1…M is the inner product of $\mathbf{w_j}$, $\mathbf{x_i}$. β= $[\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3} \ldots. \boldsymbol{\beta_N}]^{\mathbf{T}}$ is output weights' matrix, but $\boldsymbol{\beta_j} = [\boldsymbol{\beta_{j1}}, \boldsymbol{\beta_{j2}}, \boldsymbol{\beta_{j3}} \ldots. \boldsymbol{\beta_{jd}}]^{\mathbf{T}}$ where j= 1,2,3…..N is the vector of weights connecting the output neuron and a jth hidden one. $\mathbf{T}=[\mathbf{t_1}, \mathbf{t_2}, \mathbf{t_3} \ldots. \mathbf{t_M}]^{\mathbf{T}}$ is targets' matrix. The errors between the estimated value $\mathbf{t_i}$ and the real value $\mathbf{t_i}$ equal to zero using the following formula:

$$\mathbf{t_i} = \sum_{\mathbf{j=1}}^{\mathbf{N}} \boldsymbol{\beta_j}\ \mathbf{f}(\mathbf{w_j} . \mathbf{x_i} + \mathbf{b_j}) \tag{3}$$

weights that link between the hidden layer and output one are estimated by the least square solution to the linear problem using the following formula:

$$B=\mathbf{H^{-1}}T \tag{4}$$

Where $\mathbf{H^{-1}}$ is the inverse of the H matrix using the Moore Penrose method that makes ELM perform better and faster [13].

B. LS-SVM (Least Square Support Vector Machine)

It is one of deep and supervised learning that can classify data and predict values LS-SVM is the new version of SVM that solves SVM issues. Using LS-SVM, the solution can be founded by solving some linear equations rather than a Quadratic programming problem used in SVM. Let x is the matrix of input data. Using training data, LS-SVM plays a good role in creating the function that shows the dependence of the output on the input. The formula of this function:

$$f(x)=\mathbf{W^T}\ \varphi\ (x)+ b \tag{5}$$

where W, φ (x): $\mathbf{R^p} \rightarrow \mathbf{R^n}$ are column vectors with size n*1, and b ∈ **R**. Now LS-SVM calculates that function using the same minimization problem in SVM. Showing the important difference that LS-SVM contains equality constraints, unlike inequality ones in SVM. LS-SVM depends on the least square function [9]. We can minimize the error using this optimization formula:

$$\min j(w,e,b)= \frac{\mathbf{1}}{\mathbf{2}} \mathbf{w^t} w + C \frac{\mathbf{1}}{\mathbf{2}} \mathbf{e^t} \tag{6}$$

$$\mathbf{y_i}=\mathbf{W^T}\ \varphi\ (\mathbf{x_i})+ \mathrm{b} +\mathbf{e_i} \qquad (7)$$

Where e is n*1 column vector, C ∈ $\mathbf{R^+}$ is the parameter between training errors and solution size. In 2 Lagranian is formed that differs according to w,b,e, a where a is Largrangian multiplier, we have.

$$\begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & -\mathbf{Z^T} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{1^T} \\ \mathbf{0} & \mathbf{0} & \mathbf{C} & -\mathbf{I} \\ \mathbf{Z} & \mathbf{1} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{b} \\ \mathbf{e} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{y} \end{bmatrix}. \qquad (8)$$

Identity matrix and Z= $[\varphi\ (\mathbf{x_1}), \varphi\ (\mathbf{x_2}), \ldots \varphi\ (\mathbf{x_n})]^{\mathbf{T}}$

Using four from row 1, we get W= $\mathbf{Z^T}$and from row 3, we get Ce=a.

Note that: the kernel matrix K = Z $\mathbf{Z^T}$And the parameter λ = $\mathbf{C^{-1}}$, the conditions for optimality give the following solution

$$\begin{bmatrix} \mathbf{0} & \mathbf{1^T} \\ \mathbf{1} & \mathbf{K}+\lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{y} \end{bmatrix}. \qquad (9)$$

Types of the kernel function K:

First: linear kernel

$$\mathrm{K}(\mathrm{x}, \mathbf{x_i}) = \mathbf{x_i}^{\mathbf{T}}\mathbf{x}$$

Second: the polynomial kernel of d degree

$$\mathrm{K}(\mathrm{x}, \mathbf{x_i}) = (\mathbf{1}+\mathbf{x_i}^{\mathbf{T}}\mathbf{x}/\mathbf{c})^{\mathbf{d}}. \qquad (10)$$

Third: RBF kernel (radial basis function)

$$\mathrm{K}(\mathrm{x}, \mathbf{x_i}) = \mathbf{exp}(-\|\mathbf{x}-\mathbf{x_i}\|^{\mathbf{2}}/\ \sigma^{\mathbf{2}}). \qquad (11)$$

Forth: MLP kernel

$$\mathrm{K}(\mathrm{x}, \mathbf{x_i}) = \mathbf{tanh}(\mathbf{kx_i}^{\mathbf{T}}\mathbf{x}+\theta). \qquad (12)$$

C. FPA (Flower Pollination Algorithm) is an optimization technique that uses the idea of pollination of flowers. The pollination process has two directions: biotic pollination, which occurred in 90% of flowers by the spread of pollens using various insects, and abiotic pollination, which occurred in 10% of flowers where the pollens move through water or wind. Two methods do the pollination Self-pollination: occurs when the flower's pollen moves to the same flower or different flowers at the same plant.

Cross-pollination: occurs when the pollen of one flower of a specified plant moves to another flower in a different plant. The steps of the FPA algorithm:

- The cross-pollination (biotic) is called global pollination, where the Levy fight takes place
- Self-pollination is considered as local pollination
- The constancy of the flower can be treated as the probability of reproduction is proportional to two flowers similarity
- There is a switch probability p ∈ {0,1} that determines the pollination process, even local or global.

In global pollination, the pollinators can carry pollens for long distances. This ensures creating the fittest (Best) that obeys levy distribution:

$$x_i^{t+1}= x_i^{t}+ \mathrm{L}(\mathrm{Best}-x_i^{t}) \qquad (13)$$

In local pollination, the pollen can be carried by other factors, and this can be represented using the formula:

$x_i^{t+1}= x_i^{t}+ \mathrm{U}(x_j^{t}-x_k^{t})$ (14) [13],[9]

```
Algorithm 1: Flower Pollination Algorithm (FPA)
1) Initialization of:
    • Population size (n) and the flower population Xi
      (i = 1, 2,..., n) randomly chosen solutions.
    • switch probability p ∈ {0,1}
    • Maximum iterations' number.
    • The best solution in the initial population (Best)
    • Function (fn) applied to each flower
2) randomly initial population generated.
3) while (t < Max_iteration)
    for each flower do
      if (rand<p)
          Draw L (d-dimensional step vector) that
          undergoes a Levy distribution.

          Global pollination for the solution i
          $x_i^{t+1}= x_i^{t}+ L(Best-x_i^{t})$
      Else
          Draw U from uniform distribution in
          [0,1].

          From all the solutions, choose □, k
          randomly.

          Local pollination for the solution i
          $x_i^{t+1}= x_i^{t}+ U(x_j^{t}-x_k^{t})$.
      endif
      Calculate the fitness of each flower.

      Pass each flower(solution) to fn to
      evaluate the new solution.
    end for
        find Best (current best solution)
end while
```

IV. Data and Parameters Calculations

Seven parameters are mathematically and statistically calculated during a specific period of time [10] from the earthquake catalog. These parameters are the inputs for the network to predict the future expected magnitude as the output of the network.

1) Earthquake data: The Earthquake catalog source for south California is available to be downloaded for free using the website (www.data.scec.org.). The historical earthquake data of Southern California is between.

1st January 1950 and 31st May 1978 is divided into 693 periods. Each period consists of fifteen days.

2) Seismic parameters: The seven earthquake indicators are computed for each period of time. This part will show these indicators and their mathematical calculations. The first indicator is the time during the range of n events, which called T.

$$T = tn- t1 \qquad (15)$$

T1 is the time of the first event, and tn is the time for the period's nth event.

The second seismic parameter is the average magnitude of the last n events of the period, which is calculated as the following.

$$M\_mean=\frac{\sum Mi}{n} \qquad (16)$$

Mi is the magnitude of the ith event, and n is the number of events.

The third parameter is $DE^{1/2}$ the square root of released seismic energy is during time T. The $DE^{1/2}$ the parameter can be computed as follow:

$$DE^{1/2} = \frac{\sum \sqrt{E}i}{T} \qquad (17)$$

$\sqrt{E}i$ is the square root of the seismic energy (E) of the ith event where E can be calculated from the following formula:

$$Ei=10^{(11.8+1.5Mi)}ergs \qquad (18)$$

The fourth indicator is b_value, that is, the slope of the curve between the log of frequency of occurred earthquakes and the earthquake magnitudes given from Richter inverse power law.

$$\log_{10} N = a – bM \qquad (19)$$

N is the number of events with M magnitude a,b values are constants.

where b can be calculated as follow

$$b= \frac{(n\sum(Mi \log Ni)-\sum Mi \sum \log Ni)}{((\sum Mi)^2-n\sum Mi^2)} \qquad (20)$$

and a can be calculated from the following formula:

$$a= \frac{\sum(\log_{10} Ni+bMi)}{n} \qquad (21)$$

The fifth parameter is (η value), which is the sum of mean square deviation based on inverse power law. η value can be computed as follow.

$$\eta=\frac{\sum(\log_{10} Ni-(a-bMi))^2}{n-1} \qquad (22)$$

The sixth parameter is the ∆M value, defined as the magnitude deficit (the difference between the observed magnitude and the expected one). ∆M is computed as follow:

$$\Delta M= M_{observed} - M_{excepected} \qquad (23)$$

Where $M_{excepected}$ is calculated from the formula:

$$M_{excepected}={}^{a}/_{b} \qquad (24)$$

The last one μ in the meantime among the characteristic events which is calculated as follow:

$$\mu = \frac{\sum(t_{i\ characteristic})}{n_{characteristic}} \qquad (25)$$

[10].

Now a sample of the dataset from the period from 1st January 1950 to 30th May 1950 that represents ten periods of time will be presented. Each period consists of fifteen days as shown in Table I.

V. Proposed Models

The proposed methods depend on the study of historical earthquake data in earthquake catalogs. By processing these data, seismic indicators that are used as inputs for the network can be obtained. Then the ELM algorithm is optimized by FPA to enable us from an optimal prediction of the occurrence of the Earthquake, and also optimizing LS-SVM with FPA to enhance the accuracy of earthquake magnitude prediction. The network architecture contains seven input indicators, which represent the seismic indicators, and the output shows the predicted magnitude during fifteen days. The description of proposed FPA-ELM, and FPA-LS-SVM algorithms are shown in algorithm 2, and algorithm 3, respectively. The data is used in three manners. First, data is divided into 70% for training and 30% for testing. Then, data is divided into 80% for training and 20% for testing. At last data is divided into 90% for training and 10% for testing. The phases of the used models are shown in Fig. 1. The earthquake indicators which have been calculated from the datasets are shown in Table I.

TABLE I. EARTHQUAKE INDICATORS SAMPLE

time period	**input seismic indicators**							**output**
	T	DE1/2	b	H	∆M	M_mean	μ	target_mag
1/1/1950-15/1/1950	3655	1197086.433	0.585242	0.006859	0.437819	2.9305	182	3.37
16/1/1950-30/1/1950	3668	816522.7778	0.663648	0.005888	0.307195	2.59	135	2.16
31/1/1950-14/2/1950	3684	294005.8278	0.469148	0.013345	0.497108	2.255333	245	1.65
15/2/1950-1/3/1950	3699	1205497.389	0.569933	0.012245	0.034071	2.946316	194	3.05
2/3/1950-16/3/1950	3714	231294.1789	0.588185	0.009469	0.266256	2.6675	464	2.68
17/3/1950-31/3/1950	3730	483924.9513	0.566164	0.022646	0.52207	2.705333	248	2.84
1/4/1950-15/4/1950	3743	848726.3986	0.460094	0.002465	0.068989	2.713333	249	2.14
16/4/1950-30/4/1950	3758	437602.1004	0.684234	0.018921	0.311847	2.65625	234	2.44
1/5/1950-15/5/1950	3775	558145.6542	0.739229	0.026896	0.555382	2.882667	251	3.34
16/5/1950-30/5/1950	3789	534786.1837	0.52383	0.019356	0.545958	2.817692	291	3.68

Algorithm2: The Hybrid Earthquake prediction FPA_ELM

```
Input:   n number of flowers
         N_gen number of iterations
         d dimensions of search variables
         fn function applied to each flower
         Trn Training, Vld Validation datasets
Output:
         fbest optimal hidden weights and biases
         f(fbest) Sum square error for the NN
         over the validation set fbest
1) Initialization:
   • N,N_gen
   • switch probability $p \in \{0,1\}$
   • Function (fn) applied to each flower
2) randomly initial population generated.
3) while Stopping criteria not met, do
   for each flower, do
      if (rand<p)
         Draw L (d-dimensional step vector)
         that undergoes a Levy distribution.

         Global pollination for a solution i
         $x_i^{t+1}= x_i^t+$ L(Best-$x_i^t$)
      Else
         Draw U from a uniform distribution
         In [0,1]

         from all the solutions choose $j,k$
         randomly

         Local pollination for a solution i
         $x_i^{t+1}= x_i^t+$ U($x_j^t$-$x_k^t$).
      endif
      Calculate the fitness of each flower

      Pass each flower(solution) to fn to
      evaluate the new solution

      construct NN given the hidden layer
      weights and biases of flower i

      Calculate the output layer's weights using
      the MP matrix using training dataTrn and
      hidden layer weights and Biases.

      Use the validation data Vld to evaluate the
      NN model
      Compute prediction accuracy
   end for
end while
4) Pass Best to fn
5) Apply ELM.
6) Return prediction accuracy
```

Algorithm3: The hybrid FPA_SVM for Earthquake prediction

```
Input:   n number of flowers
         N_gen number of iterations
         Ub data Upper bound
         Lb  data Lower bound
         d dimensions of search variables
         fn function that applies to each flower to
         improve SVM parameters.
output:  prediction accuracy
         Best global solution
1) Initialization:
   • Population size (n) and the flower population
     Xi (i = 1, 2..., n) randomly chosen solutions.
   • switch probability $p \in \{0,1\}$
   • Maximum iterations' number.
   • Data limits Lb, Ub with dimension d.
   • The best solution in the initial population
     (Best)
   • Function (fn) applied to each flower
2) randomly initial population generated.
3) while (t < Max_iteration)
   for each flower do
      if (rand<p)
         Draw L (d-dimensional step vector)
         that undergoes a Levy distribution.

         Global pollination for a solution i
         $x_i^{t+1}= x_i^t+$ L(Best-$x_i^t$)
      Else
         Draw U from a uniform distribution
         in [0,1].

         From all the solutions choose $j$,
         randomly.

         Local pollination for a solution i
         $x_i^{t+1}= x_i^t+$ U($x_j^t$-$x_k^t$).
      endif
      Calculate the fitness of each flower.

      Pass each flower(solution) to fn to
      evaluate the new solution

      Pass output to SVM parameters

      Apply SVM

      Compute prediction accuracy
   end for
   t=t+1
end while
4) Pass Best to fn
5) Pass output to SVM parameters.
6) Apply SVM.
7) Return prediction accuracy.
```

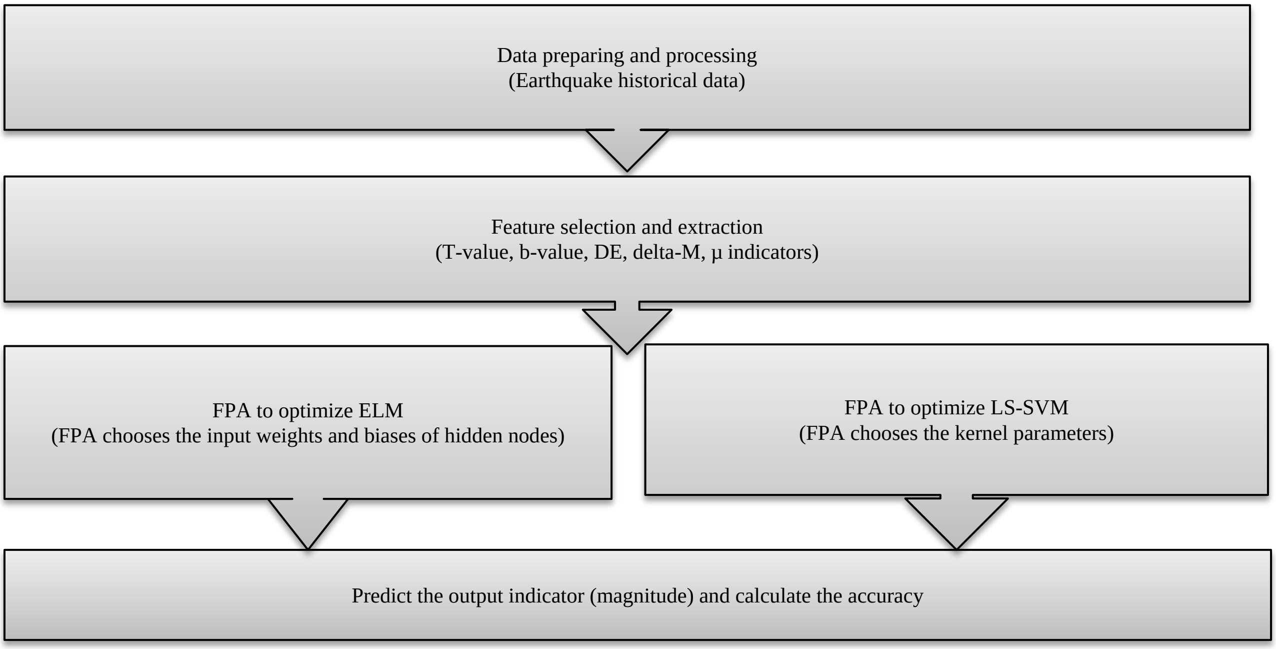

Data preparing and processing
(Earthquake historical data)
Feature selection and extraction
(T-value, b-value, DE, delta-M, µ indicators)
FPA to optimize ELM
(FPA chooses the input weights and biases of hidden nodes)
FPA to optimize LS-SVM
(FPA chooses the kernel parameters)
Predict the output indicator (magnitude) and calculate the accuracy

Fig. 1. The Phases of Proposed Model.

VI. Results and Discusssion

After processing data and introducing the indicators to the proposed models, the performance of each model is estimated using four performance evaluation criteria RMSE, MAE, SMAPE, and PMRE, which can be calculated through the following formulas [13]:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{Ai}-\text{Fi})} \tag{26}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\text{Ai}-\text{Fi}| \tag{27}$$

$$\text{SMAPE}=\frac{\sum_{i=1}^{n}|\text{Fi}-\text{Ai}|}{\sum_{i=1}^{n}\text{Ai}+\text{Fi}} \tag{28}$$

$$\text{PMRE}=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{\text{Ai}-\text{Fi}}{\text{Fi}}\right| \tag{29}$$

First data is divided into 70% for training and 30% for testing and the results showed that the accuracy of FPA-ELM is higher comparing to solely using ELM, as shown in Fig. 2.

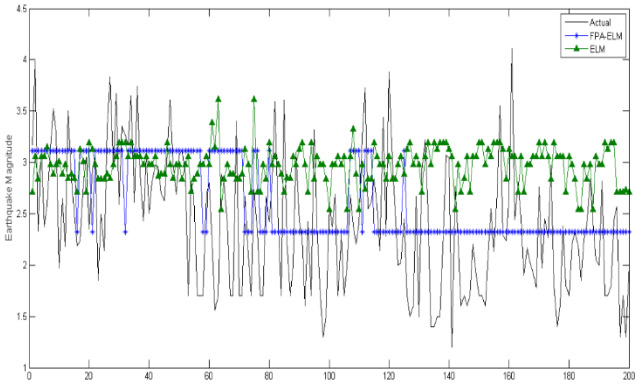


Fig. 2. ELM Accuracy VS FPA-ELM Accuracy when Data is divided into 70% for Training and 30% for Testing.

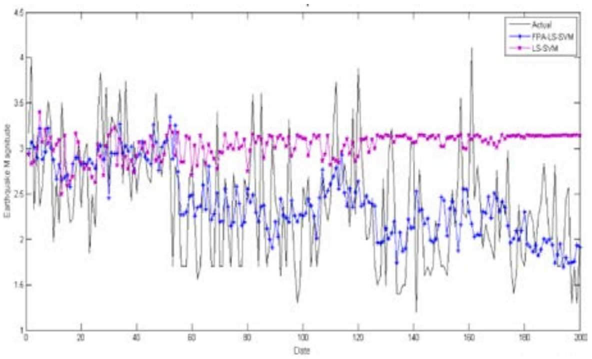


Fig. 3. LS-SVM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 70% for Training and 30% for Testing.

And the performance of LS-SVM became higher and better after optimizing LS-SVM with FPA as shown in Fig. 3.

After the experiment, the results showed that the performance of LS-SVM is better than the performance of ELM, according to this, FPA-LS-SVM accuracy is higher than FPA-ELM one. This is obvious from the following Fig. 4.

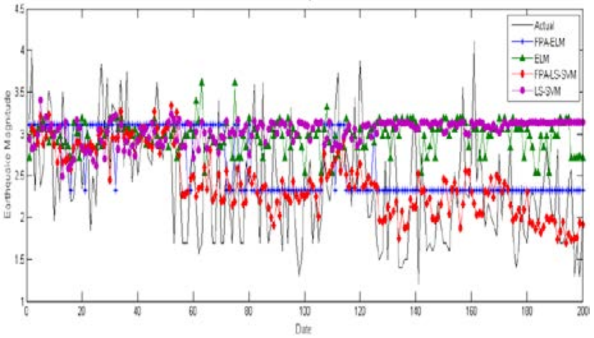


Fig. 4. FPA-ELM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 70% for Training and 30% for Testing.

TABLE II. EVALUATION CRITERIA FOR THE MODELS WHEN DATA IS DIVIDED INTO 70% FOR TRAINING AND 30% FOR TESTING

evaluation test	**Models**			
	ELM	**FPA- ELM**	**LS-SVM**	**FPA-LS-SVM**
RMSE_test	0.88057	0.645895	0.93642	0.540368
MAE_test	0.73833	0.525306	0.789076	0.434484
SMAPE_test	0.02346	0.030137	0.022373	0.034936
PMRE_test	38.0774	25.12757	40.83974	19.47929

The values of evaluation tests were as follow in Table II:

The column chart that provides the rate of RMSE evaluation criteria for all models when data is divided into 70% for training and 30% for testing is applied as shown in Fig. 5.

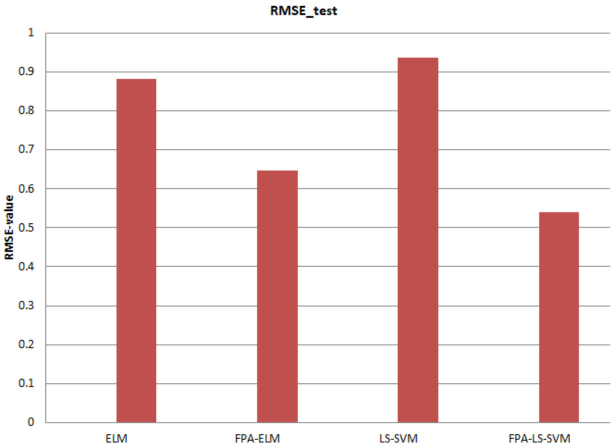


Fig. 5. RMSE Evaluation Criteria for All Models when Data is divided into 70% for Training and 30% for Testing.

The column chart that provides the rate of MAE evaluation criteria for all models when data is divided into 70% for training and 30% for testing is applied as shown in Fig. 6.

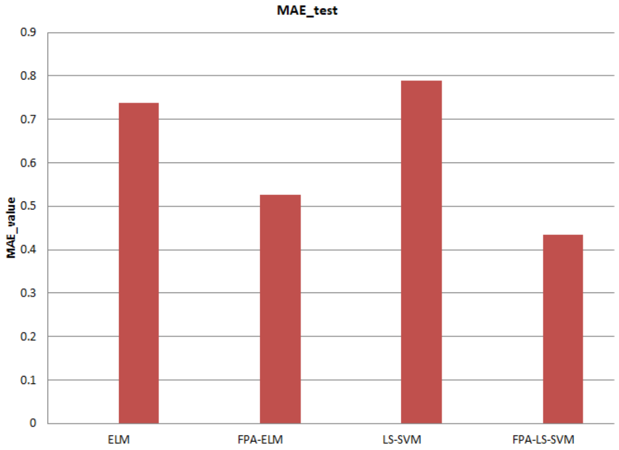


Fig. 6. MAE Evaluation Criteria for All Models when Data is divided into 70% for Training and 30% for Testing.

The column chart that provides the rate of SMAPE evaluation criteria for all models when data is divided into 70% for training and 30% for testing is applied as shown in Fig. 7.

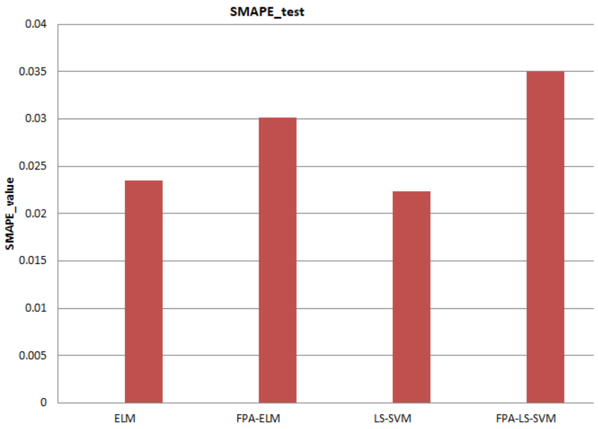


Fig. 7. SMAPE Evaluation Criteria for All Models when Data is divided into 70% for Training and 30% for Testing.

The column chart that provides the rate of PMRE evaluation criteria for all models when data is divided into 70% for training and 30% for testing is applied as shown in Fig. 8.

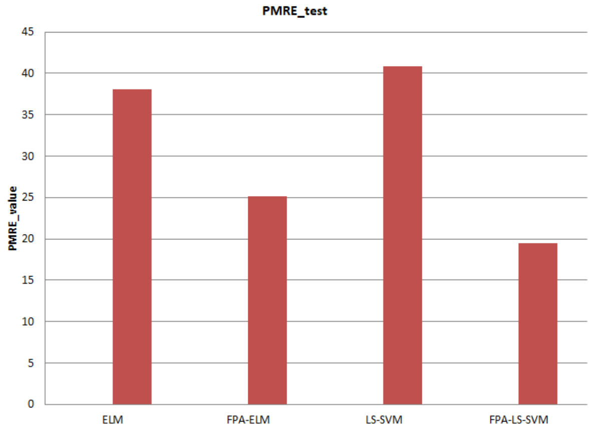


Fig. 8. PMRE Evaluation Criteria for All Models when Data is divided into 70% for Training and 30% for Testing.

Second, data is divided into 80% for training and 20% for testing and the results also showed that the accuracy of FPA-ELM is higher comparing to solely using ELM, as shown in Fig. 9.

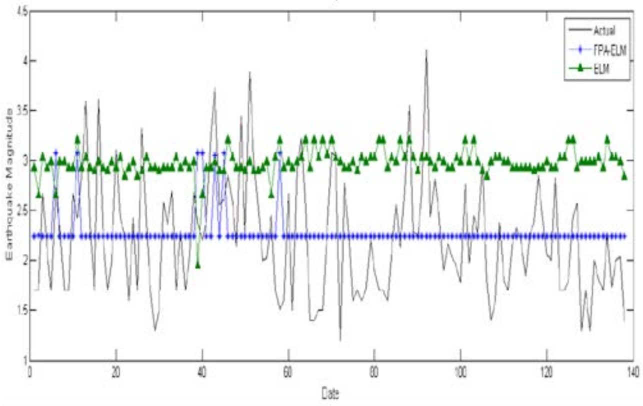


Fig. 9. ELM Accuracy VS FPA-ELM Accuracy when Data is divided into 80% for Training and 20% for Testing.

And the performance of LS-SVM became higher and better after optimizing LS-SVM with FPA as shown in Fig. 10.

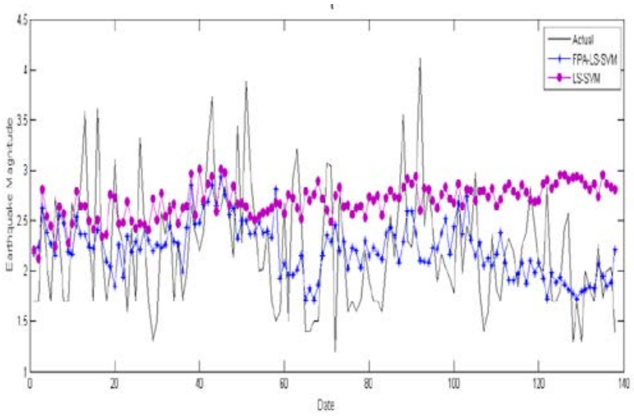


Fig. 10. LS-SVM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 80% for Training and 20% for Testing.

And FPA-LS-SVM accuracy is higher than FPA-ELM one. This is obvious from the following Fig. 11.

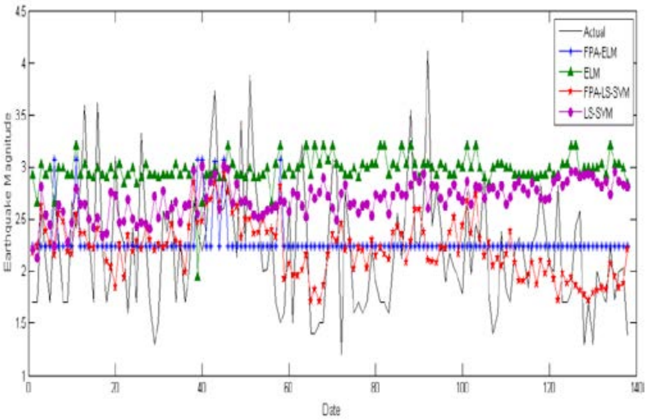


Fig. 11. FPA-ELM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 80% for Training and 20% for Testing.

In this case the values of evaluation criteria were as follow in Table III:

TABLE III. EVALUATION CRITERIA FOR THE MODELS WHEN DATA IS DIVIDED INTO 80% FOR TRAINING AND 20% FOR TESTING

evaluation test	Models			
	ELM	FPA-ELM	LS-SVM	FPA-LS-SVM
RMSE_test	0.962142	0.598239	0.769138	0.565476
MAE_test	0.845038	0.4836	0.649227	0.447226
SMAPE_test	0.03279	0.045286	0.03779	0.046786
PMRE_test	45.68055	23.49402	34.71175	20.931907

The column chart that provides the rate of RMSE evaluation criteria for all models when data is divided into 80% for training and 20% for testing is applied as shown in Fig. 12.

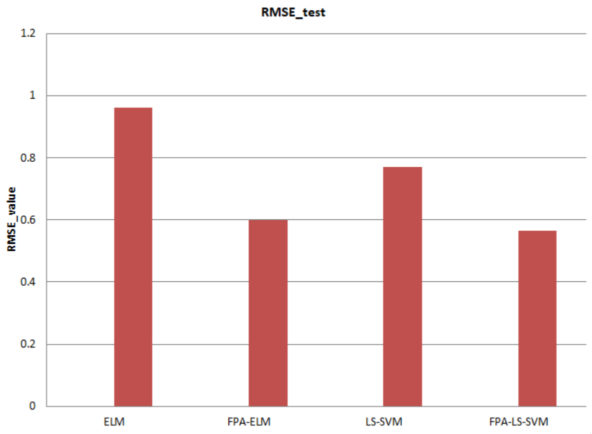


Fig. 12. RMSE Evolution Criteria for All Models when Data is divided into 80% for Training and 20% for Testing

The column chart that provides the rate of MAE evaluation criteria for all models when data is divided into 80% for training and 20% for testing is applied as shown in Fig. 13.

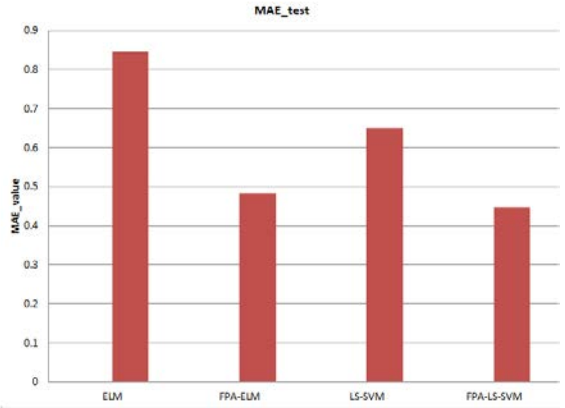


Fig. 13. MAE Evaluation Criteria for All Models when Data is divided into 80% for Training and 20% for Testing.

The column chart that provides the rate of SMAPE evaluation criteria for all models when data is divided into 80% for training and 20% for testing is applied as shown in Fig. 14.

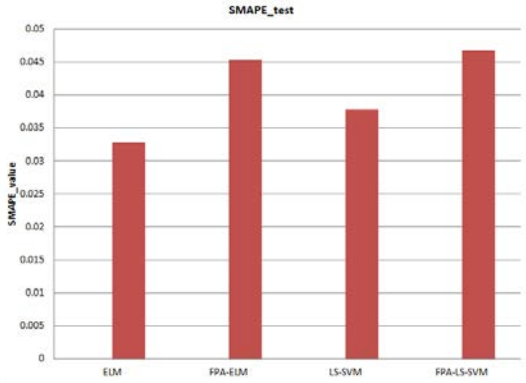


Fig. 14. SMAPE Evaluation Criteria for All Models when Data is divided into 80% for Training and 20% for Testing.

The column chart that provides the rate of PMRE evaluation criteria for all models when data is divided into 80% for training and 20% for testing is applied as shown in Fig. 15.

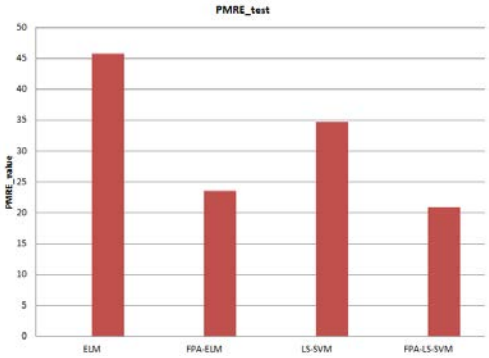


Fig. 15. PMRE Evaluation Criteria for All Models when Data is divided into 80% for Training and 20% for Testing.

At last data is divided into 90% for training and 10% for testing and the results also showed that the accuracy of FPA-ELM is higher comparing to solely using ELM, as shown in Fig. 16.

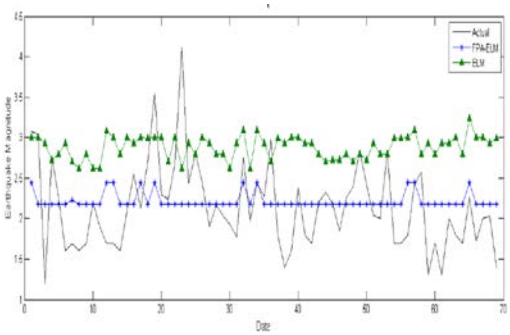


Fig. 16. ELM Accuracy VS FPA-ELM Accuracy when Data is divided into 90% for Training and 10% for Testing.

And the performance of LS-SVM became higher and better after optimizing LS-SVM with FPA as shown in Fig. 17.

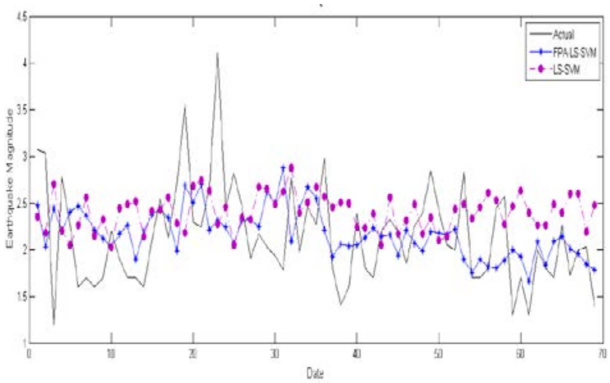


Fig. 17. LS-SVM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 90% for Training and 10% for Testing.

And also FPA-LS-SVM accuracy is higher than FPA-ELM one. This is obvious from the following Fig. 18.

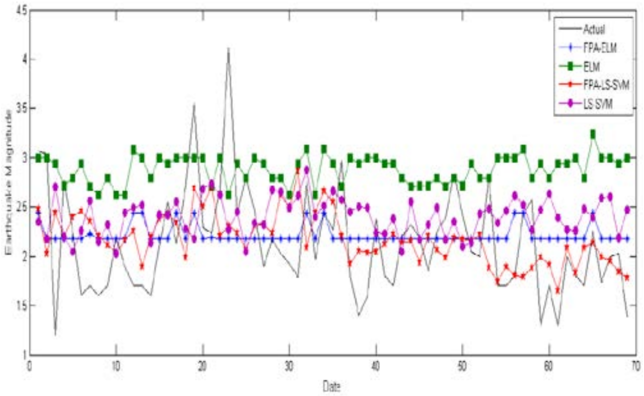


Fig. 18. FPA-ELM Accuracy VS FPA-LS-SVM Accuracy when Data is divided into 90% for Training and 10% for Testing.

In this case the values of evaluation criteria were as follow in Table IV:

TABLE IV. EVALUATION CRITERIA FOR THE MODELS WHEN DATA IS DIVIDED INTO 90% FOR TRAINING AND 10% FOR TESTING

evaluation test	**Models**			
	ELM	**FPA- ELM**	**LS-SVM**	**FPA-LS-SVM**
RMSE_test	0.929831	0.529094	0.651456	0.537101
MAE_test	0.813929	0.417803	0.53825	0.428999
SMAPE_test	0.065945	0.096061	0.082919	0.097089
PMRE_test	44.65868	21.13068	28.38148	20.80531

The column chart that provides the rate of RMSE evaluation criteria for all models when data is divided into 90% for training and 10% for testing is applied as shown in Fig. 19.

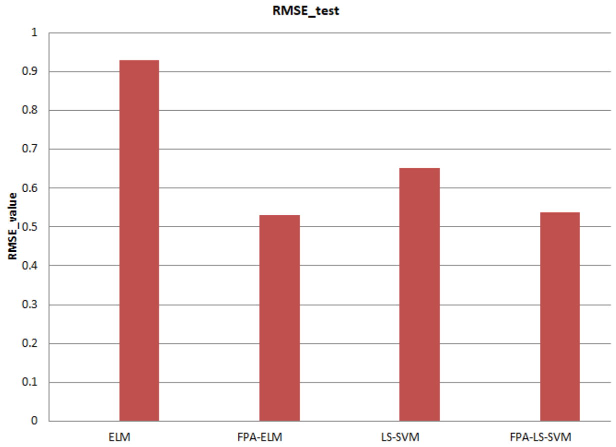


Fig. 19. RMSE Evolution Criteria for All Models when Data is divided into 90% for Training and 10% for Testing.

The column chart that provides the rate of MAE evaluation criteria for all models when data is divided into 90% for training and 10% for testing is applied as shown in Fig. 20.

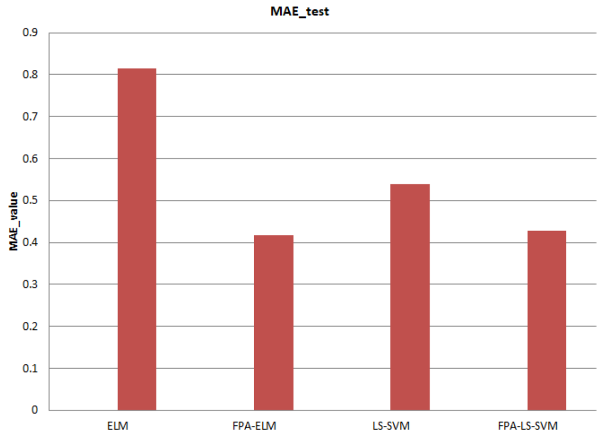


Fig. 20. MAE Evaluation Criteria for All Models when Data is divided into 90% for Training and 10% for Testing.

The column chart that provides the rate of SMAPE evaluation criteria for all models when data is divided into 90% for training and 10% for testing is applied as shown in Fig. 21.

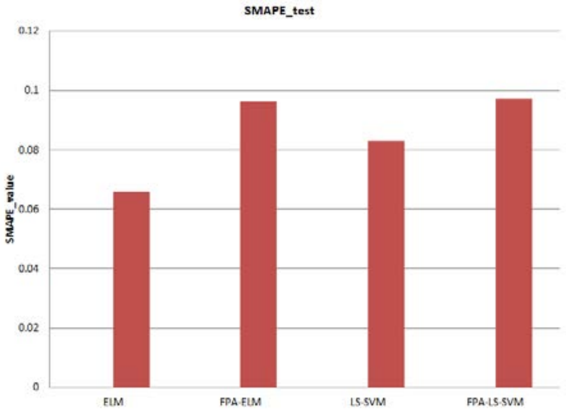


Fig. 21. SMAPE Evaluation Criteria for All Models when Data is divided into 90% for Training and 10% for Testing.

The column chart that provides the rate of PMRE evaluation criteria for all models when data is divided into 90% for training and 10% for testing is applied as shown in Fig. 22.

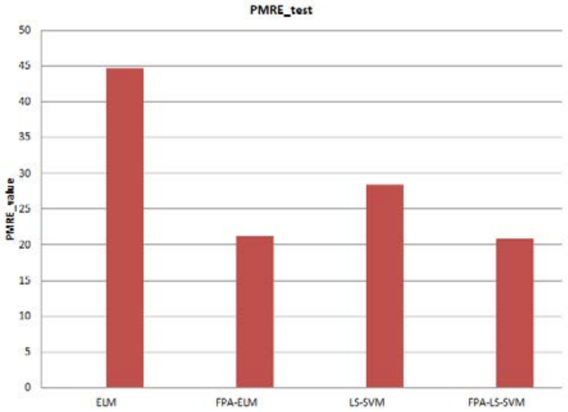


Fig. 22. PMRE Evaluation Criteria for All Models when Data is divided into 90% for Training and 10% for Testing.

VII. Conclusion

In this paper, two hybrid models, FPA-ELM and FPA-SVM, were proposed to forecast earthquake magnitude in the southern California region. Some seismic indicators were generated mathematically and statistically from the dataset to be employed as inputs for the proposed models. The proposed models were evaluated using four criteria. These criteria are RMSE, MAE, SMAPE, and PMRE. The experimental results showed that the accuracy of both ELM and LS-SVM were increased after optimizing it by FPA algorithm. The performance of proposed FPA-LS-SVM outperformed the FPA-ELM model according to all compared criteria. Also, FPA-LS-SVM is the best in reducing the false alarm ratio in earthquake prediction.

References

[1] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, “Earthquake magnitude prediction in Hindukush region using machine learning techniques,” Nat. Hazards, vol. 85, no. 1, pp. 471–486, 2017, doi: 10.1007/s11069-016-2579-3.

[2] G. Asencio-Cortés, F. Martínez-Álvarez, A. Troncoso, and A. Morales-Esteban, “Medium–large earthquake magnitude prediction in Tokyo with artificial neural networks,” Neural Comput. Appl., vol. 28, no. 5, pp. 1043–1055, 2017, doi: 10.1007/s00521-015-2121-7.

[3] T. L. Chin, C. Y. Huang, S. H. Shen, Y. C. Tsai, Y. H. Hu, and Y. M. Wu, “Learn to Detect: Improving the Accuracy of Earthquake Detection,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 8867–8878, 2019, doi: 10.1109/TGRS.2019.2923453.

[4] S. M. Mousavi, Y. Sheng, W. Zhu, and G. C. Beroza, “STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI,” IEEE Access, vol. 7, pp. 179464–179476, 2019, doi: 10.1109/ACCESS.2019.2947848.

[5] R. Mallouhy, C. A. Jaoude, C. Guyeux, and A. Makhoul, “Major earthquake event prediction using various machine learning algorithms,” 6th Int. Conf. Inf. Commun. Technol. Disaster Manag. ICT-DM 2019, pp. 1–7, 2019, doi: 10.1109/ICT-DM47966.2019.9032983.

[6] M. Maya and W. Yu, “Short-term prediction of the earthquake through neural networks and meta-learning,” 2019 16th Int. Conf. Electr. Eng. Comput. Sci. Autom. Control. CCE 2019, pp. 1–6, 2019, doi: 10.1109/ICEEE.2019.8884562.

[7] M. Monterrubio-Velasco, J. C. Carrasco-Jimenez, O. Castillo-Reyes, F. Cucchietti, and J. De La Puente, “A Machine Learning Approach for Parameter Screening in Earthquake Simulation,” Proc. - 2018 30th Int. Symp. Comput. Archit. High Perform. Comput. SBAC-PAD 2018, pp. 348–355, 2019, doi: 10.1109/CAHPC.2018.8645865.

[8] A. Vahaplar, B. T. Tezel, R. Nasiboglu, and E. Nasibov, “A monitoring system to prepare machine learning data sets for earthquake prediction based on seismic-acoustic signals,” 9th Int. Conf. Appl. Inf. Commun. Technol. AICT 2015 - Proc., pp. 44–47, 2015, doi: 10.1109/ICAICT.2015.7338513.

[9] O. Hegazy, O. S. Soliman, and M. A. Salam, “Comparative Study between FPA, BA, MCS, ABC, and PSO Algorithms in Training and Optimizing of LS-SVM for Stock Market Prediction,” Int. J. Adv. Comput. Res., vol. 5, no. 18, pp. 35–45, 2015.

[10] A. Panakkat and H. Adeli, “Neural network models for earthquake magnitude prediction using multiple seismicity indicators,” Int. J. Neural Syst., vol. 17, no. 1, pp. 13–33, 2007, doi: 10.1142/S0129065707000890.

[11] H. Adeli and A. Panakkat, “A probabilistic neural network for earthquake magnitude prediction,” Neural Networks, vol. 22, no. 7, pp. 1018–1024, 2009, doi: 10.1016/j.neunet.2009.05.003.

[12] M. Moustra, M. Avraamides, and C. Christodoulou, “Artificial neural networks for earthquake prediction using time series magnitude data or Seismic Electric Signals,” Expert Syst. Appl., vol. 38, no. 12, pp. 15032–15039, 2011, doi: 10.1016/j.eswa.2011.05.043.

[13] And M. A. S. Osman Hegazy, Omar S. Soliman, “FPA-ELM Model for Stock Market Prediction,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 2, pp. 1050–1063, 2015.

[14] M. Li, Z. Lieyuan, and S. Yaolin, “Attempts at using seismicity indicators for the prediction of large earthquakes by genetic algorithm-neural network method,” Unpubl. Manuscr., no. 1, pp. 483–489, 1998.

[15] L. MacEda, J. Llovido, and A. Satuito, “Categorization of earthquake-related tweets using machine learning approaches,” Proc. - 2018 Int. Symp. Comput. Consum. Control. IS3C 2018, no. May 2011, pp. 229–232, 2019, doi: 10.1109/IS3C.2018.00065.

[16] W. Li, N. Narvekar, N. Nakshatra, N. Raut, B. Sirkeci, and J. Gao, “Seismic data classification using machine learning,” Proc. - IEEE 4th Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2018, pp. 56–63, 2018, doi: 10.1109/BigDataService.2018.00017.

[17] G. Rajguru, Y. S. Bhadauria, and S. Mukhopadhyay, “Estimation of Earthquake Source Parameters Using Machine Learning Techniques,” 2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018, pp. 1–7, 2018, doi: 10.1109/ICCCNT.2018.8493922.

[18] Q. Wang, Y. Guo, L. Yu, and P. Li, “Earthquake Prediction Based on Spatio-Temporal Data Mining: An LSTM Network Approach,” IEEE Trans. Emerg. Top. Comput., vol. 8, no. 1, pp. 148–158, 2020, doi: 10.1109/TETC.2017.2699169.

[19] J. Reyes, A. Morales-Esteban, and F. Martínez-Álvarez, “Neural networks to predict earthquakes in Chile,” Appl. Soft Comput. J., vol. 13, no. 2, pp. 1314–1328, 2013, doi: 10.1016/j.asoc.2012.10.014.

[20] F. Martínez-Álvarez, J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero, “Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula,” Knowledge-Based Syst., vol. 50, pp. 198–210, 2013, doi: 10.1016/j.knosys.2013.06.011.

[21] F. Zhou and X. Zhu, “Earthquake prediction based on LM-BP neural network,” Lect. Notes Electr. Eng., vol. 270 LNEE, no. VOL. 1, pp. 13–20, 2014, doi: 10.1007/978-3-642-40618-8_2.

[22] A. Morales-Esteban, F. Martínez-Álvarez, and J. Reyes, “Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence,” Tectonophysics, vol. 593, pp. 121–134, 2013, doi: 10.1016/j.tecto.2013.02.036.

[23] Y. Wang, Y. Chen, and J. Zhang, “The application of RBF neural network in earthquake prediction,” 3rd Int. Conf. Genet. Evol. Comput. WGEC 2009, pp. 465–468, 2009, doi: 10.1109/WGEC.2009.81.

[24] A. S. N. Alarifi, N. S. N. Alarifi, and S. Al-Humidan, “Earthquakes magnitude predication using artificial neural network in northern Red Sea area,” J. King Saud Univ. - Sci., vol. 24, no. 4, pp. 301–313, 2012, doi: 10.1016/j.jksus.2011.05.002.

[25] G. Asencio-Cortés, A. Morales-Esteban, X. Shang, and F. Martínez-Álvarez, “Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure,” Comput. Geosci., vol. 115, pp. 198–210, 2018, doi: 10.1016/j.cageo.2017.10.011.

[26] K. Tan and X. Cai, “Prediction of earthquake in Yunnan region based on the AHC over sampling,” 2010 Chinese Control Decis. Conf. CCDC 2010, pp. 2449–2452, 2010, doi: 10.1109/CCDC.2010.5498782.

[27] E. Florido, F. Martínez-Álvarez, A. Morales-Esteban, J. Reyes, and J. L. Aznarte-Mellado, “Detecting precursory patterns to enhance earthquake prediction in Chile,” Comput. Geosci., vol. 76, pp. 112–120, 2015, doi: 10.1016/j.cageo.2014.12.002.

[28] M. Last, N. Rabinowitz, and G. Leonard, “Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries,” PLoS One, vol. 11, no. 1, pp. 1–16, 2016, doi: 10.1371/journal.pone.0146101.

[29] M. H. Rafiei and H. Adeli, “NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization,” Soil Dyn. Earthq. Eng., vol. 100, no. February, pp. 417–427, 2017, doi: 10.1016/j.soildyn.2017.05.013.

[30] T. Kerh, Y. H. Su, and A. Mosallam, “Incorporating global search capability of a genetic algorithm into neural computing to model seismic records and soil test data,” Neural Comput. Appl., vol. 28, no. 3, pp. 437–448, 2017, doi: 10.1007/s00521-015-2077-7.

[31] M. Mirrashid, “Earthquake magnitude prediction by adaptive neurofuzzy inference system (ANFIS) based on fuzzy C-means algorithm,” Nat. Hazards, vol. 74, no. 3, pp. 1577–1593, 2014, doi: 10.1007/s11069-014-1264-7.

[32] K. M. Asim, M. Awais, F. Martínez-Álvarez, and T. Iqbal, “Seismic activity prediction using computational intelligence techniques in northern Pakistan,” Acta Geophys., vol. 65, no. 5, pp. 919–930, 2017, doi: 10.1007/s11600-017-0082-1.

[33] Z. Umar, B. Pradhan, A. Ahmad, M. N. Jebur, and M. S. Tehrany, “Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia,” Catena, vol. 118, no. September 2009, pp. 124–135, 2014, doi: 10.1016/j.catena.2014.02.005.

[34] K. M. Asim, A. Idris, F. Martinez-Alvarez, and T. Iqbal, “Short Term Earthquake Prediction in Hindukush Region Using Tree Based Ensemble Learning,” Proc. - 14th Int. Conf. Front. Inf. Technol. FIT 2016, pp. 365–370, 2017, doi: 10.1109/FIT.2016.073.

[35] O. Hegazy, O.S. Soliman, and M. Abdul Salam, "A Machine Learning Model for Stock Market Prediction", International Journal of Computer Science and Telecommunications, Vol. (4), Issue (12),pp. 17-23, December 2013.

[36] O. Hegazy, O.S. Soliman, and M. Abdul Salam, "LSSVM-ABC Algorithm for Stock Price Prediction", International Journal of Computer Trends and Technology (IJCTT), Vol. (7), Issue (2),pp. 81-92, Jan 2014.

[37] O. Hegazy, O.S. Soliman, and M. Abdul Salam, "Optimizing LS-SVM using Modified Cuckoo Search algorithm (MCS) for Stock Price Prediction", International Journal of Advanced Research in Computer Science and Management Studies, Vol. (3), Issue (2),pp. 204-224, February 2015.

[38] O. Hegazy, O.S. Soliman, and M. Abdul Salam, "Comparative Study between FPA, BA, MCS, ABC, and PSO Algorithms in Training and Optimizing of LS-SVM for Stock Market Prediction", International Journal of Advanced Computer Research Vol.(5), Issue (18),pp.35- 45, March-2015.

[39] H. Abdul-Kader, MA. Salam. Evaluation of Differential Evolution and Particle Swarm Optimization Algorithms at Training of Neural Network for Stock Prediction. Int. Arab. J. e Technol. 2012;2(3):145-51.

[40] R. Salem, M. Abdul Salam, H. Abdelkader and A. Awad Mohamed, "An Artificial Bee Colony Algorithm for Data Replication Optimization in Cloud Environments," in IEEE Access, vol. 8, pp. 51841-51852, 2020, doi: 10.1109/ACCESS.2019.2957436.

[41] EM Badr, M. Abdul Salam, M Ali, H Ahmed, Social Media Sentiment Analysis using Machine Learning and Optimization Techniques, International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 41, August 2019.

[42] MA. Salam, M. Ali, Optimizing Extreme Learning Machine using GWO Algorithm for Sentiment Analysis. International Journal of Computer Applications.;2020, 975:8887.

[43] O.S. Soliman, and M. A. Salam, "A Hybrid BA-LS-SVM Model and Financial Technical Indicators for Weekly Stock Price and Trend Prediction", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 4, issue 4, pp. 764-781, 2014.

New Smart Encryption Approach based on Multidimensional Analysis Tools

Salima TRICHNI[1], Fouzia OMARY[2], Mohammed BOUGRINE[3]
Faculty of Sciences Mohammed V University in Rabat Department of Computer Science
Rabat, Morocco

***Abstract*—In the last decade, with the new situation forced by the Covide-19 pandemic, the information systems are often forced to work remotely, they must communicate and share confidential data with several interlocutors. In such a context, ensuring the confidentiality of communications becomes a complex and difficult task. Hence, the need to have a flexible system that can adapt with different parameters involved in every exchange of information. We recently presented in [1] a new smart approach to data encryption that serves the same purpose. This approach uses the concept of artificial intelligence and apply BNL skyline algorithm to decide about the most suitable algorithm to ensure the best data privacy. However, with the evolution of dimensions and criteria to be considered for this smart encryption, we find that the complexity of the BNL algorithm increase, then, the response time of the application increase and the skyline encryption quality decreases. In this work, we propose a new idea to resolve this problematic. Indeed, this contribution consists in adding another Intelligence brick to dynamically define the Skyline algorithm depending on the type and number of dimensions. Through this paper, we provide an analysis and a comparison of some skyline algorithms for the multidimensional search. The results obtained by this study show the performance of this new approach whether in terms of execution time or in the quality of the dominant encryption solution.**

Keywords—Security; confidentiality; artificial intelligent; smart encryption; cryptography; skyline

I. Introduction

With the enormous development of the communication´s means and the current circumstances caused by the covid-19 pandemic, a new management culture was brutally imposed on almost all institutions, companies and especially manufacturers who have suffered very impacting economic shock. Indeed, this new culture enforced by the health threat, is essentially based on collaboration and remote work to minimize the people´s movements and frequentation. We find ourselves with several new approaches in several areas, namely: e-administration, telemedicine, E-Learning and finally e-commerce which continues its development in the field of online shopping.

On the other hand, this situation forces us to go through telecommunications and technologies in order to share secrets, very confidential and critical data in a private network but most of the time public and not mastered. The attackers (the malicious ones) are more and more active and in constant search of vulnerabilities. Indeed, the percentage of vulnerabilities is increasing every day.

Computer security has therefore become a challenge for any company in order to ensure the continuity of its services. Research in this field has taken a new direction and aims to exploit new technologies in order to support immense digital development. New concepts have emerged, such as the concept of Identity-Based Cryptography (IBS) in which several publications are occurring as in [2] which aims to integrate authentication and integrity in the DNS and eliminates key escrow problem. At the same time, another security approach is also experiencing enormous development. This is the Blockchain. This revolutionary technology using cryptography and ensuring the security of transactions in full transparency is becoming a trend, and several researchers are thinking about how it can be applied in different needs. In [3], for example, the authors are coming up with Industrial IoT (IIoT) and Blockchain for the smart industries. The last component of development in the field of computer security is that which aims to integrate artificial intelligence into cryptographic processes. It is within this context that our research is oriented. In [1], we presented a new approach of intelligent security, adjustable with each information exchange and thus allowing communication with multiple entities regardless of their security protocol. In this system we consider the encryption algorithms as skyline points and we are used the BNL algorithm to choose the best one that best meets our security constraints. BNL works well if the size of the resulting Skyline line is small, and ideally fits into the window that causes the algorithm to terminate in a single iteration. However, this algorithm may require a large number of passes until 'that the complete Skyline is calculated and eventually terminates. So its performance is very sensitive to the number of dimensions and the distribution of the underlying data. Hence the need for a new method to stabilize the performance of this approach regardless of changes in the number of dimensions and their distributions.

The idea of this work consists in proposing a new solution concerning this approach that defines the Skyline algorithm according to the requirement of each communication. So instead of the skyline search algorithm being fixed, we modify it according to the type and the number of dimensions to be considered in each communication.

To present our work we will proceed as follows: First, we will start by citing the work related to our field of research and which integrates artificial intelligence techniques into computer security. Then, we will describe the principle behind the Intelligent Approach of Encryption. In a third section, we will present this new contribution and an experiment study of all basic Skyline algorithm to show their performance according to the parameters of each execution. Finally, in the last section we will discuss about the results and give conclusions.

II. Related Work

Artificial intelligence is a revolutionary technology that has been able to implement applications aimed at mimicking a form of real intelligence in several areas. Particularly in the area of security, several works have been carried out to strengthen and improve the security of information systems.

In general, there are two types of artificial intelligence use in the field of security. The first consists in creating decision support systems in order to control and improve the security policy of companies.

In cybersecurity, different solutions have been designed to overcome vulnerabilities and anomalies in systems. For example, in cyber-attacks, the thesis [4] offers an intrusion detection system in a completely unsupervised environment. This system is based on the Mutual Information algorithm for the selection of features and on the Deep Learning PV-DM model for reading network packets.

In this same area, other solutions are already starting to take their places in the market, such as the AI2 platform which, from the log lines, identifies suspicious activities. AI2 applies unsupervised machine learning algorithms on the input data to nominate potential attacks. Then, IT security analysts intervene to confirm and decide which incidents are real attacks. This system is also capable of continuously generating new models within hours, which can dramatically improve the speed of its cyber-attack detection capability [5].

The article [5] gives a summary and a detailed state of the art on other AI systems used in the field of security whether at the level of infrastructure, network, cloud, terminals, mobile, applications, IoT, or others such as the Web and Identity Management.

All the solutions mentioned above, come to help and facilitate the prediction and / or the detection of the problems and the faults of the systems. On the other hand, there is another category of AI solutions in the field of computer security.

This second category consists of designing systems that provide one or more properties of computer security based on AI techniques.

To ensure confidentiality, for example, in [6] the Google team proposed a solution based on the training of learning machines for the encryption and decryption of messages. Indeed, the authors consider the protagonists Alice and Bob as neural networks that try to communicate with each other in complete confidentiality while preventing a third malicious neural network named Eve from decrypting their messages. This system resembles the principles of electronic games.

The author in [7] represents another way to design an encryption system based on one of the paradigms of artificial intelligence which are on evolutionary algorithms. The Symmetrical Evolutionary algorithm (SEC) is the first variant of this type of algorithm which performs an encoding of the text in the form of positions lists. Then, it applies a set of genetic operators (mutation and crossover) on these positions at the level of each iteration in order to reproduce potential solutions. Finally, and through a well-defined evaluation function, assesses individuals and decides the safest solution. Several extensions of this system have been developed in order to increase the level of intelligence of this system in choosing the most relevant solution. Sometimes by adding difficult problems in the encryption process like the case of [8], and in other cases by modifying the evaluation function like in [9] and [10].

Our approach falls rather in the first category and it consists of a Decision Support System for the encryption of confidential data.

III. Background

A. Skyline Algorithm

1) The concept of dominance: Given a dominance relationship in a dataset, a Skyline query returns objects that cannot be dominated by any other object.

In the case of a dataset made up of multidimensional objects, one object dominates another object if it is equally good in all dimensions, and better in at least one dimension.

Skyline's computation was an algorithmic problem in nature, and all data was assumed to reside in memory.

However, nowadays we are faced with large data sets which are stored in secondary memory. Having the data on the disk, the algorithms proposed for processing Skyline requests are separated into two categories: algorithms not based on indexes and algorithms based on indexes.

2) Algorithms not based on indexes

a) Block Nested Loop (BNL)

A naive algorithm for calculating a Skyline query is to compare each object with all of the other objects in the dataset using a nested loop. However, the quadratic complexity O (n2) makes this algorithm very inefficient (n is the total number of objects in the data set).

The Block-Nested-Loop algorithm [11] applies the same idea, but uses a window (block of memory with limited space), which contains a limited number of data objects. Any candidate object p is compared to the objects of the window.

Three cases can occur:

1) p is dominated by an object in the window! p is eliminated.

2) p dominates one or more objects in the window! These objects are eliminated and p is placed in the window.

3) There are no objects in the window! p is inserted directly into the window. (In case the window is full, a temporary disk file is used to contain the candidate objects).

BNL may require a large number of passes until the full Skyline is calculated and eventually terminate because at the end of each pass the size of the temporary file will be reduced.

b) Divide & Conquer (D&C)

The D&C algorithm [12] [13] calculates the median value of a dimension, and divides the space into two partitions P1, P2.

Then it calculates the Skylines S1, S2 of P1, P2, recursively dividing P1 and P2.

Recursive partitioning stops when there are only one (or a few) objects. The overall Skyline is calculated by merging S1 and S2, and eliminating objects in S2 that are dominated by any object in S1.

c) Bitmap

The Bitmap algorithm as it was proposed in [14] is based on a vector representation of all the dataset.

In order to describe the algorithm, let p be a point in a d-dimensional space represented by a vector of m bits.

$p = \{p.d_1, ..., p._{dj}\}, 1 \leq j \leq d$

From these m bits, each p.di is represented by a number ki of bit slice. Each ki has as many bits as the number of distinct coordinate values of all points in the dataset in that dimension.

After constructing the bit slices, the algorithm performs 3 operations between 2 sets of bit parts. The first set contains the parts of bits Vx, Vy (one for each dimension) where resides the last bit of the point which is equal to 1. The second set contains the slices of bits Vx + 1, Vy + 1. In the case where the bit slices of the previous step are the last in order, we then use the bit slice zero (all bits at zero).

- The first operation A will be an AND operation between Vx and Vy.
- The second operation B will be an OR operation between Vx+1 and Vy+1.
- The third operation C will also be an AND operation between the results.

If the result of the final operation is zero, the tested point is a Skyline point.

3) Algorithms based on Index

a) Index

The index is a B-tree-based algorithm for two-dimensional data, where the data has two ordered indices.

For example, let a tree B and a tree B +, each represent a dimension. The algorithm calculates one over the entire Skyline line by scanning both indices simultaneously and stops as soon as a 'p' object is found in both indices.

Any object that has not been inspected in both indices is certainly not part of the Skyline line, as it is dominated by p. Therefore, candidate objects are those that have already been inspected in at least one index. These objects are kept in a separate set S (the superset of the Skyline line).

b) Nearest Neighbor Search (NN)

The first Skyline algorithm based on a spatial index (as R-tree), is the Skyline NN algorithm, proposed in [15]. It is called NN because of its relevance to the nearest neighbor search. It identifies Skyline objects by repeated NN search, using an appropriate distance measure.

The algorithm iteratively finds the closest object (NN) to the origin in a given region of space based on any monotonic distance function, e.g. the Euclidean distance. During the algorithmic process, entire regions dominated by a candidate object are ignored, and regions that cannot be ignored are added to a task list for further partitioning of space. And so on until the list becomes empty and, thus, the algorithm ends.

B. *Description of Intelligent Approach of Encryption*

The solution proposed in [1] is a decision-making system aimed to ensure the confidentiality of data in the most adequate way.

The following diagram shows the different steps in the framework of this strategy.

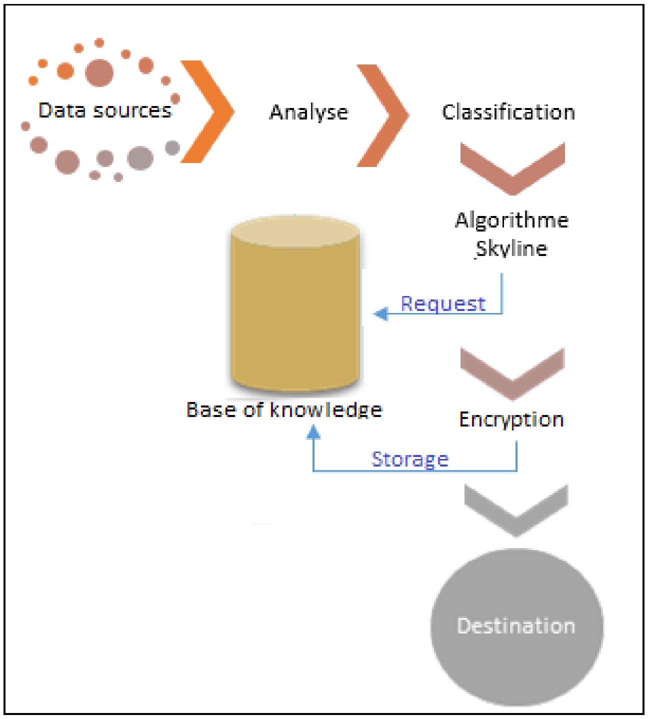


Fig. 1. Steps of the Intelligent Encryption Principle.

To explain the diagram of Fig. 1; we will detail each step in the following:

a) Step1: Analysis

Before proceeding with data encryption, in this solution we propose to rely on several characteristics of the encryption environment to decide the ciphering algorithm that best meets the security criteria to ensure in this communication.

This new encryption method begins with an analysis of the various elements that impact communication security. Indeed, we must analyze this impact at the level of:

- The source environment,
- The transmission channel,
- The destination environment

- The types of data to be transmitted,
- The generation of the keys, if not the possession of the keys.

Each part of this system represents a set of characteristics which considerably influence the security of this communication. Therefore, you have to focus on each part to extract the most relevant information.

b) Step2-Classification

Once this analysis has been carried out, we must move on to the second step which consists of classifying the data and storing all this information in an intelligent business architecture in order to subsequently decide on the most appropriate encryption.

The decisional database is a multidimensional base and knowledge base that includes a large number of experiments to cover the various cases possible.

c) Step3 - Skyline

In the third step, we launch the BNL skyline request on the previously established knowledge base by using the current communication requirement.

d) Step4 – Encryption and Storage

Depending on the chosen encryption algorithm we apply this encryption to our message and send the encrypted information.

Finally, the last step is a consolidation step in which we calculate the robustness of this encryption in terms of possible security indicators. Then, we feed the knowledge base by the feedback that we were able to complete.

IV. Contribution

A. *Problematic Description*

As noted in the previous section, the intelligent encryption approach uses the BNL algorithm to select the most dominant encryption algorithm against the specified criteria. This algorithm is very efficient in the case where one is satisfied with a very restricted set of dimensions with a dataset containing less candidates.

However, with the evolution of the knowledge base and the requirements to be taken into account for the security of communications, the complexity of this algorithm increases considerably and becomes quadratic of the order O (n ^ 2), where n is the size of the dataset to be examined; something that induces degradation of system performance.

In order to remedy this problem, we have carried out an in-depth study on certain Skyline algorithms, however we have come to the conclusion that each algorithm can be efficient and performant under certain conditions.

In what follows, we present this observation in detail and we propose a solution to circumvent this problem and grant a better flexibility of this approach.

B. *Description of Proposed Solution*

This contribution focuses more on step 3 of the Intelligent Encryption Principle. It seeks to optimize the use of Skyline algorithms for better results.

The solution is to add an extra step in this approach to dynamically choose the most suitable Skyline algorithm with respect to the data tuples to be examined.

Then the new process can be described as follow:

Step 1: Analysis of the data source/environment/channel

Step 2: Modeling and data classification in a BI database

Step 3: Building Criteria to Consider in the Skyline Query

Step 4: Skyline Algorithm Setting

Step 5: Application of the Skyline method

Step 6: Recovery of the most relevant encryption Skyline

Step 7: Evaluate the Solution and Enrich the Knowledge Base.

Below we prensent a summary diagram (Fig. 2) to illustrate how this new approach works to ensure the exchange of sensitive data between two entities (Alice and Bob) on different networks :

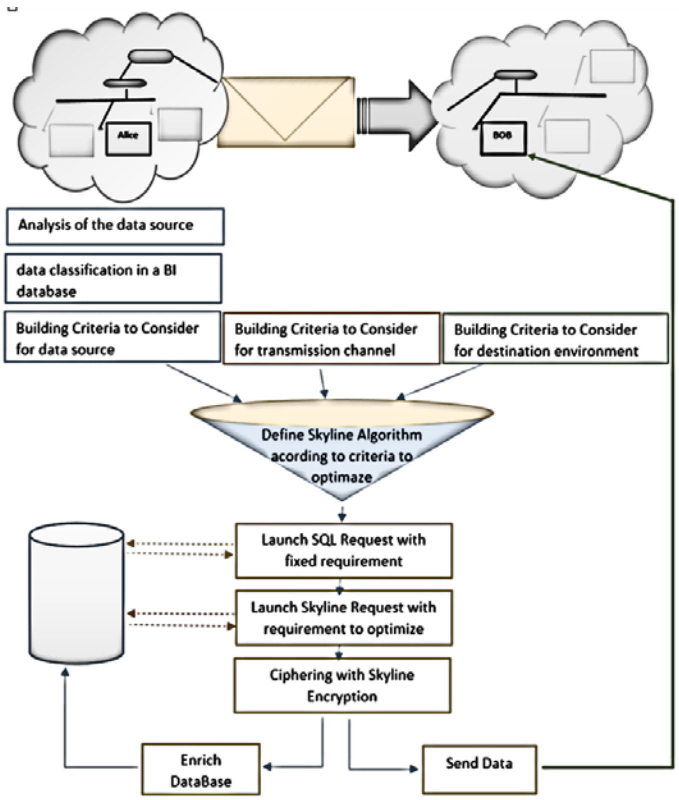

Alice
BOB
Analysis of the data source
data classification in a BI database
Building Criteria to Consider for data source
Building Criteria to Consider for transmission channel
Building Criteria to Consider for destination environment
Define Skyline Algorithm acording to criteria to optimaze
Launch SQL Request with fixed requirement
Launch Skyline Request with requirement to optimize
Ciphering with Skyline Encryption
Enrich DataBase
Send Data

Fig. 2. Diagram Summarizes the Steps of Our Contribution.

Now and after having stated the original contribution of this work which lies in adding an intelligent step allowing the configuration of the Skyline algorithm most appropriate to the criteria of our research. We chain in the following by an experimental study to analyze the impact of each criterion / dimension on the performance of the skyline algorithm in order to determine the set of parameters to be considered in the choice of Skyline.

For this, we will consider the following Skyline algorithms:

- Block Nested Loop (BNL)
- Divide & Conquer (D&C)
- Bitmap
- Index
- NN

C. Experiences &Results

To perform these experiments, we prepared a knowledge base with real data. Feeding this database has been made on the basis of an input data sample selected in an arbitrary manner which 100 texts same fixed characteristics.

Then, we applied to each text these five encryption algorithms: AES, Blowfish, DES, TripleDES [16] and ASEC which makes a total of more than a thousand test lines.

1) Specification of requirements: In an approach to optimization of processing, our method runs on two requests to attack the knowledge base as in [23].
The first query serves to minimize the volume of data to be examined in the next step. It uses fixed criteria in the SQL clause "Where" whose values are fixed at the start of each exchange.

The fixed criteria represent the properties of the current communication, for example:

- Type of data to be exchanged
- Characteristics of the source network
- Characteristics of the destination network
- The licenses available.
- The machine capacity of the encryption
- The machine capacity of the decryption
- Robustness of the transport channel between the two networks

The second query executes the skyline algorithm and considers the criteria to be optimized and from which the users' wishes in terms of the level of security to be acquired are made concrete. Hence the distinction between the two types of criteria.

For example: The criteria to be optimized:

- Quality (rate of randomness / entropy)
- The reputation of the algorithm
- encryption utilization frequency
- The cost of encryption / decryption
- The speed of encryption / decryption
- etc.

The test environment is based on an Intel (R) Core (TM) i7-6700HQ, 2.60GHz, 16GB RAM, 64bit OS machine.

2) Application of skyline algorithms: The application of Skyline algorithms was carried out by the Java programming language via the Spring Boot framework for the development of the Back End and Angular for the Front End part.
Through this application, users have the right to:

- Specify the number of dimensions.
- Specify the dimensions.
- Choose the operation to apply on each dimension.
- Choose the preferred algorithm for calculating Skyline points.

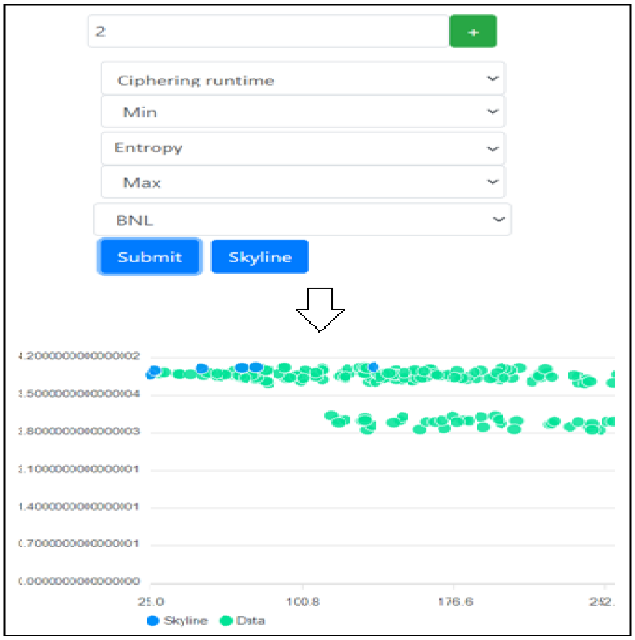
The application attacks the knowledge base and executes the chosen Skyline algorithm with the criteria entered as shown in Fig. 3, 4, 5, 6 and 7.

a) Scenario 1: On Two Dimensions

The objective of this experiment is to:

- Minimize encryption time
- Maximize entropy

Fig. 3 until Fig. 7 show the executions carried out for each Skyline algorithm and the output result.



2

Ciphering runtime

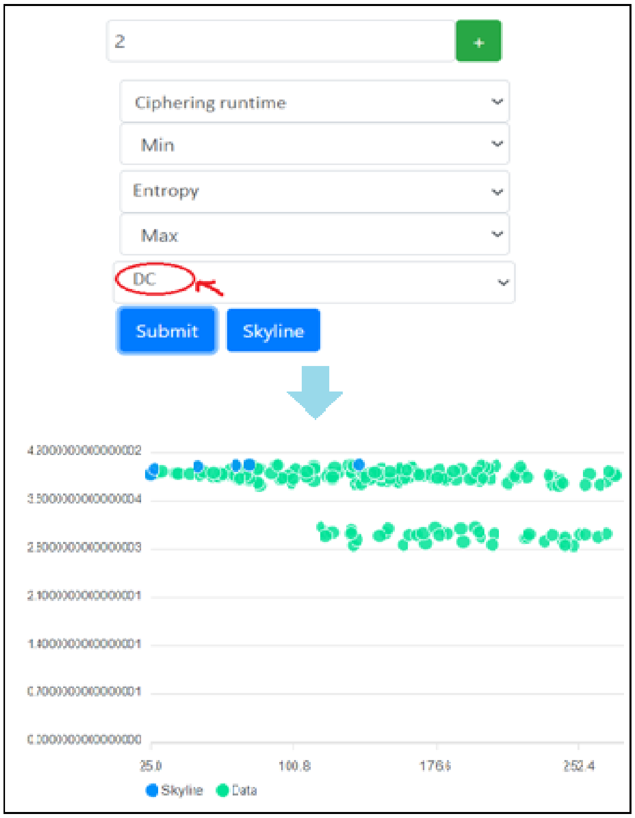
Min

Entropy

Max

BNL

Fig. 3. Application of BNL Skyline Algorithm with Two Dimensions.



2

Ciphering runtime

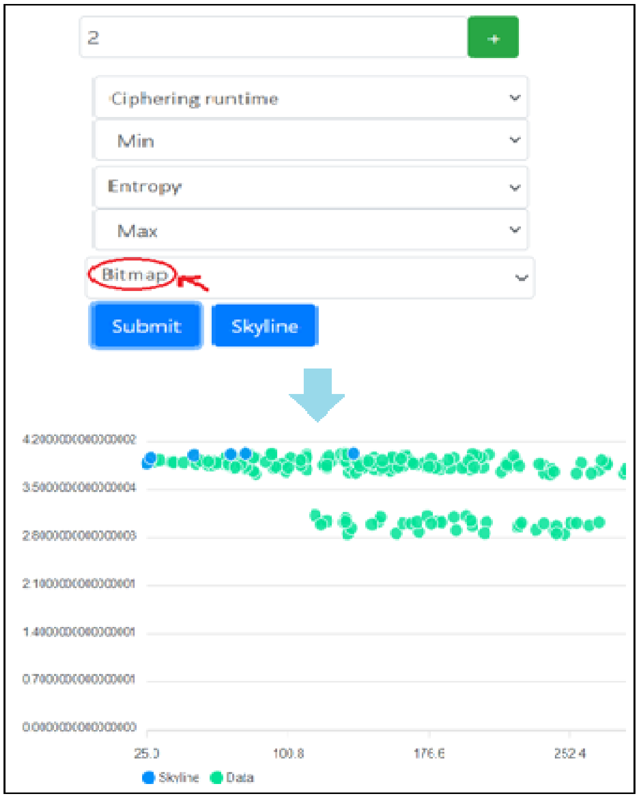
Min

Entropy

Max

DC

Fig. 4. Application of DC Skyline Algorithm to Optimize Entropy and Runtime Dimensions.



2

Ciphering runtime

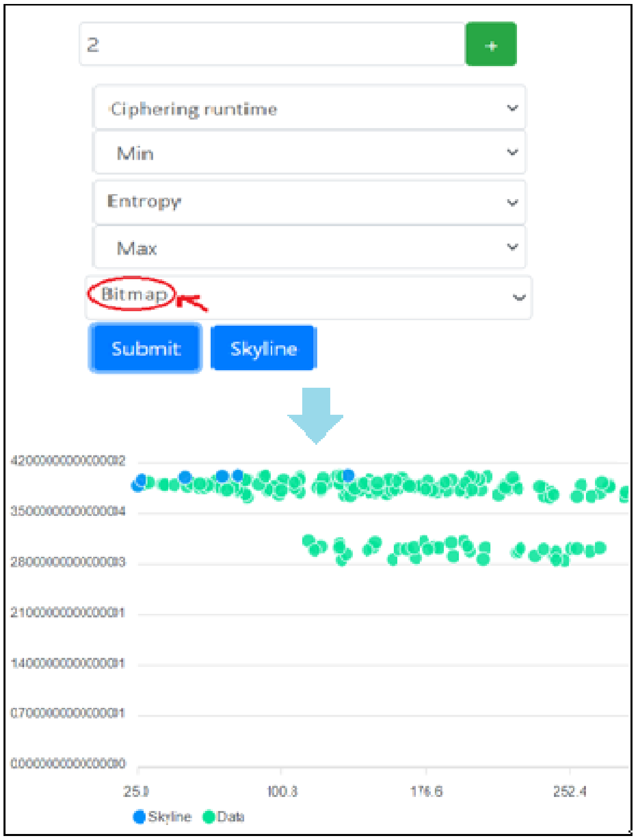
Min

Entropy

Max

Bitmap

Fig. 5. Application of Bitmap Skyline Algorithm to Optimize Entropy and Runtime Dimensions.



2

Ciphering runtime

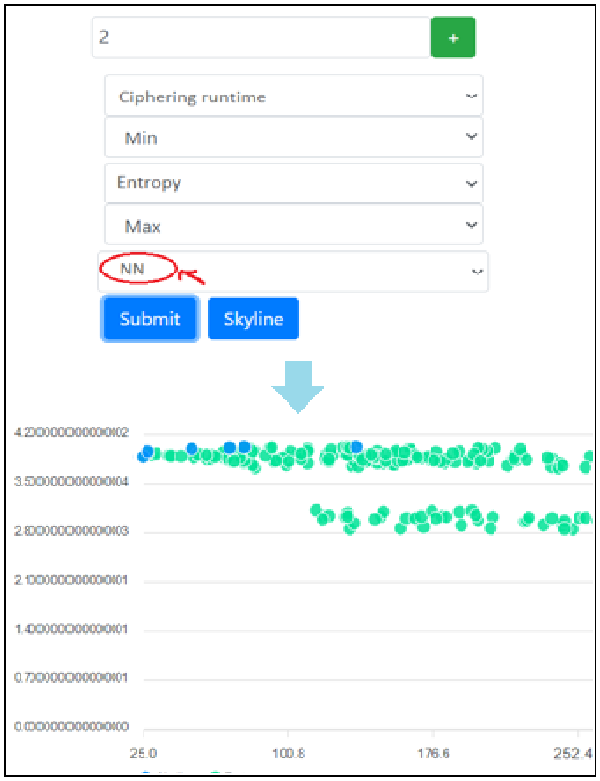
Min

Entropy

Max

Bitmap

Fig. 6. Application of Index Skyline Algorithm to Optimize Entropy and Runtime Dimensions.



2

Ciphering runtime

Min

Entropy

Max

NN

Fig. 7. Application of NN Skyline Algorithm to Optimize Entropy and Runtime Dimensions.

TABLE I. SKYLINE POINT FOR ENCRYPTION ALGORITHM THAT OPTIMIZE ENTROPY AND RUNTIME DIMENSIONS

Algorithm	Ciphering Runtime(ms)	Entropy
BLOWFISH	25	3.8870066417181426
BLOWFISH	27	3.9658710817437597
BLOWFISH	48	4.00235367813974
BLOWFISH	66	4.01642913400371
AES	73	4.026504332171881
BLOWFISH	126	4.027404704120802

- Results & Discussion of Scenario 1:

The results of the different algorithms were all the same and give out 6 skyline points as the dominant solutions. The following Table I summarizes this result:

The most efficient encryption algorithm is BLOWFISH comes after the AES algorithm in second. Indeed, in this case these two algorithms are well known in the community of computer security by their performance and their level of security [17] [18] [19] [22].

Now, if we go back to the performance of the Skyline algorithms, to come up with these results there is a huge difference in the execution time of these algorithms.

The diagram below shows the result of the execution time of each Skyline algorithm.

So from the Fig. 8, we can conclude that the BNL algorithm is the simplest and the best if we do not use the two dimensions encryption time / Entropy.

However, if we change the two dimensions considered in the first experiment and we opt for the following scenario:

b) Scenario 2: on Two Dimensions

- Objective 1: to minimize the decryption time.
- Objective 2: minimize the memory capacity required for decryption.

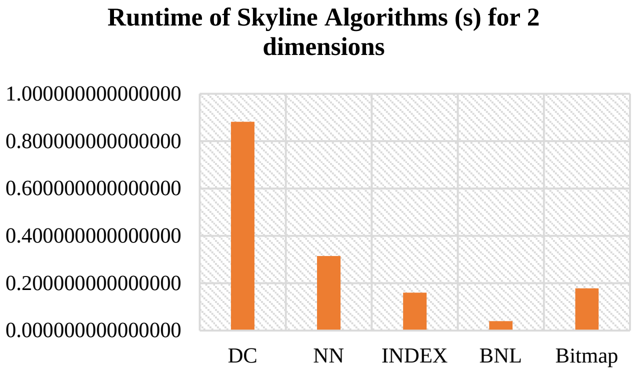


Fig. 8. Comparison of Runtime Skyline Algorithms (s) Executed with Entropy and Runtime Dimensions.

TABLE II. SKYLINE POINT FOR ENCRYPTION ALGORITHM THAT OPTIMIZE DECIPHERING RUNTIME AND USED MEMORY

Algorithm	Deciphering Runtime(ms)	Deciphering Memory used (K)
SEC	48	1695
SEC	56	1841
ASEC	93	2503
AES	243	2517

The application of the Skyline algorithms was carried out in the same way as the first experiment except for the types of dimensions which have just been modified to select the Decryption Runtime and the Memory dimensions.

In this experiment, the results in terms of Skyline points are different,

The Table II below is the summary of the results of the different executions:

- Results & Discussion of Scenario 2:

The SEC encryption algorithm is ranked first, followed by the ASEC (advanced version of SEC) algorithm, followed by the AES algorithm.

The result given by this system seems logical because SEC is an evolutionary algorithm [20] which takes enough time to perform the encryption and generate the key, however, in the decryption process it applies a single operation to the whole text at once. As a result, this type of algorithm performs best with respect to the cost of decryption.

On to the result of the skyline algorithm runtime, the following diagram illustrates the results of this scenario:

From this experience, we can see in Fig. 9 that the Bitmap algorithm performed well in this scenario.

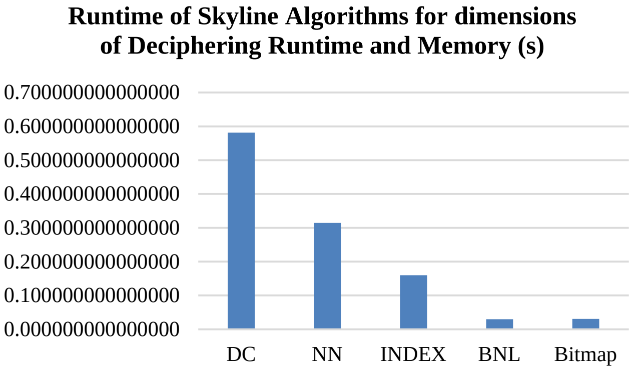


Fig. 9. Comparison of Runtime Skyline Algorithms for Dimensions of Deciphering Runtime and Memory (s).

This is because the Bitmap vector representation for these two dimensions is less complex. In fact, the more the number of values of a dimension are mastered, the algorithm becomes faster and more efficient [21]. In Scenario 1, the entropy dimension had very varied values and therefore the Bitmap vector representation will be very large and complex.

On the other hand, the BNL algorithm always remains efficient even by modifying the type of dimensions considered [21].

c) Scenario 3: for Four Dimensions

- Objective 1: minimize encryption time.
- Objective 2: maximize entropy
- Objective 3: minimize decryption time
- Objective 4: minimize the memory capacity required for encryption.

The diagram below shows the result of the execution time of the skyline algorithms.

- Results & Discussion of Scenario 3

From the results of Fig. 10, we see that the Bitmap algorithm becomes very complex however, the index algorithms become more efficient.

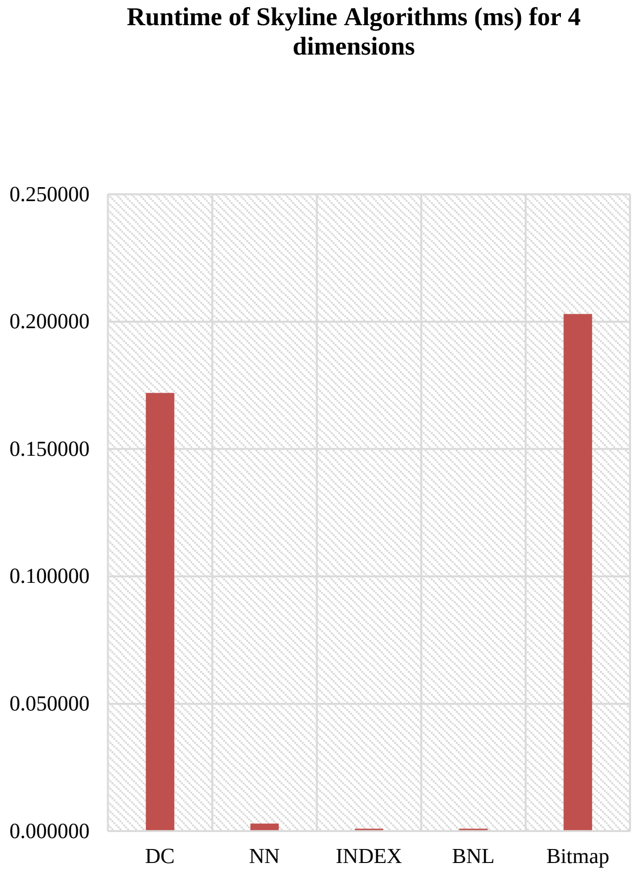


Fig. 10. Comparison of Runtime Skyline Algorithms (s) Executed with 4 Dimensions.

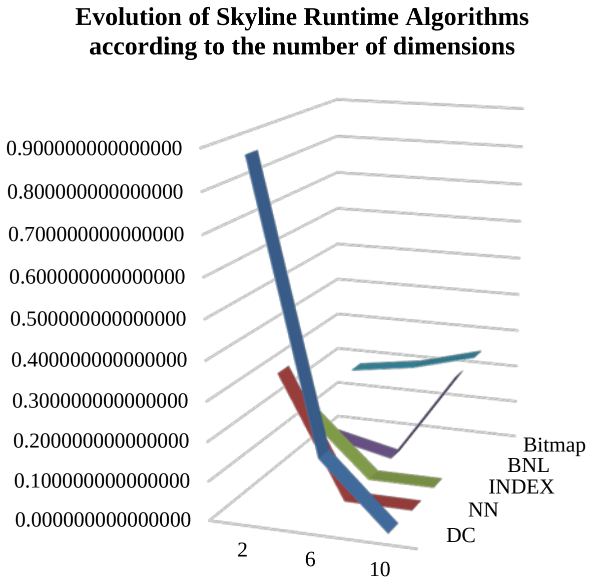


Fig. 11. Compariosn of Evolution of Skyline Runtime Algorithms according to the Number of Dimensions.

And so on, the more we add dimensions, the more the performance of the algorithms changes.

d) Scenario 4: up to 10 Dimensions

In this experiment, we ran the application on several dimensions to assess the impact of the number of dimensions on the speed of Skyline algorithms.

The Fig. 11 below shows the result of the execution time of the skyline algorithms against number of dimensions.

3) Comparison and discussion: As we have presented in previous experiences, we focused on two very important factors for the success of this approach, namely:

- The execution time of Skyline algorithms because it has a very significant impact on the speed of the entire system.
- The quality of the Skyline selected (the encryption to be implemented) because it represents the heart of this approach and embodies the robustness and security of the system.

The quality of the skyline solution can be evaluated based on the following characteristics:

- Guarantee: All returned points are skylines.
- Accuracy: All the points returned meet the criteria previously defined.
- Progressiveness: the sending of results is done instantly regardless of the size of the database are often huge.
- Completeness: at the end, all the points of the skylines are returned.

From this experiment study, we can conclude that there are two main parameters to satisfy all these characteristics and guarantee the performance of the skyline algorithms on encryption problem. These two parameters are:

- The Number of Dimensions
- The rate of variation which means the rate of the difference between numerical values of a given dimension. The greater this difference, the lower the rate of change and vice versa.

For example, to demonstrate the impact of the rate of change on the quality and performance of this approach, we can go back to experiments 1 and 2. In fact, in the first scenario we used two dimensions with different rates of change (the encryption with a normal rate of change and entropy for which the rate of change is very high). While in the second experiment we kept the same number of dimensions but with a normal rate of variation for both. The result of the first experiment favored the BNL algorithm however in the second experiment the Bitmap algorithm performed better. Otherwise, compared to the quality of the returned Skylines, they were all on the same level. In terms of:

Accuracy: criteria met

Guarantee: exact encryption

Completeness: same list of Skylines,

Progressivity: the list was returned all at once however this test is not interesting in this case since the number of dimensions is very low and the size of the base is fixed for the moment.

The two tables, Tables III and IV below summarize these results:

TABLE III. THE CRITERIA FOR CHOOSING THE SKYLINE ALGORITHM IN THE CASE OF 2 DIMENSIONS AND AN UNCORRELATED RATE OF VARIATION

Scenario 1	d=2			
	uncorrelated rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	oui	oui	oui	1
D&C	oui	oui	non	5
Bitmap	oui	oui	non	3
Index	oui	oui	oui	2
NN	oui	oui	oui	4

TABLE IV. THE CRITERIA FOR CHOOSING THE SKYLINE ALGORITHM IN THE CASE OF 2 DIMENSIONS AND A UNIFORM RATE OF VARIATION

Scenario 2	d = 2			
	uniform rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	oui	oui	oui	2
D&C	oui	oui	oui	5
Bitmap	oui	oui	oui	1
Index	oui	oui	oui	3
NN	oui	oui	oui	4

The objective behind the other two experiments is to demonstrate the impact of the dimension number on the quality and speed of the system. In fact, in these two experiments, we each time added additional dimensions to our research with different rates of variation.

From the results of these experiments, we find that the NN and D&C algorithms are starting to meet the demanded needs better than other algorithms. The Bitmap algorithm always responds very well in cases of dimensions with uniform rate of change regardless of the number of dimensions. However, the BNL algorithm is no longer favored if the number of dimensions is large.

At the end and to summarize all these results, we can use the Tables V and VI below to decide on the choice of the Skyline algorithm which gives us the most secure encryption:

TABLE V. GLOBAL CRITERIA FOR CHOOSING THE SKYLINE ALGORITHM IN THE CASE OF LESS THAN 4 DIMENSIONS

	d < 4			
	uniform rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	-	oui	oui	oui
D&C	-	oui	oui	non
Bitmap	-	oui	oui	oui
NN	-	oui	oui	non
	uncorrelated rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	-	oui	oui	oui
D&C	-	oui	non	non
Bitmap	-	oui	non	non
NN	-	oui	oui	non

TABLE VI. GLOBAL CRITERIA FOR CHOOSING THE SKYLINE ALGORITHM IN THE CASE OF MORE THAN 4 DIMENSIONS

	d >= 4			
	uniform rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	non	oui	oui	non
D&C	non	oui	oui	non
Bitmap	oui	oui	oui	oui
NN	oui	oui	oui	oui
	uncorrelated rate of variation			
	Progressiveness	Guarantee	Completeness	Speed
BNL	non	oui	non	non
D&C	non	oui	oui	oui
Bitmap	non	oui	non	non
NN	oui	oui	non	oui

Then, if we come back to our new approach (Fig. 2), for the step 4 of this new contribution that consist in 'Skyline Algorithm Setting' we can conclude that we have tree general configuration of Skyline algorithms:

Configuration 1: when performance achieved on a small number of dimensions such as BNL.

Configuration 2: performance achieved if we work on a large number of dimensions such as the DC and NN algorithms

Configuration 3: performance achieved if we work on dimensions with a limited number of different values, then we can use Bitmap algorithm.

V. Conclusion

Despite the significant development of security tools, computer systems still suffer from malicious threats especially with the huge growth of emerging technologies.

To be able to support this revolution, these new technologies must be integrated into the various security components. During this work, we tried to develop a new approach to intelligent encryption that is based on the concepts of artificial intelligence. It uses Skyline algorithms to define the policy to be followed to ensure the confidentiality of exchanges. In this work we proposed a new version of this system included an additional step to define dynamically the Skyline algorithm to be executed for choice the good Encryption Algorithm. So, from the experimental demonstration we deduced that the choice of the Skyline algorithm has a perimeter role in this approach and the use of the BNL algorithm, as in the previous contribution, risks weakening the performance and the efficiency of this encryption system, especially if the variation rate of dimension values or their number increases We applied different Skyline algorithms and we compared the results. Finally, and after several scenarios we concluded that this system must absolutely change the choice of Skyline Algorithm taking into account two main parameters which are: The number of dimensions and the rate of variation.

References

[1] S.Trichni; F.Omary; A.Idrissi; M.Bougrine; M.Abourezq : New intelligent strategy for encryption decisional support system - International Journal of High Performance Systems Architecture (IJHPSA), Vol. 9, No. 4, 2020

[2] M. Patel, R. Patel : Improved Identity Based Encryption System (IIBES): A Mechanism for Eliminating the Key-Escrow Problem - Emerging Science Journal, Vol 5, No 1 (2021), DOI 10.28991/esj-2021-01259

[3] A. Iqbal, M. Amir, V Kumar, A. Alam, M Umair : Integration of Next Generation IIoT with Blockchain for the Development of Smart Industries, Vol 4 (2020) , DOI: 10.28991/esj-2020-SP1-01

[4] Samira Douzi : Vers un Deep Learning Système de Detection des Intrusions - Doctoral thesis dissertation- Jully 2019

[5] Vähäkainu, Petri & Lehto, Martti. (2019). Artificial intelligence in the cyber security environment Artificial intelligence in the cyber security environment. https://www.researchgate.net/publication/338223306_Artificial_intelligence_in_the_cyber_security_environment_Artificial_intelligence_in_the_cyber_security_environment

[6] Mart´ın Abadi and David G. Andersen Google Brain "LEARNING TO PROTECT COMMUNICATIONS WITH ADVERSARIAL NEURAL CRYPTOGRAPHY" arXiv:1610.06918v1 [cs.CR] 21 Oct 2016

[7] F.Omary : Application Of Evolutionary Algorithms To Cryptography (Applications Des Algorithmes Evolutionnistes À La Cryptographie).Doctoral Thesis, University Mohammed V Agdal , Faculty Of Science - Rabat Marocco. (July 2006).

[8] S.TRICHNI and al: A New Approach Of Mutation Operator Applied To The Ciphering System Sec. Iccit 2011,vol 63, no. 9;sep 2013.

[9] M.Bougrine, F.Omary , S.Trichni, : A new Evolutionary Tools for New Ciphering System SEC Version, 46ième International Carnaham Conférence On Security Technology (IEEE ICCST 2012),Boston Massachusetts, USA ISBN 978-1-4673-4807-2, ISSN :1071 ; p140-146.

[10] M.Bougrine, S.Trichni, F.Omary : Improving Performance of the Symmetrical Evolutionary Ciphering System SEC - International Journal of High Performance Systems Architecture (IJHPSA), (2021) (publication in progress)

[11] S.Borzonyi, D.Kossmann, K. Stocker : "The Skyline operator", ICDE, pp.421-430, 2001.

[12] H.T. Kung, F. Luccio, F.P. Preparata : "On finding the maxima of a set of vectors", JACM, Vol.22, No.4, pp.469-476, 1975.

[13] F.P. Preparata, M.I. Shamos : "Computational geometry : an introduction",Springer-Verlag, New York, Berlin, 1985.

[14] K. Tan, P. Eng, B. Ooi : "Efficient progressive Skyline computation", VLDB, pp.301-310, 2001.

[15] D. Kossmann, F. Ramsak, S. Rost : "Shooting stars in the sky : an online algorithm for Skyline queries", VLDB, pp.275-286, 2002.

[16] W. DIFFIE, M. E. HELLMAN, "New Directions in Cryptography " IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-22, NO. 6, NOVEMBER 1976 Pp 644 –654

[17] Zimmermann, P. R. (1991) PGP User's Guide, 5th June 1991, Version 1.0, Phil's Pretty Good Software.

[18] Florin G. Et Natkin S: Techniques Of Cryptography. Cnam 2002.

[19] Menezes A.J., Oorschot P.C. Van Et Vanstone S.A.: Handbook Of Applied Cryptography.(Crc Press, 1997).

[20] Shaul Drukmann: Evolutionary Algorithms.Encyclopedia Of Computational Neuroscience 2014, Pp 1-7.

[21] K.Tan, P.Eng, B. Ooi : Efficient Progressive Skyline Computation. VLDB, 2001.

[22] AK.Diaasalama and M. Hadhoud : Studying the Effect of Most Common Encryption Algorithms," International Arab Journal of e-technology,Vol.2. No.1

[23] M. Abourezq and A. Idrissi : Introduction of an outranking method in the cloud computing research and selection system based on the skyline. In Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on, pages1–12. IEEE. .(2014b).

Natural Language Processing Applications: A New Taxonomy using Textual Entailment

Manar Elshazly[1], Mohammed Haggag[2], Soha Ahmed Ehssan[3]
Computer Science Department
Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

***Abstract*—Textual entailment recognition is one of the recent challenges of the Natural Language Processing (NLP) domain. Deep learning strategies are used in the work of text entailment instead of traditional Machine learning or raw coding to achieve new enhanced results. Textual entailment is also used in the substantial applications of NLP such as summarization, machine translation, sentiment analysis, and information verification. Text entailment is more precise than traditional Natural Language Processing techniques in extracting emotions from text because the sentiment of any text can be clarified by textual entailment. For this purpose, when combining a textual entailment with deep learning, they can hugely showed an improvement in performance accuracy and aid in new applications such as depression detection. This paper lists and describes applications of natural language processing regarding textual entailment. Various applications and approaches are discussed. Moreover, datasets, algorithms, resources, and performance evaluation for each model is included. Also, it compares textual entailment application models according to the method used, the result for each model, and the pros and cons of each model.**

Keywords—Textual entailment; deep learning; summarization; sentiment analysis; information verification; machine learning; depression detection

I. Introduction

Textual entailment[1] is the process of importing a text from another one. Textual entailment [2] is an obtaining the relationship between fragments of text when one fragment in some sense implies the other.

Recognition of text entailment (also called natural language inference) is the task of deciding on the allocation of two pieces of text whether the interpretation of one text can be given in another text. Text entailment is a relationship-related activity and sometimes there are threee relationships can occur between two sentences. Entailment- [1, 3] the meaning of one sentence can be combined with another sentence, incorrect entailment - the meaning of one text contradicts the meaning of another sentence, and non-entailment - the meaning of one text does not include or contradict the meaning of another sentence. Or may be categorized to two class (ENTAILS/NEUTRAL).

Textual entailment research mainly contains methods for developing and evaluating algorithmic identification methods of entailment [2]. Text entailment aims at a deeper understanding of text and thinking, which shares the same type of machine learning comprehension, although the structure of the functions is slightly different [4].

Similar to Semantic Textual similarity, in the task of Textual Entailment, the input also contains two sentences (pre-sentence P and hypothesis H sentence), and programs should determine whether the meaning of sentence H can be extended from the meaning of sentence P. In particular, [5, 6] systems are required to assign each line of the phrase or label of CONFIDENCE (when P comprises H, that is, H cannot be false when P is true), the CONTRADICTION label (when P contradicts H, i.e. H is false whenever P is true), or the INDEPENDENT label (where true of H could be determined on the basis of P).

Textual Entailment task is a problem of classification [5]. TE is the latest and best indicator of text acquisition in NLP applications [7].Recent work on Text Entailment work includes (1) models based on linked elements and machine-learning classifiers, (2) sentence-based models that use similar structures such as Semantic Textual similarity, sentence-based function, and (3) other neural network models that do not use Simple RRNs or CNNs as sentence encoders.

Textual Entailment (TE) is an inconsistent relationship between two expressions in which the meaning of one phrase, called Hypothesis (H), derives from another expression called Text (T). The definition of TE is strong in the sense that if H enters from T but has little or no additional information, then the two are treated as unrelated or non-entailed [8].

Text entailment is used in text encryption methods to measure text connections. TE in NLP is a direct relationship between the sections of the text. Relationships are active and effective when the fact of one text fragment follows from another fragment. In the TE framework, integrated and integrated texts are defined as text (T) and hypothesis (H) [7] .

Textual Entailment (TE) is a typical example of a semantic inference, in which the purpose is to determine where a textual hypothesis (H) can be correctly included in a given text (T). While classifying the hypothesis into genes and making an effort to determine where each is differentiated separately by T is called Partial Textual Entailment (PTE) [7].

Partial Textual entailment (PTE) is a possible solution to this problem that defines the interdependent relationships between T and H pairs. PTE relationships can play an important role in a variety of programs to use Natural Language Processing (NLP) such as text summaries and a question answering system by minimizing unwanted information [8].

Recognition of textual Entailment (RTE) [8] is an important function in Natural Language Processing (NLP) research. The key role of RTE includes a question-and-answer system where the nature of the engagement is confirmed by the given answer regarding the expected answer, a summary of multiple texts where the sentences should be deleted, and a retrieval system where only documented questions are required. However, the old definition of TE has two important implications for the relationship. First, the relationship does not coincide where H should be inserted from T. Otherwise the two pair are non-entailed. PTE [9] is a two-way relationship between two expressions in which one expression is inserted or slightly dependent from another expression. PTE expands Textual Entailment (TE) concept.

The rest of the paper is organized as follows: Section (II) provides a literature review about NLP applications that uses textual entailment, in Section (III) provides a comparison between machine learning and deep learning in textual entailment. The paper concludes in section. The paper concludes in Section (IV) with an overview of future work that it was aimed to be done

II. Natural Language Processing Applications using Entailment

Textual entailment is a relationship that obtains between fragments of text when one fragment in some sense implies the other fragment. The automation of textual entailment recognition supports a wide variety of text-based tasks, including information retrieval, information extraction, question answering, text summarization, and machine translation [2].

A. Text Summarization

NLP has several applications, the challenging one is a text Summarization. Summarization is a process of presenting a text's most important and relevant information in precise way. It helps to understand a text very quickly and precisely.

Determining the level of correlation between input sentences will help to reduce the inclusion of meaningless sentences in abbreviated text. The importance of the summary becomes even more important and effective when dealing with events on social media platforms such as Twitter, where information is fast, rich, varied and ever-growing content, no matter where, time and other circumstances.

Any user can post comments or watch, upload videos to any event very quickly. Current methods of measuring sentences only find similarities between words and sentences. These methods only state the correct information for each sentence. Summarizing text helps to understand any major text in a very short time .It is to produce the right summary, there are a few steps to follow, e.g. Lexical Analysis, Semantic Analysis and Syntactic Analysis.

The content of the summary story details depends on the needs of the user. Text Summary Methods can be classified as abstractive and an extractive summarization. The abstractive summarization incorporates key ideas into the document and conveys those ideas in clear natural language. The extractive summarization process consists of selecting key phrases, sentences, paragraphs etc. In the original document and combine those with a short form that maintains a sequential sequence [7].

The main purpose in [10] is to measure sentence similarity, which will help to summarize the text of any language, looking only at English and Bengali. The methods proposed in [10] were extensively tested using several English and Bengali texts, from many online sources. This similarity of sentence is stated. Prior to in-depth analysis the best way to prepare for the operation of both languages used is Unicode.

Model [10] represents a new proposed sentence similarity measuring model for English and Bengali language. Lexical methods have been applied, and a perfect expected result has been found.

Lexical Layer Analysis: The lexical presentation and lexical similarity are major function in lexical layer as shown in the Fig. 1. The model followed by one of the popular deep learning algorithm-Text Rank. Lexical Layer Analysis Model has the following steps as:

- Lexical analysis: brake the sentence into tokens.
- Stop words removal: remove pronoun and article words.
- Lemmatization and Stemming: return all tokens to their entry.

Sentence Similarity: The rating method helps to hear the correlation of words that occur in WordNet times. Send the distance between the two words. Other equations to be used in the same suggested algorithms, in this proposed model use python with NLTK version 3, and WS4J: java API designed for WordNet use [10].

However, Twitter's summary work has garnered a lot of attention from the research world over the past decade as a result of sharing fast and varied responses from millions of users.[9] While, summarization of Twitter event is much harder than traditional texts due to the different type of text that poses many challenges. Like:

- Processing of tweet content : this, standard Natural Language Processing (NLP) tools do not work well in this type of text and create challenges for tweet content processing, design integration and topic by topic due to the high nature of understanding features.
- Sentence Boundary Detection (SBD): A tweet does not consist of a single sentence, instead, it can include multiple sentences. Sometimes, a tweet does not include punctuation even if it includes multiple sentences.

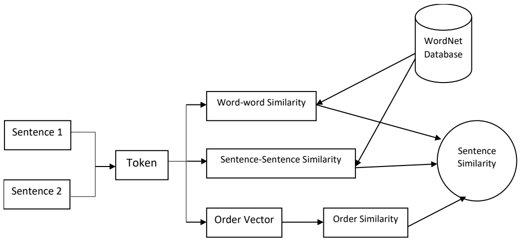


Fig. 1. Lexical Layer Analysis Model[10].

- Information redundancy problem: Events on Twitter attract a large number of tweets making it difficult to find the most important personal tweets. Every event contains incorrect details of an event in the form of a Re-tweet or titter with the full or partial content of another tweet in high volume. To reduce misunderstandings, identifying partially embedded information between tweets is harder than seeing re-tweets or embedded tweets. For example, two sentences are listed below in two separate tweets. The second sentence contains the information included in the first sentence and additional content (highlighted). To reduce the amount of non-reversible information, these two sentences can be combined into one meaningful sentence as shown below.

Model [9] proposes a summarization approach to extracting an amazing event summary from Twitter by providing a solution that can solve the above challenges. Fig. 2 illustrates this summary. This approach follows a detailed design to produce a summary of the most informative and relevant sentences from all event tweets. After they made this model they found that:

- In many previous researches, tweet is considered a summary of work. But in reality, regular tweets include many sentences. In this research the authors extract important sentences by SBD from tweets instead of treating the tweet as a sentence. In order to exclude the most important events of the event, they put all relevant tweets of the event first and convert the top tweets placed into possible sentences.

- To produce an abstract summary, authors filter out invalid data by checking the Partial Textual Entailment (PTE) relationship between sentences. They see peer-to-peer (PE) sentences and a text border that is slightly inserted between sentences. Couples of PE sentences are combined into one sentence by combining PE text. Finally, an abstract summary creates selected sentences in addition to covering high detail, diversity, consistency and readability.

In [9] the proposed model is two phase summarization approach. First phase is key sentence extractor which produces more relevant and informative sentences rather than tweets from the appropriate tweet event set, second phase is abstractive summarizer which produces a desirable summary over the selected sentences after deleting the unwanted information by pointing to the information included or partially included in the sentences. So, authors have seen the quality of the summary is limited by the accuracy of the TE determination. This [7] seeks to overcome these restrictions by using Partial Textual Entailment (PTE) as well. Hence, [7] suggested a modified approach that demonstrated competitive outcomes in standard dataset.

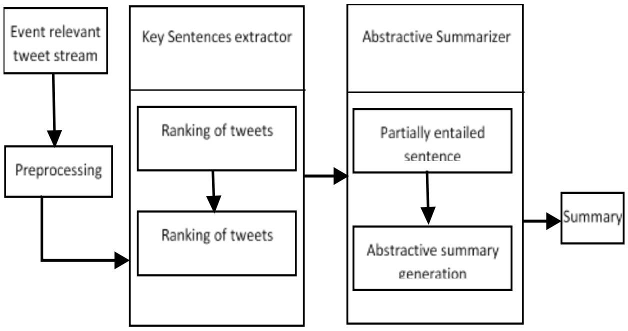


Fig. 2. [9] Summarization Approach.

Therefore model [7] suggests a new and exciting way to summarize one document. It also includes Partial Textual Entailment (PTE) as a way to get the power of expression within a couple of texts. Sentences are divided into units of text for the purpose of TE to compute which is why they are different from complete TE and are called intra-phrase sentences or PTE. Since a graph model is known to capture and present a well-defined structure of information with information shared between interconnected areas within a complex network, so authors will use a graph presentation to illustrate the relationship between different pieces of text. Then authors used the weighted (Minimum Vertex Cover) method of MVC, a graphical algorithm to get a small vertices' set (representing sentences) that make up the cover. And this cover forms their final summary [7].

The text summary function was created as a WMVC problem by them. The input document provided to the system to be compensated and rewritten and labeled as a graph network. They built a targeted, weighted graph model where border instruments could not be considered. Accordingly, the weight w is already expressed in each graph node based on the entailment values obtained. This improved graph form to date is placed under WMVC method [7]. This algorithm works by narrowing down the graphical network so that the cover C is a subset (part) of the smallest sentences weight, meaning the best-placed sentences in the text. The cover that has been found is this result - a summary of the input document. Model [7] Recommended Method of partial Text entailment Used Baseline Weighted Minimum Vertex Cover (PTE-WMVC), shown in Fig. 3 is contained the following key steps:

- Pre-processing step
- PTE score computation
- Salience scores computation
- PTE based graph construction
- Applying WMVC algorithm

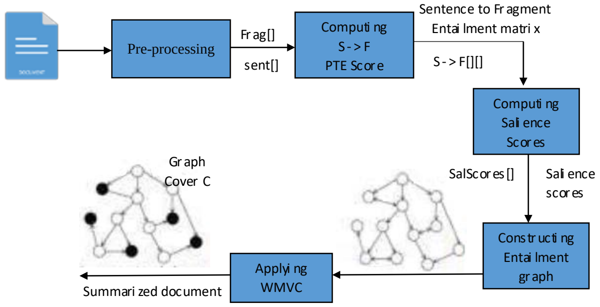


Fig. 3. PTE-WMVC Structure Diagram [7].

Finally, the authors compared the effectiveness of this approach with other leading methods in the field of text summarization. This PTE-WMVC system provided the highest recall value of 50.99% of ROGUE-1 while n-gram size was considered N = 4. The next best performance is provided by WMVC with a recall value of 38.8% followed is PDT at a rate of 34.6%. A very small amount of recall was identified by the program and ATESC as the basis for text processing in the system [7]. In [7], authors introduced an amazing and advanced way of summarizing a single text. Through extensive experiments in the database that showed in this paper the proposed algorithm uses Partial Textual Entailment at its end it uses existing techniques that look at complete textual entailment to establish relationships between sentences.

The suggested algorithm has been used successfully in small sections of sentences, uses Partial Textual Entailment and produces clear and reliable measurements. Building such a system and dealing with the same method of summarizing multiple documents by taking advantage of the functionality of the proposed algorithm involves future work. Future work of [7] will be more focused on summarizing many documents. The creation of a summary requires multiple textual sources and often involves a combination of collections of various texts.

B. *Question Answering*

Question Answer (QA) is an ancient research field in computer science that started in the sixties. In the Semantic Web, most of the suspended data is now available in the form of knowledge bases (KBs). Today, there are KBs on media, publishing, geography, life-science and more. The main purpose of the QA program over KBs is to retrieve the information you want from KB or more, using native language questions.

In summary, most QA programs only work in English and more than one KB. Multilingualism is taken for granted while carrying it can often be taken for granted. The fact that QA systems tend to use existing technologies and require several services to be presented to the end user, leads to the idea of developing QA systems in a general way [11].

Model [12] introduced new data and text implant model, which is derived from the treatment of answering multiple queries in the selection as an input problem. SCITAIL is the first input set created only from natural sentences that already exist independently "in the field" rather than sentences written specifically for the input function. In addition to the required input details, they construct add-ons from scientific questions and related answers, as well as structures from relevant web phrases taken from the larger corpus. These sentences are often a language challenge. This, combined with the high similarity of the baseline and hypothesis in both paired and non-entailed pairs, makes this new entailment work even more difficult.

Model [12] present the largest input data taken directly from the final task and contain text that appears naturally as both the basis and hypothesis. The model new data, SCITAIL, is designed for the final task of answering high school science questions. Each question and their relevant answer is converted into a supporting statement for hypothesis H. they use the retrieval (IR) method to obtain relevant text from the business of large web text documents, and use one of these sentences as Layout P. -H, not all that includes P or “supports” the statement in H. They include annotations for each hypothesis structure head as a support or not, to create SCITAIL input data with 27K examples.

Since both the foundation and the drawings in SCITAIL were written independently of each other and without the inclusion function, the linguistic diversity of the database is not limited to the coverage of handwritten rules or by crowds of clever staff two pieces of text [12].

They show that one can improve SCITAIL accuracy by 5 percent using a new neural model that exploits language structure. And found that current RTE systems, including neural input models, have interoperable functionality in this data, whether they are trained in their data or in SCITAIL. Specifically, their asymmetric Decomposed Graph Entailment Model (DGEM) raises an accuracy by 2.3%. They make the following contributions:

1) A natural input database in which the text and transcripts were written independently of each other and independently of the input function;

2) Initial data taken from the final task of answering multiple questions; and

3) A new model that uses language more recklessly in the hypothesis than the existing strategies in this data.

Recently, text submissions have received significant attention due to their extensive use of retrieval information, commendation programs and answering questions. Information extraction (IE) is an Automatic retrieval of certain information related to a selected topic from the body or bodies of text. Extraction tools enable us to retrieve data from text, information, websites or multiple sources.

Recently, various frameworks for textual entailment recognition have been proposed, ranging from traditional linguistic strategies pertaining to different deep learning methods. However, the deep neural networks that reach the state of the art in textual entailment task look only at the details of the given sentence structure rather than the real world knowledge and information beyond context. In [13], they propose a Knowledge-Context Interactive Textual Entailment Network (KCITEN) that reads graph-level presentations by tying a graph of external information through a graph-focused network. They also propose a Text-graph interactive reading system for neural learning entailment, which allows for unwanted noise and noise and focuses on informative presentations. Examination in the SciTail database shows that KCI-TEN works better than state-of-the-art methods [13].In addition to the practical application of previous studies, textual entailment recognition models have two limitations as follows:

- Underutilize the large-scale real-world knowledge: Existing RTE models may suffer from language challenges due to the high similarity of the lexical premise and the hypothesis of both blocks included and excluded.
- Perform learning and inference between text and large scale KG without explicit interaction: These methods [13] are useful for easy comprehension of concepts but can lead to a lack of communication between text-based embedding and KG-based embedding during the editing process.

To address these two limitations , in [13] proposing a novel representation framework, the Knowledge-Context Interactive Textual Entailment Network (KCI-TEN), which makes full use of the external KG to determine whether a given text incorporates a given hypothesis. Specifically, the proposed approach begins to convert the concepts and basic sentences into KG sub graph, incorporating them into the context and contexts of the environment through a graph attention network (GAT), providing an additional incentive to explore external sources of textual recognition. Thereafter, the included vectors are used to calculate the basis for comparison-hypothesis corresponding to the text level and graph level respectively. Finally, using a method to read the matching representation in size, using practical information from text and KG, and re-inserted into the classifier to get textual entailment [13]. The model concise contributions are:

- Creating an effective RTE communication framework, which directs the external KG to obtain background information and understanding.
- Investigating the feasibility of graph-based submission by approving GAT to carefully insert KG sub graphs.

The results of the experiments indicate that KCI-TEN successfully incorporates the conditions in terms of KG sub graphs [13].

Recognizing inclusion relationships between sentences is an important and common part of language communication. RTE function has been proposed as the solution to this problem. In [14], they introduce the Arabic Recognizing Textual Entailment Tool called Ar-SLoTE “Arabic Semantic Logical Textual Entailment Tool”. Model [14] suggested tool is made up of five modules: pre-processing, linguistic analysis, first-order logic representation, features extraction and entailment decision modules. It produces logical representations of hypothesis / text pairs in order to extract important and informative features, namely, predicting the dynamics of contradictions, exactly the similarities.

Ar-SLoTE is mainly based on Arabic question / answer procedures and the results obtained are very encouraging. Obtaining RTE In many NLP applications is by performing many tests, including the Question-Answering (QA) programs. Therefore, RTE helps to determine whether the given answer entails the question asked or not. There are various methods suggested in the literature to determine which entailment into this field. However, in the Arabic language, there is a shortage of RTE services provided in QA programs.

As long as, in [14] they are introducing an Arabic tool called Ar-SLotTE «Arabic Semantic Logical Textual Entailment Tool». For defining the entailment between a couple of question and answer they propose this tool. The correlation of the entailment is set as "TRUE" when it means how the question can be estimated from the answer and "FALSE" in another way.

Model [14] is based on the first logical presentations of the structure orderly to extract new semantic information from the H-hypothesis and the T-text. Three factors are excluded from the logical expression of T and H namely, the accumulation of Predicates-Arguments, the Semantic similarity and the similarity of the Name entity. A rating is given for each feature. After that, the feature vector is designed for each H / T pair. These doses are passed on to a tree decision reader to guess the category of each couple: «TRUE» if there is an entailment between them, and "FALSE» in another way. Ar-SLoTE is designed to detect the entailment relationships that exist between Arabic H/T. In particular, it sees the entailment between the true response and its answers. Noting that the real question is a question related to an entity with names.

In this situation, the hypothesis indicates the question asked and the text identifies the given answer. Thus, Ar- SLoTE takes, as an input, questions that start with test nouns (من who,ما what, أين where, متى when) and a list of related candidate responses. At the last, the entailment between the question and each answer was indicated. Research correspondent's response is considered an appropriate answer to a question [14].

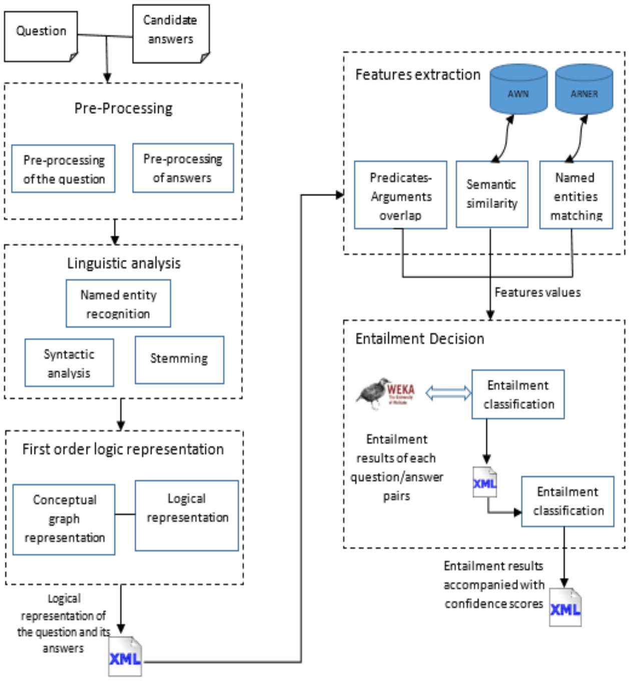

Question
Candidate answers
Pre-Processing
Pre-processing of the question
Pre-processing of answers
Linguistic analysis
Named entity recognition
Syntactic analysis
Stemming
First order logic representation
Conceptual graph representation
Logical representation
Logical representation of the question and its answers
XML
Features extraction
AWN
ARNER
Predicates-Arguments overlap
Semantic similarity
Named entities matching
Features values
Entailment Decision
WEKA
Entailment classification
Entailment results of each question/answer pairs
XML
Entailment classification
Entailment results accompanied with confidence scores
XML

Fig. 4. Ar-SLoTE Architecture [14].

As shown in Fig. 4 Ar-SLoTE architecture is built on five key modules, namely: pre-processing editing, linguistic analysis, first-order logic representation, feature extraction and entailment decision modules.

Model [14] approach focuses on the exclusion of elements from the two logical pathways taken from T and H, specifically, predicting contradictions, semantic similarities and word similarities. The Ar-SLoTE releases a file from type XML containing the accompanied results of entailment using confidence scores. Therefore, the result can be shown and translated by the user, or by machine.

As the [14] tool is designed for question / answer processes, authors have selected 500 operators of a variety of authentic questions and answers from the AQA-WebCorp Corpus. At first, they turned them into logical submissions. Second, the corpus was divided into two data sets: training set (350 pairs) and test set (150 pairs).

Reported to the following TABLE I, the results obtained by the proposed tool in [15], Ar-SLoTE and LR-ALL. Among the state-of-the-art technology that works for Arabic RTE, the tool proposed in [15] is the only function specifically intended for real-time question / answer systems. LR-All [16] Arab RTE system with highest accuracy. This paper uses traditional embedding features and words. In addition, it is based on machine learning and uses ArbTEDS dataset, the only available Arabic RTE data containing 600 pairs of T-H.

They have been able to be more accurate compared to the previous RTE system which was intended for question / answer applications (Ben-sghaier et al., 2018) [15]. Because the results obtained in [14]show that reasonable representation and appropriate relevance can improve the results of Arabic RTE. Authors said that Ar-SLoTE performance can be improved by integrating, in particular, additional cognitive information. One of authors' plans that focuses on testing their RTE tool on a large computer uses a large number of different questions / answers per pair. As a long-term quest, they aim to look at other types of questions [14].

TABLE I. COMPARISON BETWEEN AR-SLOTE AND OTHER ARABIC RTE SYSTEMS

System	Approach	Dataset	Accuracy
Ben-Sghaier, M. , W. Bakari, and M. Neji [15]	• Machine Learning based . • Use of semantic measure and word sense disambiguation.	200 Question/Answers pairs	70%
Ar-SLoTE	• Machine learning based • Use of logical representation and extraction of features	500 Question/ Answers pairs	73.33%
LR-ALL [16]	• Machine learning based • Use of traditional and word embedding features	ArbTEDS	76.2%

C. Information Verification

As the name itself implies, data verification is a process that verifies the authenticity of a person or company for a particular process that requires verified information. This process helps to verify the current status, including features such as hiring and accommodation, allowing an application that requires specific references to be completed by the preferred person or organization of applications.

Information verification is important because the level of information production is high and growing daily, usually on social media. This also results in social media being included as a mass media. Therefore, data verification in social media becomes even more important [6].

Information verification is a form of journalism investigation in which experts examine claims that have been published by others about their truthfulness. Claims can range from statements made by members of the public to stories reported by other publishers. The ultimate goal of a fact-checking system is to determine whether a claim is true, false or mixed.

In [6], a method of verifying information on Twitter is proposed. The proposed method uses textual entailment methods to improve the authentication methods on Twitter. Separating the effects of entailment methods over state-of-the-art methods can produce tweet verification results. In addition, as the writing style of tweets is incomplete and structured enough for textual entailment, they have used the language model to add tweets with official and appropriate text for textual entailment.

While the methods used to incorporate reinforcement inputs to verify data may have acceptable results, it is not possible to provide relevant sources for all tweets, especially in the past by posting tweets. Therefore, they have used other sources such as User conversational tree (UCT) without using input methods to verify tweet details. UCT analysis is based on pattern release at UCT. Test results show that using entailment methods increases tweet validation [6] as shown in the model architecture in Fig. 5.

Looking at the various challenges in obtaining rumors, they look at the findings of the rumor by the two-source analysis: (1) entailment based classifier, and (2) UCT-based classifier. Then, in each of the two sources, they train two different classifiers separately, and after that, use a weighted ensemble voting classifier, the results of these different classifiers combined to form a new classifier [6].

All in all, contribution to [6] to verify the details on Twitter is:

- Textual entailment is used to confirm rumors on Twitter
- A language model is used to make tweets more acceptable in the style of writing
- Consideration under UCT analysis
- Promote a weighted phase of voting to integrate the results of the inclusion process and UCT analysis

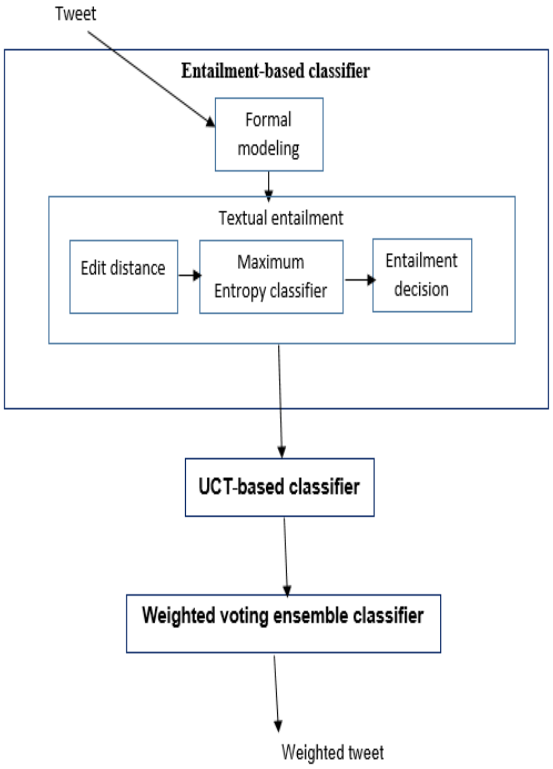


Fig. 5. Model [6] Architecture.

Information on the web is linked to rumors and unverified information. In addition, social networks as a special and broad part of the web have a lot of potential for spreading and creating inaccurate or unauthorized information.

Because of the importance of this issue, as well as improving the effectiveness of data verification, in [17] verification of data on social media is being investigated. Several features and conditions appear to be applicable to the detection of rumors.

Among the functional and structural features, authors consider two important sources of data validation in social media that include user responses and news organizations. User feedbacks as the primary source can be a user conversional tree. Some patterns can be drawn from this tree. Media agencies as a second source are also used for data verification by tools of text entailment. Finally, these two types of factors are combined to separate information into one of three categories true, false, or unconfirmed. This method was tested by checking public data sets. Test results show that the proposed hybrid information verification method can surpass highly recommended methods in data verification [17].

Today, everyone has some doubts about the information broadcast on the web or social media. This is because the amount of information that is visible on the web is wrong or unreliable. Moreover, people are exposed to all kinds of fake and uncertain information. Doubts about the facts of the information available on social media create public concern, especially in difficult situations or disasters. A large number of online false stories have the potential to cause great social problems. Stock markets has been affected by fake information on the web and social media, reduced responses during disasters and terrorist attacks [17]. Typically, bots collaborate to create and disseminate fake or wrong information on a large scale using bots accounting groups that work on a large scale for two purposes:

- by distributing the same content, e.g., by retweeting, to multiple viewers, and
- Increase social status by following each other's accounts and following trust information [17].

Those who try to stop the spread of wrong information are working to create programs that can reduce the way bots and other programs spread lies and slander. Most of the previous methods depend on the selection of the feature from the main text. In addition, the methods studied in the structure of graphs in social media [17] are popular with the rich history of graph reading. Recently, more comprehensive learning approaches have also emerged in this field. For example, recurrent neural networks (RNNs) are used to detect rumors from microblogs.

However, among current methods, the effect of resources on data verification cannot be ruled out. Therefore, in [17] to strengthen the validity of the data verification, the results of the sources are studied. The proposed method of this data in [6] uses the main sources of information, mainly feedbacks, and media sources. Initially, the effect of user feedback and media on the data verification function are studied separately.

Subsequently, these methods' results are gathered to determine the validity of the information. In the user feedbacks, the composition of the marked areas in the user conversational tree (UCT) by four types of tags, rejected, comment, supporting, and query, is considered. UCT is being studied discussed in the following three ways:[17]

- Pattern extraction: For UCT pattern extraction, the unique UCT patterns are calculated by looking at the pattern level in the tree. In these patterns, a tag for each node is selected, too.
- Statistical sequential models: By studying UCT sequencing models, two different types of sequence models are calculated
- Hidden Markov model (HMM)
- Conditional random fields (CRF).
- Edit distance: Set UCT learning methods used as standard K-nearest approaches to UCT classification with data verification function.

The second source of data verification is the media sources, which is useful for textual entailment methods. At this opinion, the text that feels authentic (called the main text) is considered to be either media or not. If not entailed, the main text is considered authentic. When the media compares with the main text, the main text is considered false. Otherwise, the main text is considered unverified. Suggested data verification methods are evaluated using public records by comparing the state-of-the-art-methods.

The results of these comparisons show that the proposed method works better than the existing alternatives presented in [17] database used.

The description of the method starts with the method's full description in the first subsection. After that, the two main parts of the method are explained, that were namely:

- Studying of user feedbacks and various components
- News sources studying

An overview of the proposed [17] approach to data verification activities is shown in Fig. 6. As shown in the figure, the input for data validation is a message from a social network.

After reviewing the input text, you are ready for verification. Using two different steps, verification work will be done: it processes feedbacks of users and new sources.

1) User feedback study: User response research is typically based on the UCT structure. UCT is a tree created when a message is communicated on social media or is often responded to. The root of this tree is the main message its authenticity is being questioned. After all, every response to this message is a child root. Previously, every response of each node in a tree is the next level of UCT.[17]

Thereafter, the UCT tagged is used to predict the final label of the input text using the following three methods in the next three paragraphs. UCT is studied in three ways including:

2) Pattern extraction: UCT derived pattern are considered features. These patterns are below each starting point with a maximum of three. Calculation of each pattern's occurrence is by looking at the weight, which is calculated by the level of the root scale relative to the height of each UCT.

3) Statistical sequential models: HMM and CRF are used in [17]proposed methods. The path label with the most voted priority is false, true, and unconfirmed is considered the last label.

4) Edit distance: Set the distance of each path at UCT calculated the most effective dynamic time strategy. this method uses a powerful planning method [17].

5) Study news sources: recognizing textual entailment

In this case, a source of information should be provided. When the main text, in which the message is confirmed as true message, is entailed from the new source, the source text is the first text. If the new source differs from the main text, then the main text is fake or wrong. Otherwise, the main text is not verified. The following methods of textual entailment are used in [17]:

- Edit distance comp (Fixed Weight Lemma/RES Word Net).
- MaxEnt Tree Skeleton RES (Verb Ocean Tree Pattern/ Word Net Tree Pattern/Word Net Verb Ocean Tree Pattern).
- P1EDA RES: Paraphrase Table.

To compare the [17] method with the state-of-the-art methods in which authors simulate tests similar to those performed in the Semeval-2017 work 8. The test methods are similar to this work, too. Test methods are Score, Confidence, Root mean Square Error (RMSE), and Final Score. The Score is ranked as the most known accuracy measure. Confidence RMSE is the RMSE Classified Confidence.

Existing new sources are critical to the success of this approach. And on some social networks, user feedback is not recorded and in this case this method is bad choice for data verification. In addition, the [17] method only applies to text messages. Examination of the this method of data verification in public dataset showed that the [17] method surpassed those who are technically competent in data verification.

Model [18] introduces ColumbiaNLP's submission of the FEVER Workshop Shared Task. This model is an end-to-end pipeline that delivers factual evidence using Wikipedia and provides a decision on the veracity of the claim based on evidence-based evidence. The FEVER shared function aims to test the capabilities of the data verification system using evidence from Wikipedia. Given a claim that includes one or more items (mapping to Wikipedia pages),the system in [18] must take written evidence (sentence sets on Wikipedia pages) supporting or disputing the claim and using this evidence, they must write a claim as Supported, Refutes or Not_Enough_Info. The shared work dataset is presented by [19] and contains 185,445 claims. *TABLE II* shows three scenarios from a set of claimant data, evidence and decision. The first system that used by [19] uses 3 main components:

- Document retrieval: using the documentation recovery component from the DrQA system that retrieves the adjacent text query using the cosine similarity between attached unigram and the bigram Term Frequency-Inverse Document Frequency (TF-IDF) vectors.

TABLE II. THREE SCENARIOS FOR CLAIM AND EVIDENCE AND DECISION OF [19]

Claim : Claim : Fox 2000 Pictures released the film Soul Food. [**wiki/Soul_Food_(film)**] **Evidence**: Soul Food is a 1997 American comedy-drama film produced by Kenneth "Babyface" Edmonds , Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures . **Verdict**: SUPPORTS
Claim : Murda Beatz's real name is Marshall Mathers. **[wiki/Murda_Beatz]** **Evidence**: Shane Lee Lindstrom (born February 11, 1994), known professionally as Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario. **Verdict**: REFUTES
Claim : L.A. Reid has served as the CEO of Arista Records for four years. [**wiki/L.A._Reid**] **Evidence**: He has served as the chairman and CEO of Epic Records, a division of Sony Music Entertainment, the president and CEO of Arista Records, and the chairman and CEO of the Island Def Jam Music Group. **Verdict**: NOT ENOUGH INFO

- Sentence Selection: [19] simple sentence selection method sets out the similarities between TF-IDF and claim. They set the most similar sentences first and then using validation accuracy on the development set, they accurate a cut-off. Then they test both DrQA and the easy use of the unigram TF-IDF to measure selection sentences. They also examined the effect of sentence selection on the RTE module by predicting the entailment given to the original texts without sentence selection.

- Textual Entailment: It is a multi-layer perceptron (MLP) with a single hidden layer that uses frequencies of term and TF-IDF cosine similarities between claims and evidence as features. To analyze the state-of-the-art in RTE, they have used the decomposable attention model (DA) between the claim and the role of evidence.

Overall, [18] end-to-end model shows an improvement of 19.56 on FEVER results compared to the previous [19] (50.83 vs. 31.27) in the development set. system [18] of utilizes changes in all modules that lead to significant improvements in both development and testing sets:

- Document Retrieval: this step is an important step in establishing an end-to-end authentication and verification system. Missing the correct text can lead to untranslated evidence, while incorrect text can add to the noise of subsequent sentence selection task and textual entailment task [18] . In [18] authors suggest a multi-step approach to retrieving documents appropriate to claims:
    - Google Custom Search API: inspired by previous research [20], the Custom Search API of Google was used first to retrieve documents having information about the claim. The token Wikipedia was added to the claim and produce a query and gather the top 2 results.
    - Named Business Recognition: Identifying an entity name using a previous activity, and after finding the entities that be named in the claim, authors use the Wikipedia python API 3 to gather a high-rated wikipedia document returned by the API for each named entity.
    - Dependency Parse: to increase the chance of finding the right entities in the claim, they get the first sentence of lowercase letters, Reason to emphasize the lower case verb phrase to avoid non-claim entities such as "Finding Dory was directed by X", where the appropriate entity is "Finding Dory" [18].Then they decide to create better accounting models to handle entity ambiguity or entity linkages in the future.
    - Combined: Authors use the documentation union returned by these three methods as the final set of relevant documents to be used by the Sentence Selection. They note that their (combined) method in [18] reached a high accuracy of 94.4% compared to the original method [19] of 55.3%. Because authors do not have best evidence of blind set test they cannot report claim coverage using their pipe.

- Sentence Selection: authors used a component to retrieve translated texts of DRQA [20] to sentences selection using the bigram TF-IDF for binning as suggested by [19]. They extract the top five sentences in the appropriate text k-most the related documents by using TF-IDF vector matching. To solve the previous problem that arose in selecting sentences of [19] authors continued to filter out the top 3 evidence from chosen 5 evidences using distributed semantic representations. Model [21] show how the deep representation of words modulates the complex aspects of word use (e.g., syntax and semantics), and its use in a variety of language contexts. Therefore, converting claim and evidence to vectors using ELMo embedding. then [18] calculated the cosine similarity between the claim and the vector of evidence and produced the top 3 sentences based on their score.

- Textual Entailment: authors did not show evidence, but trained model for each pair of evidence. For textual entailment recognition they used the previous model in their work on the supervised learning of the representation of universal sentences. They use bidirectional LSTMs for max-pooling to encode a claim and evidence. In word matters in pre-trained word embedding it is a major stumbling block to sentence representation. To resolve this they use fast text embedding based on subtitle information. Also, this embedded training in the Wikipedia corpus makes them the right choice for this method. Their final predictions are relied on the next algorithm in Fig. 7.

Because the selected evidence was naturally noisy and this pipeline [18] did not verify the evidence together they chose this rule rather than more prediction to reduce the rule of prediction of the NOT_ENOUGH_INFO class.

```
Algorithm 1 :
if count(Support)=1 and counr(Not_Enough_Info)=2 then
print(Support)
else if count(Refutes)=1 and count(Not_Enough_Info)=2 then
print(Refutes)
else
print(arg max((count(Support), count(Refutes),count(Not  Enough  Info))
```

Fig. 6. Algorithm Used To Make A Decision In Module.

TABLE III. FEVER SCORES ON SHARED TASK DEV AND TEST SET

Data	**Pipeline**	**FEVER**
DEV	[19] module	31.27
	This module	50.83
TEST	[19] module	27.45
	This module	49.06

Module [18] also tried to train a classifier that takes confidence scores from all of the claim/evidence and their status in the text and trained an improved classifier but the accuracy was not improved. Experimentally the rule has given them good results in the development set and thus they used it to get the final label. *TABLE III* shows the total FEVER points obtained by [18] pipeline in development and test sets. In that time ranking [18] system ranked 6th.

The new models cannot rely on models that rely entirely relied on semantics. Although the two sentences are similar, the similarity of the cosine between them is worse because the evidence contains many additional details that may be inconsistent with the claim and difficult for the model to understand. They also found cases where the predicted evidence was accurate but not consistent with best evidence. But this system is being penalized on not being able to match the best evidence. Module [18] has found a few annotations that are incorrect which is why FEVER scores are lower than expected. Sometimes, the lines between SUPPORT and NOT ENOUGH INFO are not cleared enough. This models need a better understanding of semantics in order to be able to distinguish between them in all cases.

D. Sentiment Analysis

Sentiment analysis (emotional AI or mining opinion) refers to the use of natural language correction, text analysis, related languages, and biometric to directly identify, extract, measure, and read the corresponding regions and information.

Sentiment Analysis tools now are slow, and it is the most difficult statistics that make the work well done. Sentiment analysis can be subdivided on the basis of a separate textual text. It can be categorized at three levels discussed briefly below:

- Document Level Analysis: analyzes the contents of the document as a whole to determine its availability.
- Speech Level Analysis: Pays attention to the various sentences of a document by continuing to break it into small word texts to analyze its structure.
- Word Level Analysis: Determines the minimum number of different words in a sentence in relation to an item or event.

Previous work defines hypothetical analysis as research to determine emotional AI. Sentiment analysis work involves finding out the feelings, status, person's opinion, object, product or event that has worked in areas such as academics, business and industry. Sentiment analysis work can be categorized on the basis of different classes - can be analyzed as positive, negative or neutral based on class.

Appropriately, authors cannot accurately measure guidance using TE for a number of reasons such as complexity of sentences and limitations of available language resources. Therefore, (PTE) is using to measure the degree of inclusion whether the text fully integrates the concept or not. The authors plan to use Partial Textual Entailment (PTE) to improve the work of Sentiment Analysis in [22]. It is expected to reduce the amount of work done in analyzing the text view thus providing a better solution for SA work. The method proposed in [23] is to carry out the task of analyzing the sentiment by exploiting the concept of partial text entailment. They suggest using the partial textual entailment to measure semantic similarities between tweets in order to combine the same tweets. The method is expected to reduce the burden of sentiment analysis and make processing faster. Then, authors propose to change the method of partial text entailment that can be adapted to most NLP applications. The purpose of [22] is improvement the task of sentiment analysis using partial textual entailment.

The authors propose a two-part Sentiment analysis approach:

a) A method to improve the PTE recognition process and

b) To improve the Sentiment analysis method using the partial textual entailment.

E. Designing a Method to Improve the PTE Recognition Process

The proposed idea is to expand the add-on model to increase due to the large number of feature builds and produce relevant results that done by deleting helpful words. Elements will be constructed only using the sentence title as the key word in the construction of the elements. This approach is expected to predict the best of text input as the text revolves around the topic of the sentence[22].

1) Choose text and Hypothesis

2) Re-analysis of the text and Hypothesis by removing helpful words

3) Subject identification: Subjects from sentences are selected from the remaining words written separately.

4) Structure construction: by matching the remaining words with the subjects

5) Composition of partial text entailment will be performed

F. Creating a Way to Improve Emotional Analysis using Partial Text Entailment: Authors Discuss the Entire SA Framework using PTE Below

1) Recognizing new models or methods for Partial Textual Entailment: this step will do the job of collecting similar tweets to reduce the burden of sentiment Analyzer which has led to the rapid implementation of updates. Integration recognition will be done using the BIUTEE tool.

2) The result obtained in Step One will be forwarded to a mood analyst to get the feelings of the speech.

The total number of sentences that entail the hypothesis will then be provided as an estimator of how strong the sentiment is[22].

Model [23] aims to increase the accuracy of the entailment of Arabic texts using resolving ignoring of the text-hypothesis pair and determining the splendor of the text-hypothesis pair whether it is Positive, Negative or Neutral. It is noteworthy that the absence of a negation detection factor provides negative results when detecting the correlation of the factor since contradictory returns true.

Contradictory words are considered termination words and are removed from the text-hypothesis pair which can lead to improper entailment decision. Another case that has not been resolved before, it is impossible for the correct text to include the wrong text and also the same for the opposite order. In [23], in order to distinguish the text of the hypothesis pair contradiction, a sentiment analysis tool is used.

Authors show that processing the essence of a text-hypothesis pair increases the accuracy of the entailment. To test [23] method of the ArbTEDS archive (ArbTEDS) dataset which comprises 618 text-hypotheses pairs and shows that the accuracy of Arabic entailment increases by resolving the inconsistencies of the entailment and analyzing the unity of the hypothesis pair. The main problem with the original Arabic texting systems is that they do not see the negation, where negation brings back the truth and does not take the critical side where negative or positive thinking does not convey the opposite feeling. To their knowledge, [23] represents the first attempt to recognize the influence of resolving negation and emotional polarity in the Recognition of Textual Entailment in the Arabic language. They conducted research on the Arab Text Entailment Dataset (ArbTEDS) [24]; shows that the more accurate the discovery obtained by resolving the negation and the more consistent finding of this text-hypothesis.

In this work, authors face the challenge of creating an Arabic text-entry system that produces relevant results. The proposed project in [23] ,that it is shown in Fig. 8,is called Sentiment Analysis and Negation Resolving for Arabic Text Entailment (SANATE) which enhances ATE method (ATE Method is described in [23]).

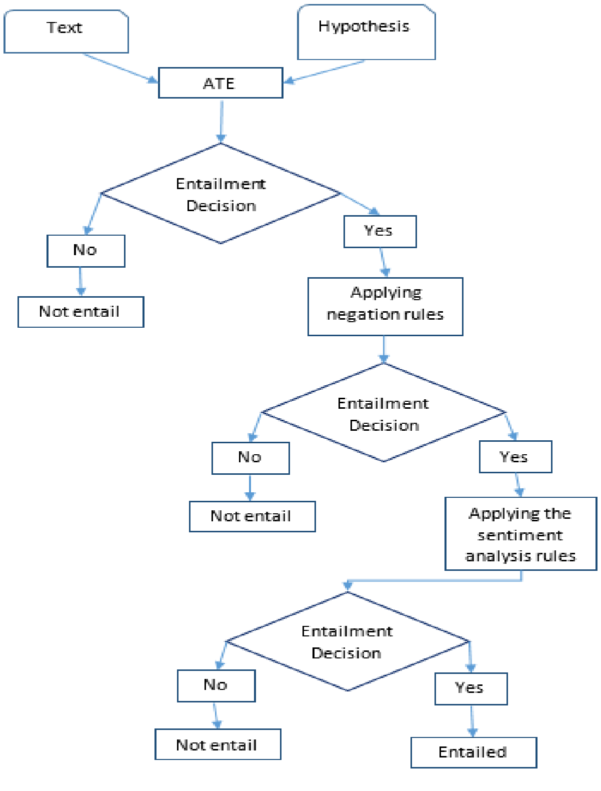


Fig. 7. General Diagram of Sanate System [23].

To improve ATE, they set some rules for negation. If the decision to include ATE is "not entail" then no further consideration is included. If the decision to include ATE says it will be "entail" then a set of rules will be considered. The rules used to do some text and hypothesis entailment include [23]:

1) If negative particles appear before the same action (normal action) in the T or in the H, the judgment will be: not entails

2) If the negative adverbs appear before the normal action in the T and in the H (both), the judgment is: entails

3) If T and H have a different verb and are negative particles appearing before the verb in the T or in the H or in both, the judgment is: NOT entails

4) If the text-hypothesis pair has more than actions (verbs), then:

a) If one of the normal actions is ignored by one of the negative particles in T or H but is not ignored in both (T and H), the judgment is not entail.

b) If the normal action is ignored by one of the negative adverbs in T and H (both), judgment will be entail.

5) If text and hypothesis contain Opinion words (R):

a) If the result R is not the same as T and H which means different variations, the judgment is not entail.

b) If output R is the same for T and H, judgment will be entail.

To analyze how they use it they have used the 618-text-hypothesis archive (ArbTEDS) database. Each pair of ArbTEDS text is integrated into the ATE system and the SANATE system. Calculation of the accuracy for each system is made by the accuracy of the equation: the measure of the determination of the entailment target to the total value of the entailment problem. ATE accuracy is 0.617 and SANATE accuracy is 0.693. From the impact of the ATE and SANATE system, it is shown that resolving disregard and dividing a text by its contradiction by analyzing emotions, improves the effectiveness of co-identification and non-entailment relationships[23].

Authors conducted research on ArbTEDS; which shows that it is a more accurate result to find entailment to negation and to analyze the unity of this text-hypothesis pair. Without resolving the entailment decision to engage in negation authors may argue, because negation gives the opposite of reality. Some texts may have entailed but the presence of negative particles alters the decision of entailment relationships [23].

Another result from this experiment, finding text and hypothesis pair polarity has a significant impact on finding entailment related to non-entailment. It is impossible for a positive idea to combine negative ideas with the opposite[23].

Fundamental topic in psychology is understanding what makes people in a good mood .Previous work has focused on the development of individual reporting estimation tools and relies on experts to analyze estimated reports from time to time. One of the objectives of the analysis is to understand what is needed to promote individual behavioral change in order to improve the well-being of all [25].

In [25], they set an integral approach; with the view that the user incorporates his or her enjoyable moments as short texts, the system can analyze these texts and provide stable user suggestions that may lead to general improvements in his or her well-being. Authors set one required part of such a program, the Happiness Entailment Recognition (HER) module, which takes as an insert a brief description of the event, a candidate proposal, and outputs the decision of whether the suggestion may be appropriate for this user depending on the event specified.

This part is used as a neural network model with two encoders, one for user input and one for sustainable suggestions, with additional layers for capturing psychologically important features during fun and suggestion. The achievement of [25] the AU-ROC of 0.831 and exceeds their base and current AllenNLP Textual Entailment model by more than 48% development, ensuring the distinction and complexity of HER work.The HER module is described in Fig. 9. This module (1) detects candidate suggestion, and (2) identifies which suggestions are reasonable for specific users. As an input in this module, users create short-term journal entries for "good times," proven interventions to improve subjective well-being.

The first task defines proposals as sustainable in the sense that they promote short-term prosperity without harm in the long run, even if the same suggestion is repeated. For example, a user may report that receiving a puppy or buying a new car has made them happy. However, they do not want to recommend doing them again because repetition is impossible. On the contrary, activities such as swimming or walking in the park have the potential to improve the well-being of the user and are repetitive, These activities were viewed as probable suggestions [25]. Such suggestions become part of this DB Proposal. Appropriately, field experts (e.g., psychologists) would be able to create those consistent suggestions, but [25] also develop a classifier that reflects exciting moments with sustainable activities, and ensure that the classifier accomplished AU-ROC of 0.900.

Second task present the problem of the happiness entailment recognition (HER), which is motivated by the problem of recognizing textual entailment (RTE) in the NLP: given a set of fun times reported by the user, and a set of possible activities, which activities can make the user happy? Model [25] is developing an improved neural network (NN) model with information about concepts, agency, and personalities to predict with the AU-ROC of 0.831 whether the suggestion matches the moment of happiness.

"Concepts" represents topics that are common to HappyDB as an important vector. “Agency” and “personality” are binary indications that the author was in control of their happy time, and that there were other people involved, respectively. They say that this classification includes psychological ideas that are important to that task [25]. The results of the experiment [25] indicate that in this task, the proposed approach exceeds the recognition model for the textual entailment, as well as their foundation. The results support that the HER function is very different from normal text entailment task, and that adding a mental view of the model improves performance.

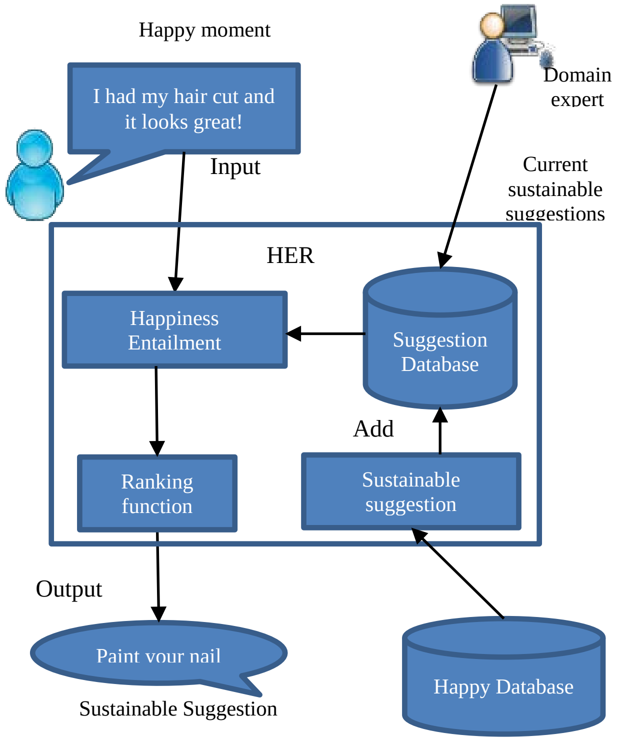


Fig. 8. Her Module and its Inputs and Outputs [25].

III. Machine Learning vs. Deep Learning in Textual Entailment

A. *Machine Learning*

Machine learning is used at different stages in the RTE process (i.e. pre-processing, alignment and decision steps), which is usually combined with previous methods. A general model presents this task as a classification problem that is supervised by two classes (entails, not entail) or three classes (entails, not entails, contradiction). Each pair (text, hypothesis) to check is represented by a feature vector, which includes the scores of several operations applied to the pair at different levels[26].Some of described papers used Machine learning to enhance their models . These techniques are Support Vector Machine (SVM), Glove Embedding, Extreme Learning Machine (ELM), decision tree classifier, Naïve Bayes, and Root Mean Square Error (RMSE). All of these techniques are really enhancing the papers result but still in small datasets, and these results compared to deep learning, deep learning enhanced the model in results and with bigger datasets was detected as shown in next table.

B. *Deep Learning*

Deep learning strategies are becoming increasingly popular in the work of entailment text, overcoming the complexity of different traditional models with complex alignment and ideas. Deep learning in neural networks has become a popular method of machine learning due to the recent success of computer vision, speech recognition, and other areas. This latest achievement is a direct result of being able to train in big data sets, from labeled images for object recognition to compatible machine translation texts. While such information has been widely available from public sources to date, confidential information collected from individuals will not only provide incentives for existing applications, but also provide new impetus for deep learning. Deep learning is a key technology behind the most recent applications. In deep learning the achievement tasks that directly detect text, image or sound can be learnt by computer models. High accuracy was achieved in deep learning models, sometimes exceeding human-level performance. Textual Entailment approaches will help in many of natural language processing applications, including question answering, summarization, text generation, machine translation, and information extraction. In previous described papers that used several deep learning techniques as Text rank, long short-term Memory (LSTM), Multilayer perceptron (MLP), bi-directional Gated Recurrent Unit (bi-GRU), and Graph Attention Network (GAT).these techniques results with shown next.

TABLE IV. TEXTUAL ENTAILMENT PAPERS

Paper	**ML vs DL**	**Method Used**	**Comments**
Sentence similarity estimation for text summarization using deep learning[10]	DL	Text Rank	Use it in sentence similarity
A New Approach for Twitter Event Summarization Based on Sentence Identification and Partial Textual Entailment[9]	ML	Rank SVM	Use Word2vec to measure word-to-vector similarity of a word give high score compared to another techniques during Rouge metrics. word2vec model is not a deep learning model. It just a library in deep learning
Text Summarization using Partial Textual Entailment based Graphs[7]	No		Used only for single document with small fragments of sentences
A textual entailment dataset from science question answering[12]. DGEM	DL	LSTM MLP	to capture the semantics compute the representation for each node This model raises the accuracy to 77.3%.
Knowledge-aware Textual Entailment with Graph Attention Network [13]	Both	GloVE embedding Bi-GRU GAT	Bi-GRU to encode the contextual patterns Using GAT to encode node contexts in various and noisy KG subgraphs instead of using fixed pre-trained embedding for nodes. Its technique raises the higher accuracy of previous method by +0.8
Ar-SLoTE: A Recognizing Textual Entailment Tool for Arabic Question/Answering Systems [14]	ML	Weka DT	Uses a small set than ArbTEDS, and their model get less accuracy than LR-ALL, but compared to another module its accuracy exceeds another module with 3.33%
Information verification improvement by textual entailment methods [6]	ML	ELM	They want to study more special pattern of UCT ELM is used in deep learning networks but researchers argue if it is considered as a deep learning technique by itself or not even it has only one hidden layer as described before
Information verification in social networks based on user feedback and news agencies [17]	Both	ELM Naïve Bayes SVM RMSE	MLP is used in deep learning techniques but it is not considerd as DNN because it only forward direction Work better than another state-of-the-art-merthod that use the same dataset. The results of hybrid method show that it is more efficent than the results of separate use of user's response approach and entailment approach individually.
Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification [18]	Yes	Bi-LSTM	Blind set, score = 49.06 > baseline's score= 27.45 dev set , score =50.83 > baseline's score = 31.27
Sentiment Analysis using Partial Textual Entailment [22]	No		This method is not implemented yet , using extended The BIUTEE for RTE
A study of the effect of resolving negation and sentiment analysis in recognizing text entailment for Arabic [23]	No		It has a major cons where it can detects polarity of a text and hypothesis pair significantly affects the detection of entailment and non-entailment
Happiness Entailment: Automating Suggestions for Well-Being [25]	Yes	Bi-LSTM	Bi-LSTM = 0.900 is more accurate by 0.133 with the best accuracy (Logistic Regression)

C. Difference between ML and DL[27]

- Data dependencies : Machine learning can trace on less data where deep learning requires large data
- Hardware dependencies : Machine learning can be traced on normal CPU but deep learning requires high performance machines with GPUs to be trained probably
- Feature processing: In ML, most app features must be determined by an expert and entered as data type. Values, shapes, textures, locations, and orientations are features of Machine learning. The performance of many ML algorithms depends on the accuracy of the extracted features. Attempts to obtain high-quality features directly from the data are significant differences between DL and traditional machine learning techniques. Therefore, DL minimizes attempts to design a feature to extract for each problem.so DL has higher accuracy than ML
- Problem solving method : machine learning use many sub-problems and solve them where deep learning is direct end-to-end problem solving
- Interpretability: DL can be tuned in various different ways, while ML has limited tuning capabilities.
- Execution Time: Deep learning in general takes a long time to be trained where some machine learning algorithms have short training and test times.

From the previous comparison, we concluded that deep learning performs better than machine learning. Unlike machine learning, deep learning has proven its ability to handle massive amounts of data. Deep learning can also solve problems that machine learning cannot provide, giving higher accuracy and reducing time. *TABLE IV* shows the textual entailment papers comparison that clarifying method used and whether machine learning or deep learning is applied, including comments on the performance of each method used.

IV. Conclusion and Future Work

Textual entailment is a remarkable field of natural language processing that is used in a variety of applications. Textual entailment applications include text summarization, information validation, question answering, sentiment analysis, etc. This paper spots on famous applications in Textual entailment. By studying previous methods and applications, this survey concludes that deep learning techniques can be used in entailment applications (such as rumor detection and depression detection) to facilitate and enhance the entailment process. It is also concluded that the textual entailment helps to extract data from social media applications and can deal with the huge amount of data more accurately with the aid of deep learning thus, the results are higher in accuracy. When dealing with massive data, more noise and detailed data are found. One text can contain several meanings, and different sentences can show the same concepts. This variation in semantic expression can be considered double trouble of language ambiguity. Textual entailment is the same but minimizes weakness the relationship to be one way. In addition, textual entailment can achieve text understanding in question answering applications. Despite using question answering applications in e-learning, it can be used in creating a clinical dataset by extracting information from regular documents. Also, it can be used to offhand extract patients' data and export it to health insurance applications.

In future, the GRU technology can be used along with textual entailment. Also, a pipeline containing more algorithms can be used. The aim is to automatically fine-tune the algorithms' parameters to achieve a higher percentage of accuracy and confidence, and discover which of those algorithms can fit the used dataset. Another future direction is to combine NLP and textual entailment with deep learning to outperform results in depression detection in social media. In conclusion, considering the promising results when using GRU in knowledge-aware applications, it is hoped to get higher accuracy when using textual entailment in depression detection.

References

[1] S.H. Jayasinghe, K. Sirts, Deep learning textual entailment system for sinhala language, Publisher, City, 2019.

[2] D.Z. Korman, E. Mack, J. Jett, A.H. Renear, Defining textual entailment, Publisher, City, 2018.

[3] P. Kapanipathi, V. Thost, S.S. Patel, S. Whitehead, I. Abdelaziz, A. Balakrishnan, M. Chang, K.P. Fadnis, R.C. Gunasekara, B. Makni, Infusing Knowledge into the Textual Entailment Task Using Graph Convolutional Networks, in: AAAI, pp. 8074-8081, 2020.

[4] Z. Zhang, Y. Wu, Z. Li, H. Zhao, Explicit contextual semantics for text comprehension, in: Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), 2019.

[5] L. Zhang, Neural Network Models for Text Understanding, in, 2019.

[6] A. Yavary, H. Sajedi, M.S. Abadeh, Information verification improvement by textual entailment methods, Publisher, City, 2019.

[7] M. Kaur, D. Srivastava, Text Summarization using Partial Textual Entailment based Graphs, in: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) , pp. 366-374, IEEE, 2019.

[8] D. Rudrapal, A. Das, B. Bhattacharya, Recognition of partial textual entailment for Indian social media text, Publisher, City, 2019.

[9] D. Rudrapal, A. Das, B.J.C.y.S. Bhattacharya, A New Approach for Twitter Event Summarization Based on Sentence Identification and Partial Textual Entailment, Publisher, City, 2019.

[10] S. Abujar, M. Hasan, S.A. Hossain, Sentence similarity estimation for text summarization using deep learning, in: Proceedings of the 2nd International Conference on Data Engineering and Communication Technology, Springer , pp. 155-164, 2019.

[11] D. Diefenbach, A. Both, K. Singh, P. Maret, Towards a question answering system over the semantic web, Publisher, City, 2020.

[12] T. Khot, A. Sabharwal, P. Clark, Scitail: A textual entailment dataset from science question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[13] D. Chen, Y. Li, M. Yang, H.-T. Zheng, Y. Shen, Knowledge-aware Textual Entailment with Graph Attention Network, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2145-2148, 2019.

[14] M. Ben-Sghaier, W. Bakari, M. Neji, Ar-SLoTE: A Recognizing Textual Entailment Tool for Arabic Question/Answering Systems, in: 2019 7th International conference on ICT & Accessibility (ICTA), IEEE, pp. 1-6, 2019.

[15] M. Ben-Sghaier, W. Bakari, M. Neji, Recognizing Textual Entailment for Arabic using semantic similarity and Word Sense Disambiguation, in: LPKM, 2018.

[16] N. Almarwani, M. Diab, Arabic textual entailment with word embeddings, in: Proceedings of the third arabic natural language processing workshop, pp. 185-190, 2017.

[17] A. Yavary, H. Sajedi, M.S. Abadeh, Information verification in social networks based on user feedback and news agencies, Publisher, City, 2020.

[18] T. Chakrabarty, T. Alhindi, S. Muresan, Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER) , pp. 127-131, 2018.

[19] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, Publisher, City, 2018.

[20] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading wikipedia to answer open-domain questions, Publisher, City, 2017.

[21] M.E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, Publisher, City, 2017.

[22] S. Gupta, S. Lakra, M. Kaur, Sentiment Analysis using Partial Textual Entailment, in: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE, pp. 51-55, 2019.

[23] F.T.J.a.p.a. AL-Khawaldeh, A study of the effect of resolving negation and sentiment analysis in recognizing text entailment for Arabic, Publisher, City, 2019.

[24] D. Majumdar, P. Bhattacharyya, Lexical based text entailment system for main task of RTE6, Publisher, City, 2010.

[25] S. Evensen, Y. Suhara, A. Halevy, V. Li, W.-C. Tan, S. Mumick, Happiness Entailment: Automating Suggestions for Well-Being, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, pp. 62-68, 2019.

[26] T. Boudaa, M. El Marouani, N. Enneya, Alignment based approach for Arabic textual entailment, Publisher, City, 2019.

[27] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, Publisher, City, 2018.

Improved Exemplar based Image Inpainting for Partial Instance Occlusion Handling with K-means Clustering and YCbCr Color Space

Deepa Abin[1], Sudeep D.Thepade[2]
Computer Department
PCCoE
Pune,India

***Abstract*—The images acquired in real time outdoor environment are often subject to uneven illumination conditions, cloudy weather, lighting conditions. The instances of partial occlusions deteriorate the background modeling of such scenes. Varying illumination outdoor scene set of images with partial occlusions are addressed through this investigative work. There is a need for a restoration method that finds improved subjective perception and execution time. The proposed work focuses on novel amended exemplar model to improve the subjective perception. The exemplar inpainting method is improved through the color quantization with K-means clustering approach in YCbCr color space. Experimental validations and proposed method results show better improvement in qualitative and objective measures than existing methods. The average “Peak Signal to Noise Ratio (PSNR)” as 28.2869 and “Structural SIMilarity index (SSIM)” as 0.9759 of the proposed method has shown better results respectively both visually and with a tradeoff in time.**

Keywords*—*Partial occlusion; exemplar inpainting; K-means clustering; YCbCr color space

I. Introduction

This paper presents an improved exemplar inpainting method for partial occlusions in illumination variant scenes. Seeing through the partial foreground occlusions is crucial in computer vision, such as detection and tracking in surveillance [1,2]. Scene understanding has been one of the significant foundations in machine perception. The real-world scenes comprise of multiple objects being occluded by other objects. The objects behind occlusions cannot be completely observed.

Occlusions are restored by regeneration of the gaps or by filling the regions in a visually pleasing way by using the image inpainting methods. Inpainting remains a progressive area in Image Processing domain. It is a kind of technology that uses the image information to remove objects or fill the missing region. After image inpainting, the image can be restored to the original and accord with the human visual sense. The subjective perception is of paramount significance in image inpainting. Substantial amount of work has been carried out in literature. However, extensive research yet needs to be carried out for further enhancing the qualitative results [1-5,8]. The Subjective and Objective Quality assessment of the inpainted images still remains an open problem.

In this work, an amended exemplar inpainting method is used for regeneration of partial occlusions through which the texture and color is preserved. An approach which can help restore the naturalness of a scene with less computation time is the primary objective of this work. The subjective perception is paramount in image restoration methods. This novel amended approach of K-means with exemplar based on the inpainting method has shown a tradeoff in subjective perception and execution time. In literature, quantization with diffusion based inpainting method [3,8,18] and PDE based method[3-4,14] has been attempted. The K-means approach has been used here for the color image quantization. The color image quantization however, is ancillary but still remains very important in image clustering. It is a process to reduce the color distortion quantitatively in the image [6,7,9]. There have been conventional image clustering algorithms in literature and out of these numerous methods, K-means clustering happens to be the simplest and widely used method. It typically focuses on minimization of sum of squared distances from cluster centres to all the points that can further relate to compact clusters. These resultant compact clusters have been further used in our improved inpainting process.

The non-uniform illumination in real scenes poses challenges towards the intrinsic inpainting performance. Visual perception is of supreme relevance in today’s era of digital imaging where numerous applications in computer vision involve illumination variant scenes. The research paper here mainly contributes to the regeneration of background during partial occlusion instances in non-uniform illumination scenes.

The three preliminary contributions of the proposed method have been as follows:

- Improved exemplar based on painting with Color Quantization using K-means clustering approach for non-uniform illumination variant scenes.
- Extension of the proposed method with YCbCr color space.
- Empirical value of ‘K’ in K-means color quantization.

The paper has been ordered as follows: The Section II mainly presents the related work and in Section III the proposed method is presented. In Section IV, the experimental results are thoroughly discussed and have been analysed in a qualitative and quantitative manner. The conclusion is summarized and presented in Section V.

II. Related Work

Here, we discuss the significant image inpainting methods, with major focus on exemplar based inpainting to provide a prerequisite for our approach.

There are different techniques of image inpainting [2-4,12-13,19-21,23,35] existing in literature. Mainly they can be classified into Texture Synthesis Inpainting, PDE Inpainting, Exemplar Inpainting, Hybrid Inpainting and Fast semi-automatic inpainting.

In 'Texture-Based inpainting methods', the missing regions are filled by sampling alongwith copying adjacent pixels [3,30]. There is a constraint in effectively handling edges and boundaries well. The natural scenes are not well addressed using this method, as it is formed of edge based structures.

'Partial Differential Equation (PDE)' is based on an iterative approach [3,14,29], produces good results when missed regions are smaller and time consumption is more when missed regions are larger.

The missing regions are filled in case of a hybrid approach using a combination of texture based synthesis and PDE method. Here, the division is mainly done into separate sections, structure based and texture based regions. The missing regions are filled up using edge propagating algorithms in combination with texture based synthesis techniques. It is computationally exhaustive when larger regions are considered.

The semi-automatic inpainting technique adapts a two-step process. Firstly, user interaction is used to specify missing information by using object boundaries followed by patch-based texture synthesis. The time consumption is relatively more in this method. However, in literature exemplar based inpainting methods has shown better image restorations. Hence, in order to deal with occlusion, the Exemplar based inpainting method is used for selected partial occlusion in the non-uniform illumination images to retain the texture and color. This exemplar based on the inpainting method, can be further improved in time and perception. The exemplar based technique is an efficient methodology to reconstruct larger target regions. The two basic steps comprise of priority assignment and selection of the patch that suits best. The patch that is best suited is selected for the target region and when obtained, the source value pixels are copied to the desired target region. The choice of varied patch matching criterias, size of patch and region filing have a great effect on results. For the missing regions wherein simple structure and texture is included, Exemplar based Inpainting method is observed to produce good results. The papers [16-18], discusses the need for improved exemplar based techniques. K-means clustering finds the the best trade-off between subjective evaluation and time. Due to its simplicity, the K-means algorithm [11], being the most popular amongst the clustering methods, has proved rewarding in enumerous applications. Though K-means is widely used as clustering algorithm, it has been actually considered as a quantization algorithm. The main goal being compression with probability distribution of K points. Higher quality images consume more storage as compared with lower quality images. Color quantization reduces the color quantity in an image where subjective error is minimal when quantized image and the original is considered. The clustering technique is used as a tool for color quantization and produces good quality results.

Out of the many techniques available for compression, K-means Clustering creates clusters comprising of major colors that further adds to computation efficiency.

As in [10-13], Exemplar based on inpainting has experiential better subjective results in comparison with other inpainting methods and this can be further improved with amendment of color quantization with compact clusters.

In the paper [6-8,16-18], the need for color quantization has been discussed to achieve lesser time consumption and the resultant images can be further used for better optimization in results. As in [7-9], there is a need to remove redundant pixels that can improve the quantization process in color space. The PDE method has been used with K-means for quantization [17]. The exemplar based on the inpainting method if further amended with color quantization, the observations can be recorded further[6-8]. In the further section, this amendment has been explored and the results have been quite impressive and thus paves way for further applications using the proposed method.

As in [30],the deep learning inpainting methods might produce meticulous results with usage of convolutional neural network for extraction of high level features. However, these high level features require high complex models with greater number of parameter adjustment, with higher accuracy especially. Recently the angle-aware patch based inpainting method[31], recovers missing regions with texture and structure components with image naturalness. However, the complexity involved in the process can be better simplified and can also be extended to the image compression domain.

In order to synthesize images, as discussed in [32,33], the patches are searched for, using the neighbour searching method. However, the major challenge still remains to maintain image naturalness in the output image, in terms of structure and texture. Though there exists conventional methods that uses image patches from the existing regions, or the diffusion based approach that propagates pixels from highly similarity regions into the whole regions. Though these methods propagate brilliant textures aimed at background inpainting, the high level semantics are however not captured perfectly and thus yields to non-realistic images. Though plausible outcomes are not achieved, but there is abundant scope for better texture regeneration. With recent advancement in generative neural network, inpainting algorithms can be trained to adapt to highly meaningful semantics, to further generate lucid structures for the missing regions, however at cost of high complexity.

Although significant progress has been made in literature, there is still work to be done to handle partial occlusions with minimum computation cost and achieve image naturalness. A systematic comparative evaluation must be made and thus determine the best combination of strategies. The main aim of our work, as discussed further is to improve the existing exemplar method with minimal execution time and retain image naturalness.

III. Proposed Method

The main task of proposed method is to handle partial occlusions using an amended exemplar approach. The performance of the partial occlusions handling relies basically on two primary aspects: the image naturalness and execution time. The contributions of the proposed method have been put forth in three stages:

1) Improved exemplar based inpainting with Color Quantization using K-means clustering approach for non uniform illumination variant outdoor scenes.

2) Extension of (i) with gradient and tensor based methods with YCbCr color space.

3) Empirical value of 'K' in K-means clustering for improved exemplar inpainting.

A. Improved Exemplar based on Inpainting with Color Quantization using a K-means Clustering Approach for Non Uniform Illumination Variant Outdoor Scenes

One of the major key points of the research problem was to handle Partial occlusion using an exemplar inpainting approach with reduced execution time. As per the literature, Color quantization is one of the significant image operations that reduces the amount of color variations in an image. The process merges lesser dominant colors in a particular image into relatively significant color. As a result, the visual error between the quantized image and the original, is substantially low.

The continuous color range can be represented using a finite subset through Quantization. Consider a function f : P → Cs representing an image with a range of colors in Cs. Here quantization would consist of a quantizer function qc : Cs → Q = {qc1, . . . , qcn}, in the color space definition Cs with a finite subset Q ⊂ Cs and qc1, . . . , qcn as quantization levels. The reduction of time and storage efficacy using color quantization could be explored further with an exemplar based inpainting. This paved way for a clustering based on color quantization amendment in an exemplar based inpainting approach and the results were appealing. This improved exemplar approach is novel with the results that could be achieved satisfying the conditions of image naturalness and improved execution time. The clustering based on color quantization method converts the quantization problem of RGB image into a problem of clustering pixels. K-means color quantization with exemplar inpainting method has not yet been explored in literature. The K-means algorithm being unsupervised clustering approach that automatically clusters based on the individual data points similarity is used for color quantization in our proposed method. The K-means algorithm is relatively simple and easier for implementation, alongwith higher efficiency and good clustering outcome. It has good adaptability to a new set of examples which could get experimented in proposed technique. The basic system proposed is putforth as in Fig. 1(a) and Fig. 1(b).

In Color quantization each of the color pixels are grouped into clusters, wherein each cluster represents a unique color in the new image. The K-means color quantization algorithm basically is basically categorized into three steps:

1) Initialization of 'K': The 'K' value depicts the quantitative value of total colors required for color quantization, to represent an image. The constraint however being that value of 'K' should be lesser than the concrete number of pixels in an image. The further classification of colors is carried out using the selected colors that act as the center colors.

2) Data point assignment to nearest centroid: Every data point needs to be assigned to its nearest centroid. Since every color is represented as RGB vector , the euclidean distance measure can be used to compute the distance between the two colors.

3) Color Remapping: Calculate the new centroid of every cluster and resdistribute every data point to the newest centroid. In color quantization, this step is the crux and in this particular step, every pixel in the original image is replaced with center color that has the closest proximity to it. This specific remapping creates an image with only 'K' colors, similar to the original image.

Repeat steps 2 and 3 until no point changes clusters, or until the centroids remain some.

These quantized cluster centres are further used and the RGB matrix with the cluster centres is further used for exemplar based inpainting.

B. Exemplar based on Inpainting on Region of Interest

Image inpainting is crucial for image restoration through the surrounding pixels information. The ideology in this method remains simple to replace the masked region with its neighboring pixels so that it looks like the neighborhood. The image reconstruction starts with identifying the damaged or missing part of an image and then search the image for similar known parts in the same image. These similar known parts of the image becomes the source region for region filling. The target region identified needs to be filled by copying the pixels from the source regions. The filling order remains crucial and determines the overall subjective quality of the image. The priority term of all pixels on boundary regions is estimated. The next step of investigation is to determine the border region pixel 'p' having the largest priority with respect to the patch centred around this point 'p'.

A Region Of Interest (ROI) primarily being a chosen subset of data points in an image defines a location in an image. The partial occlusion instances in an image can be selected dynamically and further exemplar based inpainting can be performed. Steps for gradient-based exemplar inpainting:

- Consider Region Of Interest in an image that has to be inpainted.

- It starts with boundary of the region and fills the region within.
- The neighbourhood region surrounding the pixel is selected to be painted.
- The normalized weighted sum of all known neighbourhood pixels is used to replace the identified pixel and continue for the selected patch region.

Here in the proposed system, the novel approach to improve the exemplar based inpainting method with a tradeoff in time and subjective perception has been put forth. The proposed method has shown comparatively better results as compared to conventional exemplar with gradient based method and exemplar with tensor based method. The experimentation results have been further put forth in Section IV.

C. Improved Exemplar based Inpainting with Color Quantization using K-means Clustering and YCbCr Color Space

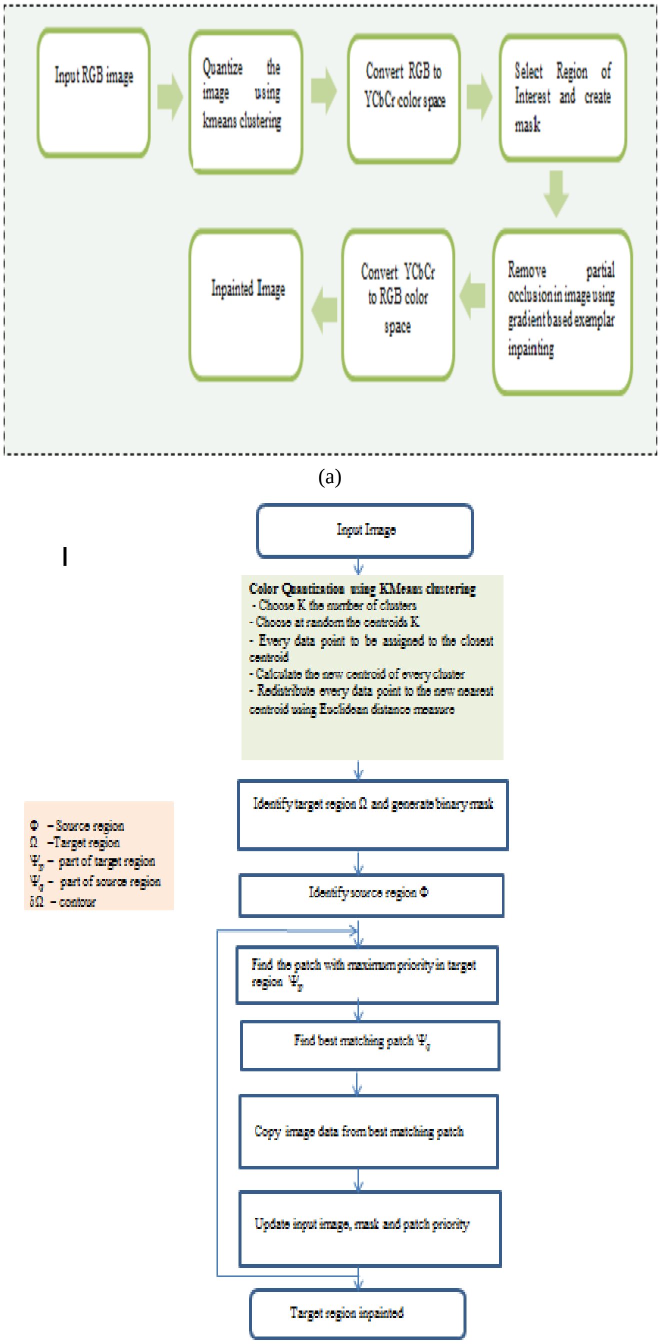
a) Color Space

The redundancies across three color channels of the source RGB image are better suppressed in YCbCr color space [34,35].

YCbCr represents the luminance (Y) chromaticity (Cb and Cr) color space, imitating a better human vision. In this color space, quantization operation shows better results due to the removal of redundancies in the R,G,B color channels. Adversarial perturbations are allocated more in the Y-channel of the YCbCr color space than in the Cb and Cr channels, whereas in the RGB color space, the perturbations are equally distributed between the three channels. Therefore, the natural step is to remove the adversarial perturbation from the Y-channel instead of targeting all channels in the RGB color space. Here in the proposed approach, the quantized outdoor scene set of images in RGB color space is transformed to YCbCr color space and then the exemplar inpainting is further applied to handle the partial occlusion instances as putforth in Fig. 1(a). After performing exemplar inpainting,the resultant image is transformed back to the RGB color space. Efficient representation of images is finally observed in this color space.

b) Empirical Value of 'K' for K-means Clustering for Improved Exemplar Inpainting

The K-means clustering method uses the Eucledian distance measure and for the K classes of the data set, through initial random sampling approach, the initial centroid is calculated. Once the initial cluster is formed, thereafter divide each sample point into the clusters such that it is closest to the nearest center point. With this iterative approach, the center point of all sample points then becomes the center point of the cluster. These steps to be repeated until the center point remains unchanged or reaches the desired set of iterations. The Fig. 1(b) illustrates the flowchart of the proposed system where all the steps have been summarized.


Input RGB image
Quantize the image using kmeans clustering
Convert RGB to YCbCr color space
Select Region of Interest and create mask
Remove partial occlusion in image using gradient based exemplar inpainting
Convert YCbCr to RGB color space
Inpainted Image
Input Image
Color Quantization using KMeans clustering
- Choose K the number of clusters
- Choose at random the centroids K
- Every data point to be assigned to the closest centroid
- Calculate the new centroid of every cluster
- Redistribute every data point to the new nearest centroid using Euclidean distance measure
Identify target region Ω and generate binary mask
Identify source region Φ
Find the patch with maximum priority in target region Ψ_p
Find best matching patch Ψ_q
Copy image data from best matching patch
Update input image, mask and patch priority
Target region inpainted
Φ – Source region
Ω –Target region
Ψ_p – part of target region
Ψ_q – part of source region
δΩ – contour

(a)

(b)

Fig. 1. (a). Block Diagram for Proposed Improved Exemplar based Inpainting and (b). Flowchart for Proposed Improved Exemplar based Inpainting with Color Quantization using K-means Clustering Approach for Non Uniform Illumination Variant Scenes with YCbCr Color Space.

IV. Results and Discussion

In the section here, the testbed is introduced and then the proposed method is compared with state of the art methods are presented with interpretation and analysis. The sample outdoor scene images were taken form MOT[25] and VV[26] dataset. The proposed method was applied on the selective images form these datasets where partial occlusion evidences could be obtained. A total of twelve outdoor images with non uniform illumination were selected for experimentation. The experimentation analysis is carried out in MATLAB 2020 environment under 64-bit Windows 8.1 operating system, with Intel® Core (TM) i3-4030U and CPU (1.90 GHz) with 4 GB RAM.

The Fig. 2 here illustrates the sample images considered for the experimentation. In this work, the color quantization was performed on each of the images from the testbed for different values of 'K' such as 20, 30, 40, 50 and 60 to find optimal value of 'K' for effective color quantization amendment with exemplar inpainting method and the objective quality metrics with CPU time was recorded.

The efficacy of objective evaluation of proposed method is done using the PSNR[24] and SSIM[24]. As evident from the literature stated [24,36-38], higher PSNR signifies better subjective evaluation of inpainting results. This can be correlated with the subjective perception of image naturalness. Generally, larger the PSNR, lesser the diversity between the inpainted image and the original. Additionally, SSIM measures the similarity between the two images. It has proved to be a sufficient metric for inpainting results. In exemplar based inpainting method, it becomes difficult to recover the object that needs removal completely using any mathematical method. It can hold maximum resemblance to the original and hence the PSNR and SSIM values are crucial objective evaluation parameters [36-38]. Hence, PSNR and SSIM values are relevant to evaluate our work of partial occlusion handling inpainting results. In further subsections, the visual efficacy of proposed method is discussed with execution time and feasibility.

Here, the quantized color image with K-means clustering is used further to determine the partial occlusion instances with ROI method. The mask required for the exemplar inpainting method is generated using this selected ROI. The exemplar inpainting method with structure tensor [4,5] comprises of the intensity information of the neighbourhood region as well as the gradient's transformation directions and degree of coherence of the same. The priority function includes this structure tensor. The linear structure is thus preserved using this priority function along the geometric structures in an image. In large regions, object removal becomes more effective when using structure tensor and also ensures accurate propagation [27,28].

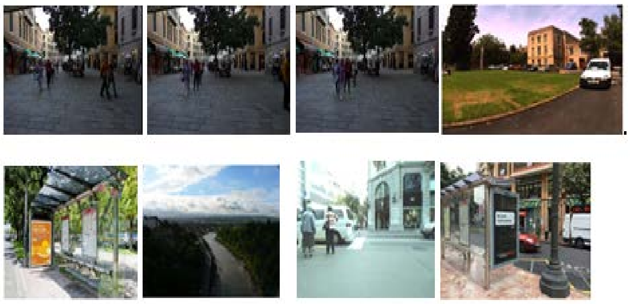


Fig. 2. Sample Outdoor Images with Non Uniform Illumination and Partial Occlusion for Experimentation [23, 24].

The objective evaluation has been put forth as seen in Table I. Here, a clustering approach with 50 region clusters has been considered. The objective and subjective evaluation results with the proposed improved exemplar inpainting method with color quantization using K-means clustering approach is discussed. Here as seen in the Table I, the gradual trend in the proposed method with gradient-based exemplar has shown better average PSNR of 27.8275 and SSIM of 0.9419. The graphical visualization of average PSNR and average SSIM is further putforth in Fig. 3 and Fig. 4.

TABLE I. COMPARATIVE ANALYSIS OF OBJECTIVE EVALUATION OF PROPOSED METHOD

Images	**Exemplar with gradient based method [15]**		**Exemplar with tensor based method [28]**		**Proposed K-means color quantization with gradient based exemplar inpainting**		**Proposed K-means color quantization with tensor based exemplar inpainting**	
	PSNR	***SSIM***	***PSNR***	***SSIM***	***PSNR***	***SSIM***	***PSNR***	***SSIM***
a	28.8594	0.9805	28.7718	0.973	30.0989	0.9403	30.0228	0.9363
b	30.2172	0.9844	29.2833	0.9783	31.3876	0.9417	29.83	0.9383
c	29.9121	0.9563	28.1554	0.9372	29.0497	0.9198	28.9	0.9171
d	29.5671	0.9834	27.5867	0.9812	29.9817	0.9662	29.0978	0.9636
e	30.9792	0.9895	29.4162	0.9813	31.0264	0.905	30.9832	0.9845
f	20.3264	0.9357	18.908	0.8638	21.3028	0.8725	20.4466	0.92093
g	25.9154	0.953	26.5642	0.9585	26.8061	0.9758	25.9662	0.9643
h	21.3499	0.9452	21.2661	0.9416	21.6572	0.9257	21.5304	0.9266
i	28.8805	0.9814	28.1732	0.9799	28.0979	0.9666	28.9001	0.9971
j	30.63	0.9833	30.63	0.9686	31.3565	0.9694	30.4123	0.989
k	28.239	0.9766	26.4779	0.9546	28.9829	0.9565	28.9392	0.9765
l	23.2706	0.9733	22.7972	0.9611	24.1827	0.9636	24.9772	0.9836
Average	27.3456	0.9703	26.5025	0.9566	**27.8275**	**0.9419**	27.5004	0.9582

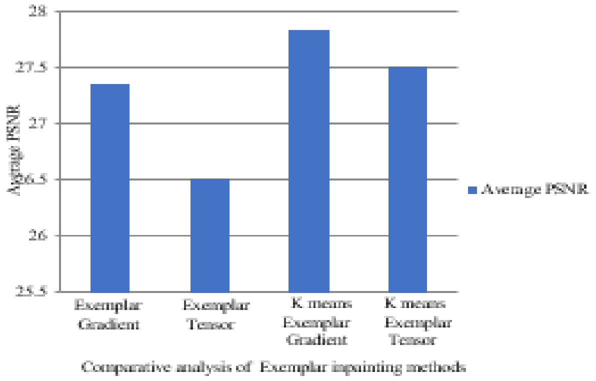


Fig. 3. Comparative Analysis of Average PSNR of Proposed Method with Existing Exemplar Inpainting Methods.

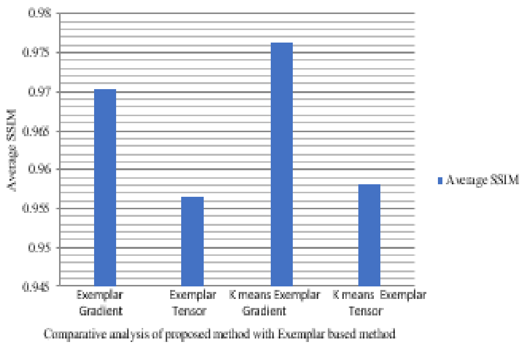
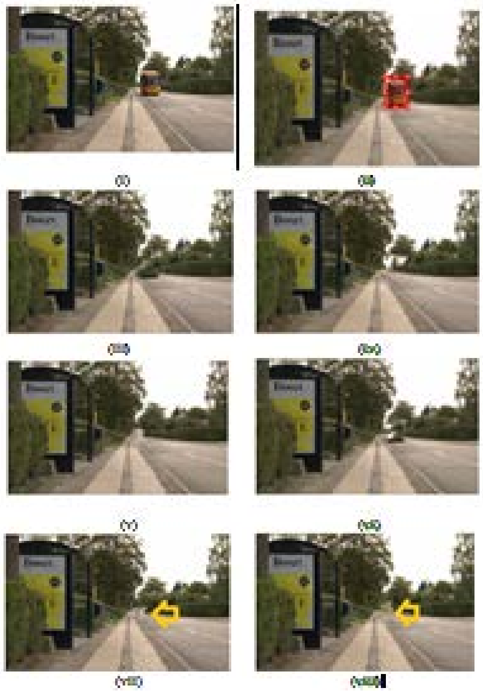


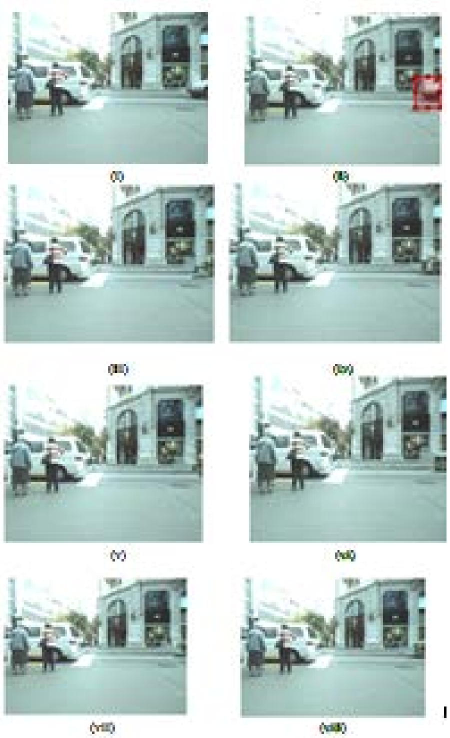
Fig. 4. Comparative Analysis of Average SSIM of Proposed Method with Exemplar Inpainting Methods.

The proposed method when extended to YCbCr color space [22, 34] also demonstrated comparatively better results than the RGB based variant of proposed method. Here as seen in Table II again, the gradient based exemplar has shown better average PSNR of 28.2869 and SSIM of 0.9759 as compared to gradient based exemplar method, tensor based exemplar method and proposed K-means color quantization with tensor based exemplar inpainting method. Here Kmax was fixed to 50 empirically as it was quite enough from a subjective point of view where more number of colors may pose difficulties in color identification itself.

Here as seen in Fig. 5(a) and 5(b) the subjective evaluation results for partial occlusion on selective frames have been put forth. As observed, the proposed method with K-means color quantization with gradient based exemplar method has shown better visual perception than the other considered existing methods.



(a)



(b)

Fig. 5. (a). (i) Original Image (ii) Color Quantized Image with K-means Clustering and ROI (iii) Exemplar Gradient based Inpainting Method (iv) Exemplar Tensor based Inpainting Method (v) Proposed K-means Color Quantization with Gradient based Exemplar Inpainting Method in RGB Color Space (vi) Proposed K-means Color Quantization with Tensor based Exemplar Inpainting Method in RGB Color Space (vii) Proposed K-means Color Quantization with Tensor based Exemplar Inpainting Method in YCbCr Color Space. And (b). (i) Original Image (ii) Color Quantized Image with K-means Clustering and ROI (iii) Exemplar Gradient based Inpainting Method (iv) Exemplar Tensor based Inpainting Method (v) Proposed K-means Color Quantization with Gradient based Exemplar Inpainting Method in RGB Color Space (vi) Proposed K-means Color Quantization with Tensor based Exemplar Inpainting Method in RGB Color Space (vii) Proposed K-means Color Quantization with Gradient based Exemplar Inpainting Method in YCbCr Color Space (viii) Proposed K-means Color Quantization with Tensor Based Exemplar Inpainting Method in Ycbcr Color Space.

As putforth in Table III, the method proposed has demonstrated better computation time in seconds as compared with all the other methods. The average time of proposed method with K-means amendment in YCbCr color space has been an average of 0.7580 sec.

Here as the proposed method has performed comparatively with gradient-based exemplar method, further experimentation was conducted to find an empirical value of ‘K’ [17] for the K-means clustering approach. As can be observed, the image naturalness and visual quality of the inpainted image is better preserved using our method. The results were varied for different values of ‘K’ from 20 to 60 and as observed experientially the optimal value of ‘K’ has been 50. The average PSNR and SSIM for K=50 has shown better results as compared to varied values of ‘K’ and has been tabulated as in Table IV.

TABLE II. OBJECTIVE EVALUATION OF GRADIENT BASED AND TENSOR BASED EXEMPLAR INPAINTING METHOD WITH PROPOSED K-MEANS COLOR QUANTIZATION WITH GRADIENT BASED EXEMPLAR INPAINTING AND TENSOR BASED EXEMPLAR INPAINTING METHOD IN YCBCR COLOR SPACE

	Exemplar with gradient based method [15]		**Exemplar with tensor based method [28]**		**Proposed K-means color quantization with gradient based exemplar inpainting**		**Proposed K-means color quantization with tensor based exemplar inpainting**	
Images	***PSNR***	***SSIM***	***PSNR***	***SSIM***	***PSNR***	***SSIM***	***PSNR***	***SSIM***
a	12.1778	0.2019	12.1339	0.2011	30.3264	0.9799	30.2436	0.9352
b	12.2206	0.2009	12.1722	0.2000	31.6080	0.9856	31.2089	0.9397
c	12.2055	0.2104	12.2932	0.2107	30.7342	0.9595	29.2634	0.9182
d	11.7090	0.2080	11.7104	0.2064	30.1510	0.9782	29.3028	0.9622
e	10.8579	0.2509	12.9048	0.1931	31.1825	0.9868	30.3999	0.9046
f	12.9608	0.1918	12.6109	0.2509	21.3047	0.9726	19.5922	0.8642
g	12.4174	0.2872	10.7362	0.1486	27.8264	0.9653	26.9662	0.9843
h	10.7800	0.1459	10.6843	0.1438	21.8751	0.9459	21.5074	0.9287
i	11.5253	0.1097	11.5220	0.1104	27.9440	0.9851	26.7446	0.9608
j	10.8929	0.2403	10.7354	0.2065	33.6176	0.9892	31.2825	0.9689
k	10.8608	0.2482	10.7586	0.2094	28.4694	0.9786	26.2098	0.9531
l	10.9080	0.2427	10.6483	0.2103	24.4039	0.9844	22.9983	0.9557
Average	11.6188	0.2114	11.5758	0.1909	**28.2869**	**0.9759**	27.1433	0.9396

TABLE III. COMPARATIVE ANALYSIS WITH TIME AS PERFORMANCE MEASURE FOR GRADIENT BASED AND TENSOR BASED EXEMPLAR INPAINTING METHOD WITH PROPOSED K-MEANS COLOR QUANTIZATION IN YCBCR COLOR SPACE

Images	**Exemplar with gradient based method [15]**	**Exemplar with tensor based method [28]**	**Proposed K-means color quantization with gradient based exemplar inpainting**	**Proposed K-means color quantization with tensor based exemplar inpainting**
a	0.3049	0.3092	0.1993	0.4541
b	0.30174	0.2967	0.3172	0.542
c	1.1590	2.1725	1.1159	1.978
d	0.6649	0.2509	0.1351	0.2812
e	9.2533	8.9804	5.5015	8.88
f	1.3130	1.6597	1.1435	2.5031
g	0.3621	0.4079	0.516	0.327
h	0.0884	0.0573	0.0657	0.0866
i	0.0555	0.0465	0.0186	0.0612
j	0.6647	0.0159	0.0207	0.0364
k	0.0305	0.0537	0.042	0.0536
l	0.0282	0.0462	0.0204	0.0405
Average	1.1855	1.1914	**0.7580**	1.2703

As observed through the experimentations carried out, the proposed approach with YCbCr color space has shown substantial visual perception and improved execution time that are crucial in inpainting methods. The relative comparison with original images for partial occlusion instances is better handled using the proposed approach. The proposed approach, however, has not shown good results but for the lesser value of ‘K’ below 20 where distortions could be evident. This method can be further applied for image restoration applications where the tradeoff between time and visual perception is crucial.

In this paper, after comprehensive discussion along with subjective and objective evaluation performed on testbed, the improved exemplar based technique with color quantization using K-Means approach has shown comparatively better performance than the existing variation of the exemplar method. The average PSNR for the proposed method has been 28.2869 and SSIM as 0.9759 with an average time for the sample non-uniform illumination variant scenes as 0.7580 sec. An experimental analysis of the obtained results yielded the following conclusions:

- K-means color quantization amended with exemplar based inpainting in YCbCr color space has demonstrated qualitative and quantitative efficacy.
- Improvement in image naturalness and execution time was better achieved.

The experiments performed could demonstrate plausible results for subjective perception. The tradeoff achieved here is of paramount significance. The results achieved here can offer a context for added research that may be undertaken further to improve upon, the presented methods here. In future, the different quantization approaches can be further amended with an exemplar based inpainting method and the results can be observed. Automated restoration methods for complex and overlapping structure with curves and larger curvatures can be further addressed.

TABLE IV. EXPERIENTIAL RESULTS OF PSNR AND SSIM FOR EMPIRICAL VALUE OF 'K' FOR PROPOSED K-MEANS COLOR QUANTIZATION

Images	**PSNR**					**SSIM**				
	K=20	***K=30***	***K=40***	***K=50***	***K=60***	***K=20***	***K=30***	***K=40***	***K=50***	***K=60***
a	28.894	27.7337	28.8998	30.3264	28.9952	0.8875	0.9122	0.9273	0.9799	0.9399
b	29.1588	30.296	30.3361	31.6080	30.1142	0.894	0.9211	0.9313	0.9856	0.942
c	28.1037	28.9613	28.4288	30.7342	28.7446	0.8735	0.8986	0.9112	0.9595	0.9167
d	28.6113	28.1891	30.0093	30.1510	29.2277	0.9429	0.9542	0.9613	0.9782	0.9641
e	27.1098	29.1605	30.5192	31.1825	29.5521	0.8275	0.8696	0.891	0.9868	0.9085
f	20.4527	21.5886	20.0316	21.3047	20.1994	0.8178	0.8573	0.8543	0.9726	0.8842
g	24.3368	26.1326	27.5005	27.8264	24.3205	0.8918	0.9218	0.929	0.9653	0.9231
h	20.9770	21.6913	20.8654	21.8751	22.2709	0.9124	0.9188	0.9234	0.9459	0.9214
i	24.9995	27.2451	26.8494	27.9440	27.7503	0.9417	0.9531	0.9613	0.9851	0.9667
j	29.1623	30.4572	31.1235	33.6176	32.4004	0.9494	0.9603	0.967	0.9892	0.9753
k	25.7069	23.0402	22.8276	28.4694	26.2181	0.9305	0.93	0.9309	0.9786	0.9515
l	23.526	23.9969	24.3097	24.4039	24.6357	0.942	0.9525	0.9606	0.9844	0.9637
Average	26.5062	26.9556	27.3986	**28.2869**	27.0358	0.9009	0.9208	0.9291	**0.9759**	0.9381

V. CONCLUSION

In this paper, after comprehensive discussion along with subjective and objective evaluation performed on testbed, the improved exemplar based technique with color quantization using K-Means approach has shown comparatively better performance than the existing variation of the exemplar method. The average PSNR for the proposed method has been 28.2869 and SSIM as 0.9759 with an average time for the sample non-uniform illumination variant scenes as 0.7580 sec. An experimental analysis of the obtained results yielded the following conclusions:

- K-means color quantization amended with exemplar based inpainting in YCbCr color space has demonstrated qualitative and quantitative efficacy.
- Improvement in image naturalness and execution time was better achieved.

The experiments performed could demonstrate plausible results for subjective perception. The tradeoff achieved here is of paramount significance. The results achieved here can offer a context for added research that may be undertaken further to improve upon, the presented methods here. In future, the different quantization approaches can be further amended with an exemplar based inpainting method and the results can be observed. Automated restoration methods for complex and overlapping structure with curves and larger curvatures can be further addressed.

REFERENCES

[1] Bora, D. J. (2017). Importance of image enhancement techniques in color image segmentation: A comprehensive and comparative study. arXiv preprint arXiv:1708.05081.

[2] Bhangale, M. S., & Thorat, A. P. P. (2016). Image Inpainting Using Modified Exemplar-Based Method.

[3] Bertalmio, M., Vese, L., Sapiro, G., & Osher, S. (2003). Simultaneous structure and texture image inpainting. IEEE transactions on image processing, 12(8), 882-889. https://doi.org/10.1109/TIP.2003.815261

[4] Awati, M. A. S., & Patil, M. M. R. Review of Exemplar Based Image Inpainting using Structure Tensor. IJEEE, Vol. No.6, Issue No. 02, July-Dec., 2014. ISSN- 2321-2055 (E)

[5] Siadati, S. Z., Yaghmaee, F., & Mahdavi, P. (2016, May). A new exemplar-based image inpainting algorithm using image structure tensors. In 2016 24th Iranian Conference on Electrical Engineering (ICEE) (pp. 995-1001). IEEE. https://doi.org/10.1109/IranianCEE.2016.7585666

[6] Kaur, G., Singh, D., & Kaur, G. (2013). 'RGB' Color Image Quantization using Pollination based Optimization. International Journal of Computer Applications, 78(9). International Journal of Computer Applications (0975 – 8887) Volume 78 – No.9, September 2013

[7] Mota, C., Gomes, J., & Cavalcante, M. I. (2001). Optimal image quantization, perception and the median cut algorithm. Anais da Academia Brasileira de Ciências, 73(3), 303-317. https://doi.org/10.1590/S0001-37652001000300001

[8] Vreja, R., & Brad, R. (2014). Image inpainting methods evaluation and improvement. The Scientific World Journal, 2014. https://doi.org/10.1155/2014/937845

[9] Braquelaire, J. P., & Brun, L. (1997). Comparison and optimization of methods of color image quantization. IEEE Transactions on image processing, 6(7), 1048-1052. https://doi.org/ 10.1109/83.597280

[10] Cheng, Y., Liu, W., & Xing, W. (2019). A Novel Algorithm for Exemplar-based Image Inpainting (S). In SEKE (pp. 630-777). https://doi.org/10.18293/SEKE2019-152

[11] Naouel, O., & Kholladi, M. K. (2013). An Image Inpainting Algorithm based on K-means Algorithm.

[12] Newson, A., Almansa, A., Gousseau, Y., & Pérez, P. (2017). Non-local patch-based image inpainting. Image Processing On Line, 7, 373-385. https://doi.org/10.5201/ipol.2017.189

[13] Chhabra, J. K., & Birchha, M. V. (2014). Detailed survey on exemplar based image inpainting techniques. International Journal of Computer Science and Information Technologies, 5(5), 6350-635.

[14] Hoeltgen, L., Peter, P., & Breuß, M. (2018). Clustering-based quantisation for PDE-based image compression. Signal, Image and Video Processing, 12(3), 411-419.

[15] Criminisi, A., Perez, P., & Toyama, K. (2003, June). Object removal by exemplar-based inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. (Vol. 2, pp. II-II). IEEE.

[16] Blas, M. R., Agrawal, M., Sundaresan, A., & Konolige, K. (2008, September). Fast color/texture segmentation for outdoor robots. In 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 4078-4085). IEEE. https://doi.org/10.1109/IROS.2008.4651086

[17] Ray, S., & Turi, R. H. (1999, December). Determination of number of clusters in k-means clustering and application in colour image segmentation. In Proceedings of the 4th international conference on advances in pattern recognition and digital techniques (pp. 137-143).

[18] Nirali Pandya, Bhailal Limbasiya, "A Survey on Image Inpainting Techniques," International Journal of Current Engineering and Technology,vol.3,no.5, Dec.,pp.1828-1831, 2013.

[19] A.Criminisi,P.Perez, K.Toyama, "Region Filling and Object Removal by Exemplar based Inpainting,"IEEE Transactions on Image Processing,vol.9,pp.1-31,2004.

[20] Drori, I. (2007). Fast minimization by iterative thresholding for multidimensional NMR spectroscopy. EURASIP Journal on Advances in Signal Processing, 2007, 1-10.

[21] Oliveira, A., Fickel, G., Walter, M., & Jung, C. (2015, April). Selective hole-filling for depth-image based rendering. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1186-1190). IEEE.

[22] Kekre, H. B., Thepade, S., Das, R. K. K., & Ghosh, S. (2012). Image classification using block truncation coding with assorted color spaces. International Journal of Computer Applications, 44(6), 9-14.

[23] Pritika Patel,Ankit Prajapati,Shaliendra Mishra, "Review of Different Inpainting Algorithms," International Journal of Computer Applications,vol.59,no.18,Dec.,pp.866 –870, 2012.

[24] Hore, A., & Ziou, D. (2010, August). Image quality metrics: PSNR vs. SSIM. In 2010 20th international conference on pattern recognition (pp. 2366-2369). IEEE.

[25] https://motchallenge.net/data/MOT17

[26] Ying, Z., Li, G., & Gao, W. (2017). A bio-inspired multi-exposure fusion framework for low-light image enhancement. arXiv preprint arXiv:1711.00591.

[27] Baghaie, A., & Yu, Z. (2015). Structure tensor based image interpolation method. AEU-international Journal of Electronics and Communications, 69(2), 515-522. https://doi.org/10.1016/j.aeue.2014.10.022

[28] Liu, K., Qing Tan, J., & Yue Su, B. (2013, August). Exemplar-based image inpainting using structure tensor. In 2013 International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013) (pp. 619-623). Atlantis Press.

[29] Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000, July). Image inpainting. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques (pp. 417-424). https://doi.org/10.1145/344779.344972

[30] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G.-S. Hsu, and M. H. Yap, 'A comprehensive review of past and presentimage inpainting methods', Comput. Vis. Image Underst., vol. 203, p. 103147, Feb. 2021. https://doi.org/10.1016/j.cviu.2020.103147.

[31] N. Zhang, H. Ji, L. Liu, and G. Wang, 'Exemplar-based image inpainting using angle-aware patch matching', EURASIPJ. Image Video Process., vol. 2019, no. 1, p. 70, Jul. 2019. https://doi.org/10.1186/s13640-019-0471-2.

[32] A. Akl, C. Yaacoub, M. Donias, J.-P. Da Costa, and C. Germain, 'A survey of exemplar-based texture synthesis methods', Comput. Vis. Image Underst., vol. 172, pp. 12–24, Jul. 2018. https://doi.org/10.1016/j.cviu.2018.04.001.

[33] R. B. M. Salman and C. N. Paunwala, 'Semi automatic image inpainting using partial JSEG segmentation, in 2017 International Conference on Inventive Systems and Control (ICISC), pp. 1–6, Jan. 2017. https://doi.org/ 10.1109/ICISC.2017.8068728.

[34] Siadati, S. Z., Yaghmaee, F., & Mahdavi, P. (2016, May). A new exemplar-based image inpainting algorithm using image structure tensors. In 2016 24th Iranian Conference on Electrical Engineering (ICEE) (pp. 995-1001). IEEE.

[35] Mahdavi, P., Yaghmaee, F., & Alilou, V. K. Digital Image Inpainting Using Bilateral Filtering.

[36] Sangeetha, K., Sengottuvelan, P., & Balamurugan, E. (2011). A comparative analysis of exemplar based image inpainting algorithms. European Journal of Scientific Research, 60(3), 316-325.

[37] Salman, R. B. M., Goswami, P., & Paunwala, C. N. (2014, December). Comparative analysis of exemplar based image inpainting techniques. In 2014 2nd International Conference on Emerging Technology Trends in Electronics, Communication and Networking (pp. 1-6). IEEE.

[38] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. (2019). Image inpainting: A review. Neural Processing Letters, 1-22.

Predicting the Appropriate Mode of Childbirth using Machine Learning Algorithm

Md. Kowsher[1]
Department of Applied Mathematics
Noakhali Science & Technology University
Noakhali 3814, Bangladesh

Anik Tahabilder[3]
School of Engineering + Technology
Western Carolina University
Cullowhee, NC 28723, USA

Nusrat Jahan Prottasha[2]
Md. Abdur-Rakib[5], Md. Shameem Alam[6]
Department of CSE
Daffodil International University
Dhaka 1207, Bangladesh

Kaiser Habib[4]
Department of RET
University of Dhaka
Dhaka 1206
Bangladesh

***Abstract*—A woman's satisfaction with childbirth may have immediate and long-term effects on her health as well as on the relationship with her newborn child. The mode of baby delivery is genuinely vital to a delivery patient and her infant child. It might be a crucial factor for ensuring the safety of both the mother and the child. During the baby delivery, decision-making within a short time becomes very challenging for the physician. Besides, humans may make wrong decisions selecting the appropriate delivery mode of childbirth. A wrong decision increases the mother's life risk and can also be harmful to the newborn baby's health. Computer-aided decision-making can be an excellent solution to this problem. Considering this scope, we have built a supervised machine learning-based decision-making model to predict the most suitable childbirth mode that will reduce this risk. This work has applied 32 supervised classifier algorithms and 11 training methods on the real childbirth dataset from the Tarail Upazilla Health complex, Kishorganj, Bangladesh. We have also analyzed the result and compared them using various statistical parameters to determine the best-performed model. The quadratic discriminant analysis has shown the highest accuracy of 0.979992 with the F1 score of 0.979962. Using this model to decide the appropriate labor mode may significantly reduce maternal and infant health risks.**

Keywords—Childbirth; labour mode; supervised machine learning; maternal death; infant

I. INTRODUCTION

In baby delivery, we want to make sure that the mother and the child are safe. For this safety, the method of baby delivery is very significant. Usually, the corresponding physician chooses the mode of delivery from two options, includes (i) vaginal birth or (ii) Cesarean area (c-section) birth. So, the patient herself cannot contribute to the decision-making procedure. When a child contains a low-risk pregnancy and is in the head-down position, and the patient is at least 37 weeks pregnant, gynecologists suggest attempting a vaginal birth. In this case, a newborn baby usually gets essential gut bacteria from the mother. Besides, it can help press liquid out of a baby's lungs, decreasing the risk of the baby's breathing problem. This way of birthing also helps for breastfeeding and reduces the baby's risk of asthma and obesity. Additionally, parents will be able to avoid the cost and potential risk of surgery. That is why normal birth is most suitable for both the baby's and the mother's health.

On the other hand, there are cases like twins, or the mother has diabetes, high blood weight, HIV, or active herpes, or the baby is not in a head-down position, which complies the patient to have a c-section delivery. However, it increases the risk of asthma from the early childhood of the baby. In other instances, including delivering a comparatively larger baby for the maternal pelvis, or if the baby is not in a head-down position, the c-section delivery becomes an essential mode of childbirth.

However, many times, physicians are more biased for a c-section delivery than a vaginal delivery. The number count of c-sections is increasing day by day, and it got doubled during 1980. Another record says, as of 2015, the cesarean section rate is exceeding by 35.5% than WHO's recommendation [1]. Lately, during the period 2017-2018, the c-section rate in Bangladesh has increased by 51%, and in 2018, 77% of those c-sections were unnecessary [2]. Save the Children, a popular magazine, has recently documented a 51% increment in extraneous c-section delivery in Bangladesh [3]. In addition to that, maternal mortality is also shown and has been a big problem for most South Asian countries. Compared to the developed countries, the maternal mortality rate in Bangladesh is extremely high. In 2017, the maternal mortality rate in the USA was 0.000017 percent [4]. In the same year, this rate was 0.000113 percent in Bangladesh, which is a few times more than in other developed countries. WHO has reported maternal mortality of 194 per 1000 in Bangladesh. This high maternal death rate can be significantly reduced by selecting the birth mode appropriately. "National Low Birth Weight Survey Bangladesh, 2015" has reported that Maternal mortality and cesarean delivery rates have doubled compared to regular deliveries [5]. It gives us the scope to develop a decision-making model to choose the appropriate mode of childbirth.

Both of the processes of childbirth have advantages and disadvantages based on the particular patient's situation. However, the gynecologist decides the birth mode considering the mother's biological factors, including counting, age, ANC, para, partograph, AMTSL, blood circulation, birth weight, BP, PNC-1 presentation, cervix(OS), membrane, and so on. This research proposed a scientific method to decide childbirth mode considering the mother's present situation and earlier records.

The following points denote the main contributions of this research paper:

- We have proposed a computerized method of decision-making for selecting the appropriate mode of childbirth.
- Since this process is computerized and machine learning-based, it will be less error-prone.
- We have used 32 different classifier algorithms to make the decision more accurate and reliable.
- This model can analyze and use such big data for decision-making that it is merely impossible for a human being to analyze.

The rest of the sections are organized as follows: Section II describes the related work. Section III describes the methods and materials used, Section IV depicts the experimental procedure and the model. Section V examines and evaluates the results, and Section VI describes the conclusion and proposes the future direction of this research.

II. Related Work

A lot of research is being done in the machine learning domain for biomedical decision-making. Mboya IB et al. have proposed a machine learning-based method that can predict perinatal death using supervised machine learning algorithms [6]. ML-based model is also being used for predicting a lot of factors of childbirth. For example, Abraham, Abin, et al. described a new technique for gathering various information from EHRs in order to predict singleton preterm birth by applying various machine learning models [7]. Recently, Islam, Muhammad Nazrul, et al. has presented research regarding childbirth mode with two-fold findings: first, the potential highlights for deciding the method of labor, and second, machine learning algorithms for anticipating the suitable way of labor (vaginal birth, crisis cesarean, cesarean birth) [8]. Kowsher, M. et al. has reported good accuracy in applying machine learning-based recommendation system to predict the most appropriate childbirth mode [9]. Also, Khan, Nafiz Imtiaz, et al. have a similar test to anticipate whether the cesarean area is essential with the assistance of information mining and subsequently expand the mother and infant's security during and after labor by staying away from a pointless cesarean segment [10]. Besides, Fu, Yuanqing, et al. had described a model to recognize early life hazard factors for youth overweight/stoutness among preterm babies and decided to take care of practices that could alter the distinguished danger factors [11]. Other researchers also have applied machine learning models to classify various biomedical factors and hence to conclude the adverse effect of c-section delivery. For instance, Siddiqui, Mohammad Khubeb, et al. described that machine learning classifiers could use EEG information and identify seizures alongside uncovering applicable reasonable examples without trading off execution [12]. In addition, Soh, Yan Xi, et al. had explained the relationships among sociodemographic and medicine factors, the concern of parturition, psychosocial wellbeing, and childbirth self-efficacy employing a structural equation modeling approach [3]. c-section delivery may have a postbirth adverse health effect on a mother. Chen, Yanfang, et al. showed the relationship between conveyance and post-traumatic stress problems that yielded conflicting outcomes. This examination is expected to research the relationship between conveyance and post-traumatic stress in an associate of Chinese ladies with a high pace of cesarean conveyance [14].

Zhang, Yiye, et al. propose a machine learning structure for PPD hazard expectation utilizing information extricated from electronic wellbeing records (EHRs) [15]. Later on, Lipschuetz, Michal, et al. presented to decide the customized forecast of vaginal birth after cesarean conveyance utilizing 30 an AI calculation that might help patient-doctor dynamic and 31 increment paces of preliminaries of work [16]. Also, Serçekuş, Pınar, Okan Vardar, and Sevgi Özkan proposed to recognize and think about the dread of labor and related to variables among pregnant ladies and their accomplices [17]. Onchonga, David et al. described a new investigating ladies' experience from maternity specialists drove incorporated pre-birth preparing and its effect on the dread of labor [18]. After that, Liu, Ligue, et al. proposed an expectation model of undeveloped improvement by using machine learning algorithms dependent on authentic case information. In this way, specialists can make more exact ideas on the quantity of patient subsequent meet-ups and give choice help to moderately unpracticed specialists in clinical practice [19]. On the other hand, Lindblad Wollmann, Charlotte, et al. described the predicting vaginal birth in ladies with one earlier cesarean and no vaginal conveyances utilizing machine learning strategies [20].

Unlike their works, we showed and analyzed various methodology of supervised classifiers based on a real dataset of childbirth to figure out the best model to predict the suitable mode of delivery.

III. Methodology

To build our proposed model, we went through four significant phases: Dataset formation, data preprocessing, training the models, the performance analysis of the model. We have collected the data from the Tarail Upazilla Health Complex, a specialized clinic for maternal care located in Tarail, Kishorganj, Bangladesh. First, we determined the features that influence our targeted feature, the mode of childbirth. We kept the most significant of them, and some other features having less impact were deleted as they don't contribute much to the targeted variable. Then the data was split into two sections, i.e., training set and test set, which are later used for training and testing correspondingly. After collecting raw data, we had preprocessing to make it suitable for the machine learning model. Data pre-processing techniques have made the data outliers free and more solid, and it also increases the accuracy. As a result, we used several preprocessing steps such as cleaning data, missing value handling, categorical data handing, feature selection, feature scaling. Having completed all the preprocessing steps, the data becomes ready for the machine learning models. We have used several groups of supervised learning classifiers such as Tree, Ensemble, Neighbors, Naive Bayes, Calibration, Discriminant Analysis, SVM, Linear model, Gaussian Process, and Deep Neural Network. Most of those classifiers have shown good performance with this preprocessed training and test dataset. The methodology of our proposed model is depicted in Fig. 1.

A. Dataset Description

We have used a dataset that is containing the medical records of 13527 women. It has 21 diverse observation values for every pregnant woman counting title, age, address, admission time and time, ANC number of shrouds (by therapeutically prepared supplier), para, the reason of confirmation, amid pregnancy (week), cesarean, breech conveyance, partograph, blood circulation, AMTSL, birth weight, PNC-1(postnatal administrations and the status of the patient), PNC1 (postnatal delivery administrations), BP, introduction, layer and cervix (OS).

Para alludes to the total number count of pregnancies that the lady has carried on the last twenty weeks of pregnancy. This number includes both live births and pregnancy misfortunes after twenty weeks. Gravida referred to the number of times of affirmed pregnancies of a lady, including both live birth and interrupted pregnancies. ANC (ante-natal check-up) implies a routine checkup for the mother to ensure appropriate facility for further safety. ANC is usually conducted in three stages, the first one is between 4th to 28th week, the second one is between 28th to 36th week, and the last one is between 36th to 48th week. There is a possibility that the infant can open its mouth almost 10 cm or over, then it can be conveyed in an ordinary way. On the other hand, if curved (OS) isn't 10 cm, even it is over 12 hours, this patient requires a cesarean conveyance.

PNC implies postnatal care. After conveyance, this is often done by checking the typical state of the mother. Cephalic is for typical conveyance, but in some cases, there's a breach at multi-case, but typical conveyance is done. Cesarean conveyance happens sideways or transverse. Pantographs are utilized to decide the physical condition of the mother and child. After children's birth, Placenta is extricated within the AMTCL strategy. Blood circulation is given on the off chance that the quiet is Iron deficient. Most of the time, the layer remains completely intact. Some of the time, it endures from spilling, burst. Blood weight is watched to screen the typical arrangement of the mother.

B. Data Pre-processing

In machine learning, data preprocessing is in the approach of transferring or encoding the raw data in a phase where algorithms can use the data for building the model. We need to preprocess the data accordingly to make it fit for the machine learning model. A well-processed data gives high accuracy and makes the model more reliable. Here, we have used several stages of preprocessing, which have been illustrated in Fig. 2.

In our dataset, there are lots of incomplete, null, and duplicate values. For this reason, we took three steps to correct these data. Those three steps are described below:

- We have noticed that there are many data points that are repeated in a row. Therefore, we simply removed all the duplicate data other than one single observation.
- We also notice some of the rows and columns are empty. We also erased the entire empty rows or columns from our dataset.
- There were some rows and columns that have 50% or more incomplete or null values. We removed the entire rows and columns to fix this issue.
- There were some columns that have very low variance. We also ignored those columns to make the dataset better suited for our machine learning model.

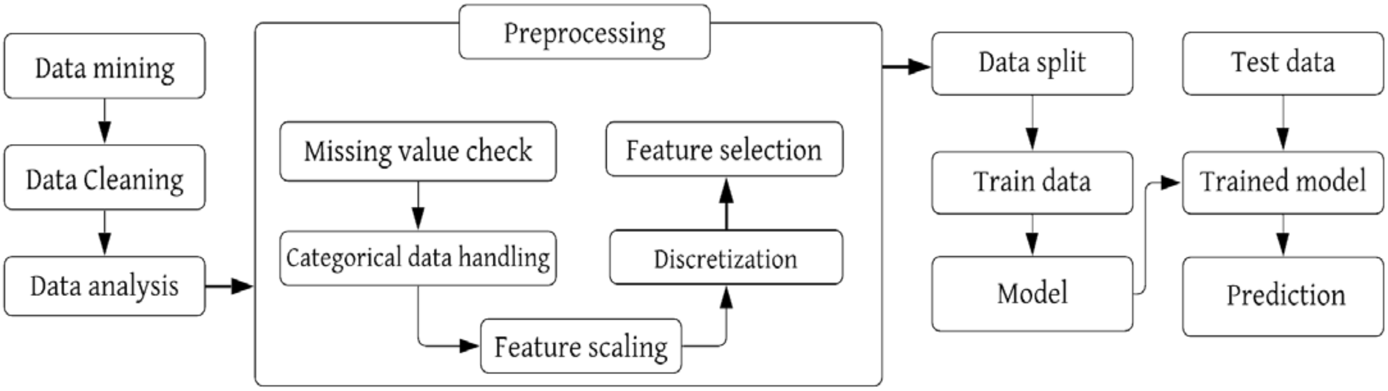


Fig. 1. Workflow of the Proposed Model.

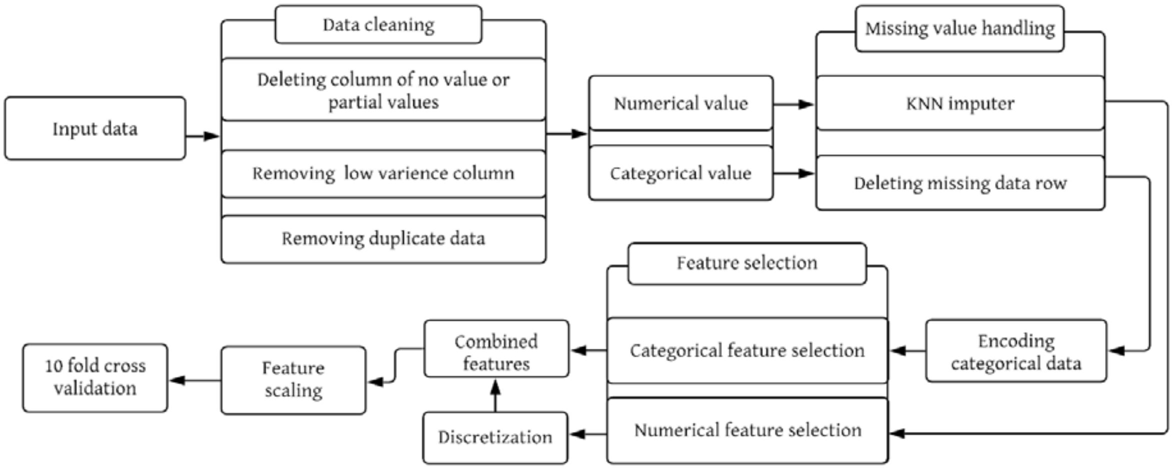


Fig. 2. Data Preprocessing Stages.

Generally, a missing value is defined as the value which is either not stored in the sample or missing partial information. The missing value is commonly seen in any kind of dataset. Our dataset also had some missing values. However, most predictive modeling methods can't handle any missing value. Hence, this issue must be solved before we feed this data into the machine learning model. Sometimes, median, mean, mode methods are used to update a missing value. The selection of the methods depends on the data types and the totals number of observations. However, the most straightforward procedure for dealing with the missing value is to remove the whole row for categorical features and replace the missing value by selecting the nearest neighbors for numerical data. We have used the K Nearest Neighbor aka KNN based method for a more accurate missing value imputation and replaced the NAN data by getting the nearest neighbor value. We have considered three neighbors for KNN algorithm implementation and completed all the missing values using KNN imputer to build a perfect feature matrix. The following Fig. 3 has illustrated the working procedure of the KNN imputer.

Categorical data is a qualitative feature whose values are taken based on the value of labels. So, we need to encode this type of data into numbers so that the machine learning model can implement mathematical operations on it. In our dataset, there are a total of three categorical variables, including "PRESENTATION", "REASON" and "MEMBRANE". We have used one-hot encoding, one of the most popular encoding algorithms, to encode the categorical values into numbers. It is the most general approach, and it works well unless any categorical variable takes a large number of different values. After this encoding, a binary matrix is formed where 1 indicates the presence of any value and 0 indicates the absence of the value.

Feature selection is a critical stage of implementing a machine learning model. It is the process of determining the mathematical relation between the feature variable and the target variable. We have kept the most significant features and dropped some features with less significance on the targeted variable. Reduction in features reduces the computational cost. Our dataset contains 21 features, and we have considered the p-value for finding the probability of the null hypothesis. The features with a p-value less than 0.05have been taken out. After checking multicollinearity, we have maintained a strategic distance from those components, which show repetition and don't back the p-value assumption. Besides, to handle the numerical feature, we took the help of the Pearson correlation coefficient, which is defined in equation -1, and for categorical features, we used the ANOVA F measurement, which is described in equation -2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{\mathrm{X}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

$$F = \frac{n\sum_{i=1}^{n}(\overline{x_k} - \overline{x_G})^2/(k-1)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_k})^2/(N-k)}} \qquad (2)$$

After performing the feature selection, we had the most relevant nine features, including Para, Age, Cervix (OS), Gravida, Systolic, Diastolic, Reason, FHR(BPM), and Presentation.

In data analysis, it can often be observed that the numerical data are mostly like skewed or non-standard deviation due to outliers, multi distributions, very exponential distributions, and more. We converted the numeric value into categorical behavior to solve this problem. We have applied the discretization method to converts the numerical value into a distribution function.

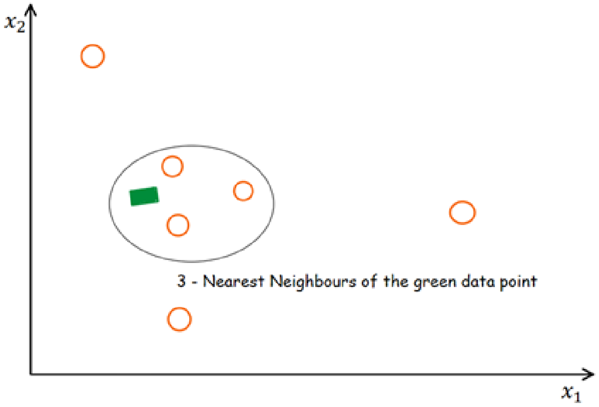


Fig. 3. The KNN Imputer for Missing Value Handling

Feature scaling is one of the crucial techniques that are mandatory to standardize the working data's independent features. Nevertheless, there exist various methods like Min-Max Scaling, Variance Scaling, Standardization, Mean Normalization, and Unit vectors for feature scaling. In our work, we have applied the Min-Max scaling as a feature scaling technique. Here, the transferred range between 0 and 1. The min-max scaling can be written as shown in equation- 3 below,

$$\acute{x} = \frac{x-\min(x)}{max(x)-\min(x)} \quad (3)$$

Cross-Validation is used to assess the models' predictive performance and judge the performance of a new data set. This is often fundamentally a variation of Recursive Feature Elimination (RFE) with cross-validation in each iteration. In RFE, the features are disposed of using a backward selection of the features. It uses a base classifier for selecting the features of data. It starts by building a model with the complete set of features and computes an importance score for each variable. The features with the least imperative score are deleted. The method recursively finds the optimal set of features that gives the most excellent model accuracy. With the cross-validation, the training and test set is divided into k number of folds where the k-1 fold of them is used for training and the rest 1 fold of data is used for testing. After running iteratively for K times, we actually get the average accuracy. This method takes more computation, but we can ensure good accuracy with comparatively fewer data. In our project, we used ten-fold cross-validated features to build our classification models.

C. Model Description

Implementing a lot of diverse models can ensure the best possible accuracy. So, we have applied the 32 most sensible machine learning classifiers, including Tree, Ensemble classifier, Neighbors, Naive Bayes, Calibration, Discriminant Examination, SVM, Linear model, to predict the birth mode Gaussian Process, and deep neural networks. All models are shown in Fig. 4.

Tree-based algorithms are considered to be one of the leading and most used supervised learning methods. In this work, we have implemented a decision trees and additional tree classifiers. In these two algorithms, we utilized "gini" for the Gini impurity, and the splitter is chosen as 'best' to select the part at each node.

Ensemble methods are procedures that make multiple models and combine them to create moves forward. Here, we utilized five ensemble-based classifiers [21]. These are AdaBoost, Stowing, Gradient Boosting, enable hist gradient boosting, Random Forests classifiers. In all these classifiers, in AdaBoost classifiers, the number of boosting estimators is 50 with the SAMME.R as a real boosting algorithm. In the bagging classifier, we used ten estimators. The loss function of Gradient Boosting is 'deviance', logistic regression with probabilistic outputs with the 100 boosting stages. Besides, in the section of random forest three, we use 100 trees as a forest with Gini impurity.

Afterward, we have utilized three neighbors' classifiers of statistical pattern recognition [22]. These are radius neighbor, k-neighbor, and nearest centroid. In KNN, we used five neighbors for every iteration. Besides, the Makowski metric is chosen for all neighbor classifiers.

Naïve Bayes is based on an estimate of the age of the "naive" with a set of learning calculations guided by an application and an estimate of the accuracy between them [23].

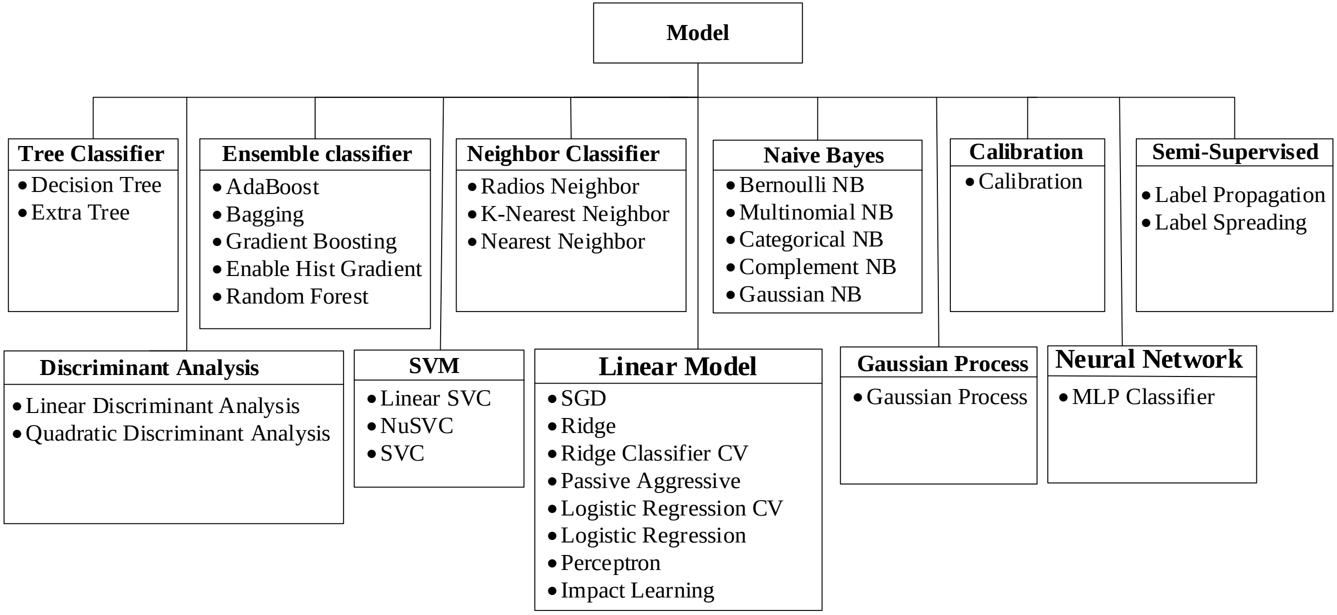


Fig. 4. Tabulation of all the Models.

Here Bernoulli, Multinomial, Categorical, Complement, and Gaussian Naive Bayes are executed to compare the Bayes algorithm on childbirth mode detection. For each method, we have also used the additive smoothing parameter.

Model calibration implies the process where we take a model that is already trained and apply a post-processing operation, which improves its probability estimation. Thus, if we were to inspect the samples that were estimated to be positive with a probability of 0.85, we would expect that 85% of them are in fact positive.

The label propagation and label spreading are used in the area of semi-supervised algorithms. The Gamma Parameter for RBF bit is utilized as 20. The maximum emphasis is 1000, and the neighbor parameters are 7.

The researchers commonly use discriminant analysis to analyze the data when the criterion or the dependent variable is categorical, and the predictor or the independent variable is the interval in nature. Dependent variables should categorize at the moment as well as include predictive or distinct variable natures such that researchers can use them to analyze research data (quadratic inequality analysis (QDA) and linear inequality analysis (LDA) [24].

The support vector machine, aka SVM, is used mainly for exploring a hyperplane in d-dimensional space that notably classifies the data points [25]. In the linear SVC, we used hinge as loss function with l_2 penalty. The numerical value three is used as the polynomial kernel in NuSVC with the RBF kernel type.

The linear model could be a module lesson if it contains diverse functions for performing machine learning linearly [26]. We utilized the eight classifiers such as SGD, Ridge Classifier, Ridge Classifier CV, Passive-Aggressive Classifier, Logistic Regression, Logistic Regression CV, Perceptron, and Impact Learning [27].

The Gaussian process is a stochastic method in probabilistic hypotheses and statistics, such as a common multivariate distribution containing which is a limited random sample collection. The kernel of the gaussian process classifier specifies the covariance function, and the accessible internal optimizers are 'fmin_l_bfgs_b'.

A neural network could be an arrangement of algorithms that endeavors to recognize basic relationships in a set of data through a method that imitates the way the human brain works. The Artificial Neural Network (ANN) is a computing system where neurons inspire people [28]. There are three layers, and these are the input layer, hidden layer, and output layer. The input layer usually takes the input data into the network. The hidden layer is the layer where input and output are connected based on conditions. The output level is decided by considering the respect action, weight, and hidden level. There's no rule of the thumb to select the hidden layer in ANN. We have used sixty-four hidden layers between the input and output layers.

IV. Experiment

In furtherance of our experiment from the proposed work, we have first assembled the model and trained it. Thirty-two classifiers from supervised learning based on different learning methodologies have been implemented to predict childbirth's most applicable mode. This section described different experimental tasks for the performance analysis and evaluation and compared all algorithms. Besides, we have illustrated the experimental setup used to execute the whole task and used 11 statistical evaluation metrics for analysis performance. Finally, we have also compared with other works related to this issue regarding the best version of our work.

A. *Experimental Setup*

We have completed the whole computation in google colab, a python simulation environment provided by Google. This environment comes with parallel computation facilities for fast execution. We have used the most popular libraries to make easy and expressive data structures work well and intuitively with fast, flexible, and time-series data. Finally, the scikit-learn Library contains specialized machine learning and statistical modeling tools, including classification, regression, and clustering algorithms for modeling. We have used a machine learning framework named sci-kit learn and deep learning framework Keras to implement the classification algorithm. Finally, we used Matplotlib for data visualization, graphical representation, and also for data analysis.

B. *Measurement Metrics*

We have used several Statistical metrics [29] for measurements, evaluation, and analysis of the performances and compared all the algorithms. We will define and describe all of those in the following section.

We have analyzed the 11 statistical measurements, including accuracy, F1 score, precision, recall, and so on. Accuracy and F1 score are the most important of them.

Accuracy is a metric that evaluates the matric for the correct prediction rate for the positive class. The expression is shown below in equation 4.

$$Accuracy = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)} \tag{4}$$

F1 score conveys the balance between the precision and the recall. It is also called the F Score or the F Measure. A good F1 score indicates that we have low false positives and low false negatives in the results. The expression of the F1 score is shown in equation 5.

$$F1 = \frac{TP}{TP + 0.5(FP + FN)} \tag{5}$$

Recall considers the percentage of correct predictions for all the positive categories. In other words, recall is how many of the true positives were recalled (found), i.e., how many of the correct hits were also found. The expression of recall score is shown below in equation 6.

$$RS = \frac{TP}{TP + FN} \tag{6}$$

The F-beta score is evaluated in the binary classification model based on a configurable single-score for the positive class's forecasts. It's also calculated utilizing precision and recall. The value of the F-beta score can be calculated using equation 7 below.

$$FBS = \frac{(1+\beta^2)\,.(Precision\,.Recall)}{(\beta^2\,.Precision\,+Recall\,)} \tag{7}$$

Hamming loss is designed for multiclass while Precision, Recall, F1-beta score represents one clear single-presentation-value for multiple-label cases compared to the precision/recall/f1beta score that can be assessed only for independent binary classifiers for each label. The expression of Hamming loss is shown below in equation 8.

$$HL = \frac{1}{|D|}\sum_{i=1}^{|D|} \frac{|Yi\,\Delta\,Zi|}{|L|} \tag{8}$$

In the Jaccard similarity coefficient, the union of the two label sets is used to compare the set of labels predicted in y_true to mark the separate intersection as a measure by calculation. The equation of Jaccard similarity coefficient is shown below in equation 9.

$$\mathrm{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

Matthews Correlation Coefficient, aka MCC, is used as a standard for binary and multiclass classification in machine learning. The equation of Matthews Correlation Coefficient is shown below in equation 10.

$$\mathrm{MCC} = \frac{TP\,.FN\,-FP\,.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \tag{10}$$

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

The balanced accuracy in binary (BAC) and multiclass classification is usually used to measure the performance if the dataset is imbalanced. The equation of balance accuracy is shown below in equation 11.

$$\mathrm{BAC}=\frac{1}{2}\left(\frac{TP}{P}+\frac{TN}{N}\right) \tag{11}$$

Cohen's kappa (CKS) is a statistic that measures inter-annotator agreement. We can consider Cohen's Kappa as a quantitative measure of reliability for two raters that are rating the same thing. The equation of Cohen's kappa (CKS) is shown below in equation 12.

$$\mathrm{K} = \frac{P_{0-}\,P_E}{1-\,P_E} \tag{12}$$

V. Result Analysis

Our experiment improved the methods of decision-making. We have implemented thirty-two classifier parameters to gain the best possible performance. After getting the performance matrix from all the models, we have tabulated the data into a table. We have analyzed the 11 statistical measurements, including accuracy, F1 score, RS, PS, FBS, HL, JS, MS, AUC, BAC and CKS. The statistical measure and for each of the algorithms is shown in Table I below. In addition, we have compared the performance of all the proposed models and determine the best suitable model that can be used in real-life decision-making or selecting the most suitable mode of childbirth.

From Table I, we can see that the decision tree classifier has predicted the best accuracy of 0.918307, and the F1 score is 0.918198 from the branch of the Tree. Secondly, the hist gradient boosting classifier has gained the best accuracy of 0.959158 with an F1 score is 0.959071 from the section of Ensemble's algorithm. Thirdly, the KNN classifier has acquired better accuracy, which is 0.961015, along with an F1 score of 0.960853 from the area of neighbor's classifiers. Also, from the section of naive Bayes algorithms, we can figure out that Gaussian naive Bayes has shown the best accuracy of 0.874381 and its F1 score is 0.872635 among all naive Bayes classifiers. Next, calibration also has placed the best accuracy of 0.877063, and its F1 score is 0.875664 from the branch of naive Bayes. After that, we can see from the semi-supervised classifiers, the label propagation has gained the best accuracy of 0.906972, and the F1 score is 0.906223. Besides, it can also be seen from the discriminant analysis section that the quadratic discriminant analysis has proved the best position of accuracy 0.979992 with an F1 score of 0.979962. Moreover, SVC is the best for predicting childbirth mode with an accuracy of 0.956477 and an F1 score of 0.956302 from all algorithms of SVM. Furthermore, we also can find out that the Gaussian process classifier placed the accuracy of 0.891708 and the F1 score is 0.890333. In the neural network, the multilayer perceptron classifier has acquired a good performance with an accuracy of 0.954404, and the F1 score is 0.954299.

Overall, by considering all the sections of algorithms for the prediction of childbirth mode, we can observe that the quadratic discriminant analysis is the winner with an accuracy of 0.979992 and the F1 score is 0.979962. The neighbors' classifier also comes up with the second-best with an accuracy of 0.961015, and the F1 score is 0.960853.

TABLE I. PERFORMANCE ANALYSIS

Name	Accuracy	F1S	RS	PS	FBS	HL	JS	MCC	AUC	BAC	CKS
Tree Classifiers											
DtC	0.918317	0.918198	0.918236	0.918166	0.918178	0.081683	0.850266	0.877476	0.951797	0.918236	0.877473
ETC	0.902228	0.902208	0.902168	0.902268	0.902242	0.097772	0.823259	0.853352	0.93791	0.902168	0.853342
Ensemble Classifiers											
AdC	0.772896	0.760404	0.77305	0.79518	0.774519	0.227104	0.637223	0.681965	0.857822	0.77305	0.659515
BC	0.94637	0.946173	0.946315	0.946145	0.946142	0.05363	0.898545	0.919609	0.970062	0.946315	0.919553
GBC	0.950083	0.94991	0.950032	0.949875	0.949878	0.049917	0.905279	0.925165	0.971867	0.950032	0.925123
HGBC	0.959158	0.959071	0.959129	0.959056	0.959057	0.040842	0.921829	0.938758	0.977144	0.959129	0.938737
RFC	0.956271	0.95605	0.956208	0.956053	0.956032	0.043729	0.916453	0.934484	0.975667	0.956208	0.934404
Neighbors Classifiers											
RNC	0.796823	0.793171	0.796323	0.795863	0.794037	0.203177	0.667912	0.697534	0.878434	0.796323	0.695157
KNC	0.961015	0.960893	0.960944	0.960897	0.960889	0.038985	0.925345	0.941547	0.977879	0.960944	0.94152
NC	0.808375	0.803996	0.80814	0.804359	0.803641	0.191625	0.681652	0.714516	0.886066	0.80814	0.712552
Naive Bayes Classifiers											
BNB	0.331271	0.165918	0.333126	0.110469	0.127515	0.668729	0.110446	-0.00883	0.499843	0.333126	-0.00031
MNB	0.808168	0.803797	0.808056	0.805796	0.804182	0.191832	0.681258	0.715031	0.894046	0.808056	0.71227
CNB	0.336015	0.16767	0.333333	0.112005	0.129157	0.663985	0.112005	0	0.5	0.333333	0
CoNB	0.718647	0.681511	0.718514	0.74856	0.699551	0.281353	0.539377	0.609911	0.834682	0.718514	0.577959
GNB	0.874381	0.872635	0.874237	0.872807	0.872524	0.125619	0.777803	0.81239	0.93513	0.874237	0.81156
Calibration Classifier											
CC	0.877063	0.875664	0.87692	0.87552	0.875444	0.122937	0.782584	0.816105	0.935097	0.87692	0.815584
Semi-Supervised Classifier											
LP	0.906972	0.906223	0.906743	0.90729	0.906667	0.093028	0.831271	0.861163	0.947339	0.906743	0.860439
LS	0.902847	0.901956	0.902615	0.902912	0.902333	0.097153	0.824296	0.854988	0.945149	0.902615	0.85425
Discriminant Analysis Classifiers											
LDA	0.870462	0.868522	0.870233	0.868103	0.868115	0.129538	0.774095	0.806278	0.922529	0.870233	0.805682
QDA	**0.979992**	**0.979962**	0.979964	0.979987	0.979974	0.020008	0.960818	0.97	0.990303	0.979964	0.969987
SVM Classifiers											
LSVC	0.880982	0.879968	0.880805	0.879558	0.879671	0.119018	0.790406	0.821664	0.930283	0.880805	0.821466
NuSVC	0.924711	0.923732	0.924601	0.924279	0.923889	0.075289	0.860045	0.887741	0.960271	0.924601	0.887061
SVC	0.956477	0.956302	0.956403	0.956278	0.956279	0.043523	0.917007	0.934751	0.975629	0.956403	0.934713
Linear Model Classifiers											
SGDC	0.871493	0.869171	0.871364	0.870912	0.869713	0.128507	0.774591	0.809034	0.922391	0.871364	0.807246
RdC	0.865924	0.863561	0.865752	0.864239	0.863613	0.134076	0.764468	0.800186	0.919536	0.865752	0.798875
RdCV	0.865924	0.86358	0.865754	0.864204	0.863609	0.134076	0.764519	0.800154	0.920406	0.865754	0.798875
PAC	0.828383	0.82831	0.82787	0.867552	0.846668	0.171617	0.714729	0.761057	0.893194	0.82787	0.742496
LRCV	0.891502	0.890906	0.891326	0.89062	0.890718	0.108498	0.808021	0.837313	0.932856	0.891326	0.837249
LR	0.883457	0.882394	0.883287	0.882026	0.88211	0.116543	0.79431	0.825425	0.931482	0.883287	0.82518
Pr	0.837871	0.836795	0.837884	0.842252	0.839304	0.162129	0.727938	0.760099	0.903731	0.837884	0.756858
IL	0.880982	0.879634	0.880853	0.879964	0.879639	0.119018	0.790078	0.822198	0.930677	0.880853	0.821475
Gaussian Process Classifiers											
GPC	0.891708	0.890444	0.891566	0.890383	0.890278	0.108292	0.80634	0.838053	0.938326	0.891566	0.837557
Neural Network Classifier											
MLPC	0.954414	0.954259	0.954337	0.954276	0.954258	0.045586	0.913242	0.931665	0.975341	0.954337	0.931619

VI. Conclusion and Future Work

Selection of the best baby delivery methods is crucial for protecting both the mother and the newborn baby. But it still remains to explore the best sets of features when making this decision in a computerized way. That's why we try to leverage AI to recognize the best mode of baby delivery. Nowadays, machine learning, deep learning, and other computerized computation models are ubiquitously being used in medical decision-making. Here, we have used machine learning-based binary classification algorithms for decision-making between two methods of childbirth. This model will assist the doctor in making a more accurate decision within a very short time. This machine-learning-aided decision will not replace the necessity of the doctors for decision-making. Instead, it will help the physician to gain a deeper insight into the patient's information available. The way of decision-making using this model is very computerized and less likely to have an error.

The dataset we have used in this project is not very robust. In the future, we want to add much more observation to our dataset and make this model much more general. We believe a large set of data will produce better accuracy and less overfitting. Besides, we will implement a more in-depth learning-based classification to expand the investigation and make the top choice for record-breaking performance. After childbirth, we plan to implement this system to predict other real-life biomedical factors in advance. In addition, the data can be collected during the whole nine months of the mother's pregnancy. In the future, we plan to make a GUI of this model available to physicians, who can use it as like medical device for decision making.

References

[1] Maeda, Eri, et al. "Cesarean delivery rates for overall and multiple pregnancies in Japan: A descriptive study using nationwide health insurance claims data." Journal of Obstetrics and Gynaecology Research (2021).

[2] Chen, Qian, et al. "The impact of cesarean delivery on infant DNA methylation." BMC pregnancy and childbirth 21.1 (2021): 1-8.

[3] Ali, Nazia Binte, et al. "Are childbirth location and mode of delivery associated with favorable early breastfeeding practices in hard to reach areas of Bangladesh?." Plos one 15.11 (2020): e0242135.

[4] Serçekuş, Pınar, Okan Vardar, and Sevgi Özkan. "Fear of childbirth among pregnant women and their partners in Turkey." Sexual & Reproductive Healthcare 24 (2020): 100501.

[5] Shaheen, Razia, et al. "Prevalence of Low Birth Weight in Urban Dhaka and its Association with Maternal Age and Socioeconomic Status." Dr. Sulaiman Al Habib Medical Journal 2.4 (2020): 162-166.

[6] Mboya, Innocent B., et al. "Prediction of perinatal death using machine learning models: a birth registry-based cohort study in northern Tanzania." BMJ Open 10.10 (2020): e040132.

[7] Abraham, Abin, et al. "Dense phenotyping from electronic health records enables machine-learning-based prediction of preterm birth." medRxiv (2020).

[8] Islam, Muhammad Nazrul, et al. "Exploring Machine Learning Algorithms to Find the Best Features for Predicting Modes of Childbirth." IEEE Access (2020).

[9] Kowsher, M., Prottasha, N. J., Tahabilder, A., & Islam, M. B. (2020, February). Machine Learning Based Recommendation Systems for the Mode of Childbirth. In International Conference on Cyber Security and Computer Science (pp. 295-306). Springer, Cham.

[10] Khan, Nafiz Imtiaz, et al. "Prediction of Cesarean Childbirth using Ensemble Machine Learning Methods." (2020).

[11] Fu, Yuanqing, et al. "Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: a prospective birth cohort." BMC medicine 18.1 (2020): 1-10.

[12] Siddiqui, Mohammad Khubeb, et al. "A review of epileptic seizure detection using machine learning classifiers." Brain informatics 7 (2020): 1-18.

[13] Soh, Yan Xi, et al. "Determinants of childbirth self-efficacy among multi-ethnic pregnant women in Singapore: a structural equation modeling approach." Midwifery 87 (2020): 102716.

[14] Chen, Yanfang, et al. "Association between mode of birth and post-traumatic stress disorder following childbirth: a prospective cohort study of Chinese women." (2020).

[15] Zhang, Yiye, et al. "Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women." Journal of Affective Disorders 279 (2020):1-8.

[16] Lipschuetz, Michal, et al. "Prediction of vaginal birth after cesarean deliveries using machine learning." American journal of obstetrics and gynecology 222.6 (2020): 613-e1.

[17] Serçekuş, Pınar, Okan Vardar, and Sevgi Özkan. "Fear of childbirth among pregnant women and their partners in Turkey." Sexual & Reproductive Healthcare 24 (2020): 100501.

[18] Onchonga, David, et al. "Midwife-led integrated pre-birth training and its impact on the fear of childbirth. A qualitative interview study." Sexual & Reproductive Healthcare 25 (2020): 100512.

[19] Liu, Lijue, et al. "Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor." Computer Methods and Programs in Biomedicine 196 (2020): 105624.

[20] Lindblad Wollmann, Charlotte, et al. "Predicting vaginal birth after previous cesarean: using machine ‐ learning models and a population ‐ based cohort in Sweden." Acta obstetricia et gynecologica Scandinavica (2020).

[21] Pintelas, Panagiotis, and Ioannis E. Livieris. "Special issue on ensemble learning and applications." (2020): 140.

[22] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." International Conference on Machine Learning. PMLR, 2018.

[23] Zhang, Huan, Liangxiao Jiang, and Liangjun Yu. "Class-specific attribute value weighting for Naive Bayes." Information Sciences 508 (2020): 260-274.

[24] Charles, Vincent, Juan Aparicio, and Joe Zhu. "The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis." European Journal of Operational Research 279.3 (2019): 929-940.

[25] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.

[26] Stroup, Walter W. Generalized linear mixed models: modern concepts, methods and applications. CRC press, 2012.

[27] Kowsher, Md, Anik Tahabilder, and Saydul Akbar Murad. "Impact-learning: a robust machine learning algorithm." Proceedings of the 8th International Conference on Computer and Communications Management. 2020.

[28] Kukreja, Harsh, et al. "An introduction to artificial neural network." Int J Adv Res Innov Ideas Educ 1 (2016): 27-30.

[29] Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." Electronics 8.8 (2019): 832.

Power-based Side Channel Analysis and Fault Injection: Hacking Techniques and Combined Countermeasure

Noura Benhadjyoussef[1], Mouna Karmani[2], Mohsen Machhout[3]
Faculty of Sciences of Monastir, Electronics and MicroElectronics Laboratory (LEME)
Monastir 5019, Tunisia

***Abstract*—Over the last years, physical attacks have been massively researched due to their capability to extract secret information from cryptographic engine. These hacking techniques are based on exploiting information from physical implementations instead of cryptographic algorithm flaws. Fault-injection attacks (FA) and Side-channel analysis (SCA) are the most popular techniques of implementation attacks. Aiming to secure cryptographic devices against such attacks, many studies have proposed a variety of developed and sophisticated countermeasures. Hence, the majority of these secured approaches are used for precise and single attack and it is difficult to thwart hybrid attack, such as combined power and fault attacks. In this work, the Advanced Encryption Standard is used as a case study in order to analyse the most well-known physical-based Hacking techniques: Differential Fault Analysis (DFA) and Correlation Power Analysis (CPA). Consequently, with the knowledge of such contemporary hacking technique, we proposed a low overhead countermeasure for the AES implementation that combines the concept of correlated power noise generating with a combined-approach based fault detection scheme.**

Keywords—Advanced encryption standard; fault attack; power attacks; combined countermeasure; hardware implementation

I. Introduction

From a data security viewpoint, securing secret information requires using algorithms that resist theoretical hacking techniques. Though, treating an algorithm in a purely mathematical way or, in other words, shying away from its physical implementation opens the door to numerous threats in the real-world security. In the modern age of electronics, cryptanalysis attempts to reveal sensitive data based on physical property of a cryptographic device rather than making use of the theoretical flaws in the implemented cryptographic algorithm. Indeed, as the cryptographic algorithms are implemented on a physical platform, they are susceptible to well-known physical attacks, namely Fault Attacks (FA) and Side-Channel Analysis (SCA). These two classes of attacks exploit the physical interactions with cryptographic systems to break their security and extract secret information. These attacks are practical due to their methods to reveal the secret information from most cryptosystems which supposed to be cryptanalytically secure.

In SCA hacking technique, a passive adversary observes platform side-channel information to reveal the sensitive information. Indeed, these devices leak sensitive correlated information in the form of electromagnetic emissions (EM), time execution, power consumption, allowing a hacker to reveal the secret key from the cryptographic device [1]–[4].

On the other hand, FA hacking technique is an active cryptanalysis based on perturbing the cryptographic device processing in order to obtain an abnormal behaviour. The hacker then exploits the erroneous of the cryptographic device result to retrieve the secret key [5], [6]. Many studies combine SCA and FA in order to form even more sophisticated attacks [7], [8].

Aiming to secure cryptographic devices against such real-world attacks, countermeasures must be thus designed to harden cryptographic implementations before they are used in the wild. Many studies have proposed a variety of developed and sophisticated countermeasures. In particular, countermeasure for the AES crypto-core has been massively researched for many years against both fault injection Attack and power analysis attacks.

For SCA, most implemented countermeasures aim to decreasing the signal-to-noise ratio (SNR) by using two key approaches; the noise insertion or the leaked sensitive correlated information destruction. These countermeasures are categorized as logical [9], architectural [10], [11], and circuit-level countermeasures [12]–[14]. For the logical and architectural countermeasures, the used approach is specific to the crypto-core and design. On the other hand, physical countermeasures are nonspecific and can be used to protect any crypto-core by providing cover around it [2].

Error detection schemes against FAs are, generally, based on some redundancy approaches. Either using the temporal redundancy, where a given operation is computed twice, or using hardware redundancy by executing the same transformation at the same time to compare the obtained results and check whether a fault was induced. Adding correction and error detection codes to intermediate values is another option to protect the considered cryptographic system named information redundancy [15], [24].

Although the secured approaches have been widely studied for each kind of physical attack, the study of combined countermeasures has not been well explored in the existing literature. In this paper, we perform an in-depth study of Differential fault attack (DFA) and Correlation power analysis

(CPA) and we propose a dual complementary AES cryptographic circuit to defend against both fault and power SCA attacks.

The main contributions of this paper are as follows:

1) We firstly present a fault based Hacking technique to indicate how a fault injection can be useful to extract the sensitive data of the AES crypt core. In the suggested case study, we clarify the main procedures that can threat the security of the considered AES design basing on DFA attack.

2) We study the power based Hacking technique and we perform a successful CPA on the FPGA based AES implementation using the Side-channel Attack Standard Evaluation Board (SASEBO).

3) We develop a combined fault detection scheme to secure the AES cryptocore based on an error detection code for linear AES transformation and temporal redundancy for the nonlinear SubByte transformation. This proposed scheme can be applied for both LUT and GF SubByte transformation implementations.

4) To avoid information leakage, the proposed fault detection scheme is enhanced using a correlated power noise generator. This enhanced countermeasure eliminates the AES cryptocore power correlation with the secret key by adding an interfering power signal which depends on the manipulated data and a nosey key.

5) Finally, we implement our AES cryptocore with the proposed countermeasure on a Virtex V Xilinx field-programmable gate array (FPGA) device. Moreover, to investigate the practicality and effectiveness of the proposed architecture, we compare our implementation results to similar secured AES implementations presented in literature.

The paper is organized as follows. Section 2 introduces the proposed AES Faults-based Hacking technique using the DFA attack. Section 3 studies the AES power Side-Channel analysis using the CPA attack. Section 4 presents the proposed fault-resilient AES implementation. The proposed power based SCA-Countermeasure for AES implementation is presented in Section 5. The experimental results and comparison with previous works reported in the literature is given in Section 6. Section 7 concludes the paper.

II. The Faults-based Hacking Technique using the DFA Analysis against AES Implementation

A. Related Works

Fault attacks exploit the possibility to inject a fault into cryptographic devices in order to reveal the secret key. The fault injection is done by means of Electromagnetic field, supply voltage variation, laser beam, or temperature control. In particular, Differential Fault Attacks on AES has become a widely research topic using different fault-models; single-bit, single-byte fault and multibyte fault.

In [16], two DFA attacks on AES are proposed, the first attack inject a theoretical single-bit fault into an intermediate result allowing hacker to extract the AES-128 secret key with 50 faulty ciphertexts. While the second attack inject a byte-fault and reveal the key with less than 250 faulty ciphertexts.

Reference [17] presents an improved DFA attack approach on AES using unknown and random multi-byte faults. The authors focused on combined fault model that inject single-byte and multi-byte faults. Their attack showed that about 97.3% of the attacks can be completed within 3 pairs of correct and faulty ciphertexts.

In [18], authors presented a DFA combined fault model combining a single-byte random faults model in encryption process with a single-byte faults model in the key schedule process. Their experimental results showed that 6 pairs of correct and faulty ciphertexts could reveal the AES-128 secret key. Reference [19] shows that a FA can break the advanced encryption standard (AES) by exploiting the existing target devices Input/Output signals instead of the artificial triggering implementation. Indeed, authors identify fault injection time by employing target devices electromagnetic emission. Using one-byte fault model, the attack was successfully executed 55 times out of the 1000 conducted fault injection attacks overall.

In this paper, we aim to propose a low-cost fault-resilient AES architecture that resists side-channel attack. So, as a first step, we must study how a fault injection can be used to extract the cryptographic key. In this step, the Giraud's Single-Bit Differential Fault Analysis [16] is adopted as a case study to give details of the main procedure that can menace the security of the considered AES implementation. Consequently, with the knowledge of the considered Fault-hacking technique, we will propose the adequate DFA countermeasure for AES cryptocore.

B. The Fault Injection Step

In this step, we have adapted the single-bit fault model to simulate the physical real defects. This model assumed that only one bit in the considered circuit is faulty and supposed that one line in the circuit behaved as if it is at logic 0 or logic 1. For the multiple-bit fault model it is assumed that the same basic assumptions as the single-bit fault models, except allowing two or more lines in the circuit to be faulty at the same time.

For our considered DFA attack, the Single-Bit model is adopted, where only one bit was injected at the beginning of the final round (see Fig. 1).

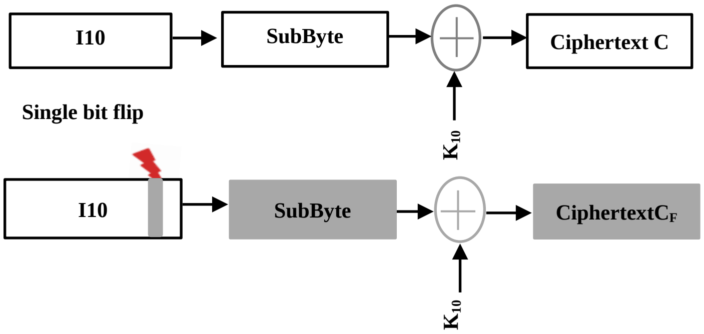


Fig. 1. The Single-Bit Fault Injection into the Input of the 10th Round.

C. The Fault Propagation Step

Using the DFA Single-Bit model, only a faulty one-bit *'e'* was injected in the output of the 9th AES round (I_{10}). When flipping a single bit between the MixColumns of the ninth round and the SubBytes of the tenth round, the changed bit spreads in the last round and generates a faulty Ciphertext (CF) with single faulty byte.

As presented in Fig. 2, the injected fault into the AES-128 bloc modifies 8-bit through the SubBytes operation. Indeed, this non-linear operation is a byte substitution and executes each 8-bit input separately. The affected byte is Xor-ed with another round key byte of the tenth round and produces one differential fault.

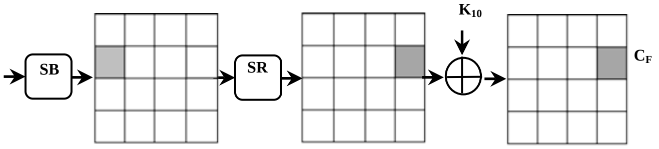


Fig. 2. Propagation of Injected Fault in $I_{10.}$

D. The Fault Exploitation Step

In order to reveal the AES secret key, the hacker exploits the observable fault by exploiting some relation between the two obtained ciphertexts. Indeed the hacker must repeat the experiment with the same plaintext and same key but without inducing fault. As a result, two ciphertexts derived from the same plaintext and key are obtained, where one of the ciphertext is fault-free (C) and the other is faulty (CF).

As a first step, the hacker tries to solve these equations.

$$C = SB\,(I_{10}) \oplus k_{10} \tag{1}$$

$$C_F = SB\,(I_{10} \oplus e) \oplus k_{10} \tag{2}$$

$$\Delta = C \oplus C_F \tag{3}$$

where $SB(I_{10})$ is the result of the SubByte transformation applied on one byte of the 9th round inputs (I_{10}) and k_{10} is the 10th round key corresponding byte (k_{10}). While Δ is the injected fault differential. As the single bit flip on I_{10} is the adapted fault model, the Δ Hamming weight (HW) must equal 1. In order to reveal the *k10* value, the hacker must ensure a full exploration of all possible key values. Therefore, the hacker first computes for each key-assumption ($\tilde{k}$) of the real-key byte (k), the corresponding hypothesized fault differential $\tilde{\Delta}$ as follows:

$$\widetilde{I_{10}} = SB^{-1}(C \oplus k\widetilde{10}) \tag{4}$$

$$\widetilde{I_{10F}} = SB^{-1}(C_F \oplus k\widetilde{10}) \tag{5}$$

$$\tilde{\Delta} = \widetilde{I_{10}} \oplus \widetilde{I_{10F}} \tag{6}$$

where $\widetilde{k_{10}}$ denotes the hypothetical key and $\widetilde{\mathrm{I}_{10}}$ is the input of the corresponding tenth round. Finally, the hacker must verify if the calculated $\tilde{\Delta}$ is identical to Δ. Indeed, the hypothetical key $\widetilde{k_{10}}$ may be a correct assumption for the real key k_{10} if the $\tilde{\Delta}$ Hamming weight is equal to 1. Otherwise, the hypothetical key $\tilde{k}$ is rejected. This process is recomputed for each k_{10} byte to reveal the overall round key k_{10}.

Fig. 3 present the number of injected fault needed to retrieve the whole 16-byte last-round key of AES-128. As shown in Fig. 2 the considered hacking technique needs only 32 fault injections to extract the whole 128-bit tenth round key. Finally, since the AES key expansion is invertible, the hacker can compute the original key (k_0) going backwards.

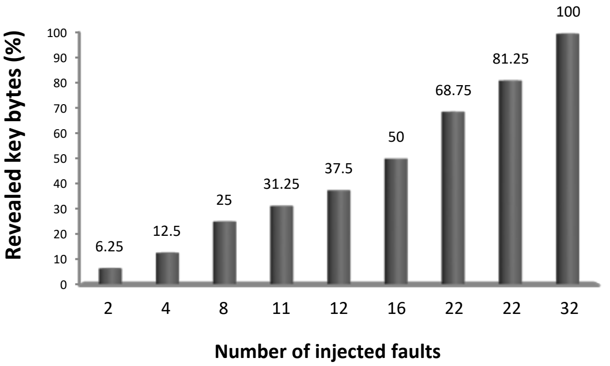


Fig. 3. The DFA Attack Results.

III. The Power Analysis-based Hacking Technique using the CPA Attack Against AES Implementation

Power-based side-channel attacks assume that there is a correlation between the level of power consumption and cryptographic operations manipulated by the cryptocore.

Simple power analysis (SPA) [20], differential power analysis (DPA) [21], and correlation power analysis (CPA) [22] are three fundamental techniques of power-based SCA attacks. The CPA hacking technique requires the least power traces to extract the secret-key and it has been considered as the most powerful power-based SCA. In this paper, the CPA-based Hacking technique was adopted as a case study in order to indicate the main procedure that can threat the AES cryptocore security.

A. CPA based Hacking Theory

The goal of CPA-based hacking technique is to accurately model the power consumption of the cryptographic circuit under attack in order to find correlation between characteristics of real power consumption traces and a predicted power trace. Therefore, choosing the accurately power model enable hackers to predict correctly the secret key by obtaining highest level of correlation.

1) The CMOS device power consumption model: For the cryptographic platforms based on the SOC design, the CMOS technology still the principal hardware solution due to its various advantages. The total power consumption of a CMOS device (P_{total}) is composed of two components: the static power (P_{static}) and the dynamic power ($P_{dynamic}$) [21]. P_{static} is the result of the transistors leakage current and depends on the circuit design. Hence, $P_{dynamic}$ is consumed when switching occurs. Indeed, if a CMOS cell changes from 0 to 1 or from 1 to 0, switching happens in transistors and Pdynamic is consumed. Therefore it depends on the manipulated data and the operation being done.

$$P_{dynamic} = P_{0\rightarrow 1}.C_{L.}\ f.\ V_{dd} \quad (7)$$

where C_L denotes the gate load capacitance, f denotes the clock frequency, VDD is the supply voltage and P0→1 the probability of a 0→1 output transition. As shown in (7), at a given time, the dynamic power dissipation depends upon the number of bit switching from one position to another [23][22]. Power consumption-based SCA uses a leakage model in order to define a relationship between the device power consumption and the secret information employed.

Various power models are proposed to estimate the power consumption of device under attack when processing the target data. The most well-known models are the Hamming distance and Hamming weight models.

a) The Hamming Weight Model

The Hamming weight model (HW) is the most basic power consumption model [21]. This model computes, in a data word, the number of bits set to 1. Considering multiple bits at a time, it is important to understand that the power consumption is, exclusively, based on the numbers of bits that are at logical 1 and not the number those bits are meant to represent. [25]. So, the predicted power consumption PW in an n-bit microprocessor is computed simply by:

$$PW= a*HW(D)+b= a * \sum_{j=0}^{n-1} d_{j} +b \quad (8)$$

where dj = 0 or 1 is the bit values of the binary data D ($D=\sum$) handled by the cryptographic device under attack. is a scaling factor between the power consumed and the Hamming weight. and b is a term for everything like static power dissipation, the variation from one clock cycle to another, and time dependent components.

b) The Hamming Distance Model

The Hamming Distance model (HD) was proposed by Brier et al in [22], where the leakage is assumed to depend on switching activity in CMOS device. This model supposes that the power consumption in the target circuit correlates to the bits number changing from one state to another. Indeed, to estimate the device power consumption, the hacker uses the HD model and count number of 1→0 and 0→1 transitions that occur in a register or bus of a cryptographic device when it changes from one state to the next state. The consumptions for a bit switching from 0 to 1 or from 1 to 0 are further assumed to be same. Let R the reference state for a data word from which the bits are switched and D the current state manipulated by the target device. The power consumed PW is described by the mathematical equation for the hamming distance model as follows:

$$PW=a*HD(D)+b = a*HW(R \oplus D) + b \quad (9)$$

where $HW(R \oplus D)$ is the number of flipping bits from binary data *R* to *D*.

2) Pearson correlation coefficient: To evaluate the correlation between the estimated power consumption and the real power trace, the Pearson coefficient, $\rho_{W,PW}$ is considered as an efficient way. This correlation coefficient calculates the correlation between estimated power consumption PW of target data D and the equivalent real power traces measured W during processing the target cryptographic operation. $\rho_{W,PW}$ is described as follows:

$$\rho_{W,PW} = \frac{Cov(PW,W)}{\sigma_{PW}*\sigma_{W}} \quad (10)$$

where Cov denotes the covariance between PW and W, and and are standard deviation for PW and W respectively. When manipulated, the selected data D must depend on the desired secret key and the correlation coefficient is adopted as a distinguisher. Therefore the hacker predicts the unknown key and calculates the correlation coefficient $\rho_{W,HD}$ for every key candidate. The values will respect the inequality $0 \leq |\rho_{W,HD}| \leq 1$ and the right key assumption is supposed to indicate the biggest value.

B. CPA based Hacking Practice

In this section, we demonstrate the efficiency of the power analysis-based side-channel attacks on AES-128 engine implemented on FPGA (see Fig. 4). The steps involved in a successful CPA-hacking technique are:

1) Attack point selection step: In this step, a hacker chooses the attack point which can be a register or some function manipulating an intermediate result of the algorithm. This point must depend on both the known variable (e.g. the output of S-box) and secret keys K. In this case, the 10th round encryption is attack because the latter has been isolated from the other rounds and have relatively clear power signals [26]. Then, we calculate the original secret key, K0, going backwards since the AES Key Schedule is invertible. Fig. 5 shows the selected intermediate node D denoted as the output of Subbytes transformation and the reference point R defined as the corresponding Ciphertext. The AES-128 is used as a case study but this power-hacking technique can be applied to AES-192 and AES-256.

2) Power Assumption step: This step consists of predicting the target device power consumption with certain leakage model to estimate the dynamic power consumption reflecting secret-data moving and manipulated operations. As explained in previous section, power models present the correlation between the power consumed of the cryptographic CMOS device and data processed by this device at the same time. Indeed, bus value switching or registers value switching from 0 to1 or from 1 to 0 consumes some energy amount to achieve the transition. Hence, by counting the bits transition number at a given time, the hacker may predict the device under attack power consumption.

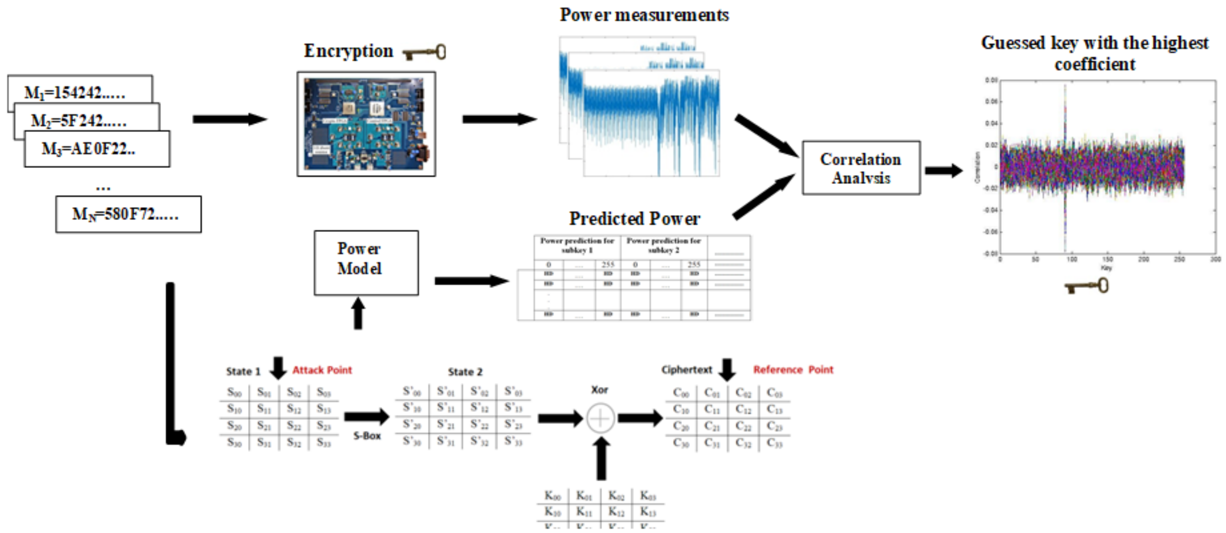


Fig. 4. Power Side-Channel Attack on 128-AES.

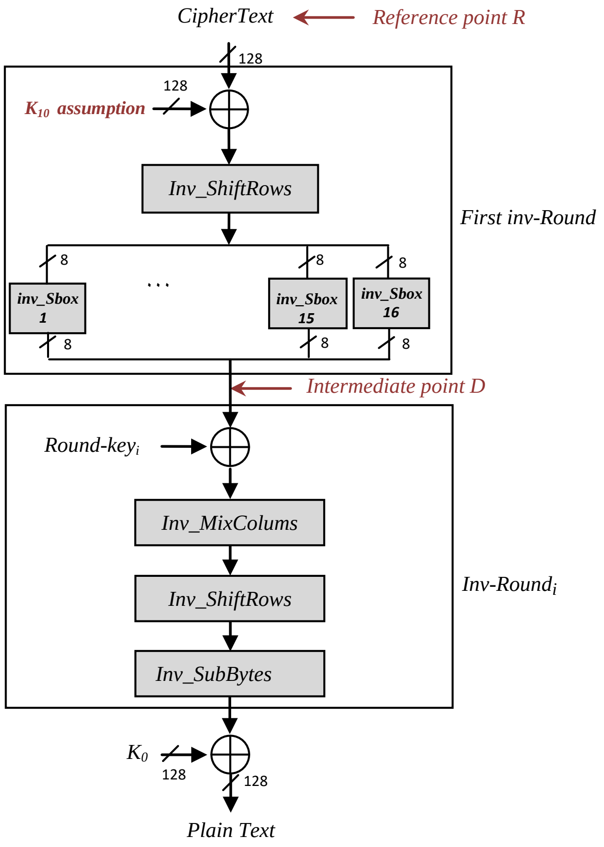


Fig. 5. The Selected Node for CPA Attack.

In this step, we have adapted the hamming distance model to predict the power consumption of the last AES-128 round encryption. The AES-128 decryption is an inv_round based encryption algorithm that process 128-bits data blocks as 16 bytes using 128-bits cipher keys. Each inv_rounds manipulates 128-bits round keys (K1 to K10) calculated from the original AES key, K_0 [27]. Indeed, this secret round-key is Xor-ed with the previous inv_round output, followed by an Inverse-ShiftRows transformation and Inverse-SubBytes transformation. The Inverse SubBytes operation divides the resulting 128-bits into 16 bytes and passing each through a Substitution S-Box. The S-box block takes 8-bit as input and produces 8 bite as output. Therefore, predicting one byte of the considered key is simple to calculate. For N Ciphertexts (N=20 000 in our case study) we predicts a subkey (The number of sub-key guess is limited to 28 assumptions: 256) and we calculate *HD(D)* predicted power consumption of the selected point *D* by the hamming distance model.

This step is repeated for 16 S-box outputs. So, we obtain a predicted power matrix P of size N x 256 x 16 as shown in Table I. *HD* value can be 1,2,3 or 8.

3) Measuring Power consumption step: The common setup for all Power-based side-channel attacks uses a PC in order to send known plaintexts to the target cryptographic circuit, trigger a device and save its power measurements traces. Therefore the hacker must obtain a matrix with a data pair of same Ciphertext used in the Power assumption step and their corresponding power measurements. The power measurements traces are normalized using pre-amplifier and gathered by oscilloscope during the AES encryption process. In this work, power measurements were performed by the "DPA contest v2" competition from the COMELEC Telecom department. The platform used to perform the power measurements acquisition is the Xilinx FPGA based Side-channel Attack Standard Evaluation Board (SASEBO)[28].

4) Correlation analysis step: This step evaluates the correlation between the predicted power and the power measurements using the Pearson coefficient. In this CPA statistical step, the measured power traces, denoted W, are compared to the predicted power consumption, denoted PW, for each 256 sub-key guesses. The correlation coefficient ρ_{WPw} is applied as follows:

TABLE I. PREDICTED POWER MATRIX P

Power prediction for subkey 1			**Power prediction for subkey 2**			**……………**	**Power prediction for subkey 16**		
0	…	255	0	…	255	……………	0	…	255
HD	…	**HD**	**HD**	…	**HD**	**……………**	**HD**	…	**HD**
HD	…	**HD**	**HD**	…	**HD**	**……………**	**HD**	…	**HD**
. . .									
HD	…	**HD**	**HD**	…	**HD**	**……………**	**HD**	…	**HD**

$$\rho_{W,Pw} = \frac{E(W,Pw)-E(W)*E(Pw)}{\sqrt{V(W)V(Pw)}}, \qquad (11)$$

If the sub-key assumption is correct, we expect that only one value, corresponding to the correct sub-key prediction, leads to a high correlation coefficient. The experimental result with only 3000 power measurements is shown in Fig. 6. As illustrated, the correlation power traces do not reveal the correct secret-key. Indeed there is no high correlation value in the obtained trace. The same correlation coefficient is calculated for the 256 sub-key assumption using 20000 power measurements. As indicated in Fig. 7, a unique correlation value, corresponding to the correct sub-key assumption, have high correlation value. Besides, the correct key assumption regularly stands out with a notable difference leading to a sure verdict of a successful attack. Fig. 9 corresponds to the correlation coefficient of correct sub-key assumption of. Hence, Fig. 8 presents the correlation value when the key guess is incorrect. This trace proves that there is no correlation between the predicted trace P and the corresponding measured trace W.

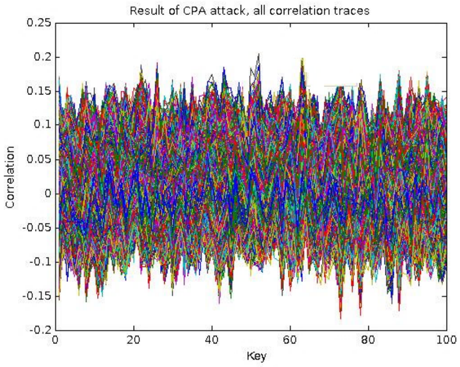


Fig. 6. Failed CPA using 3000 Inputs.

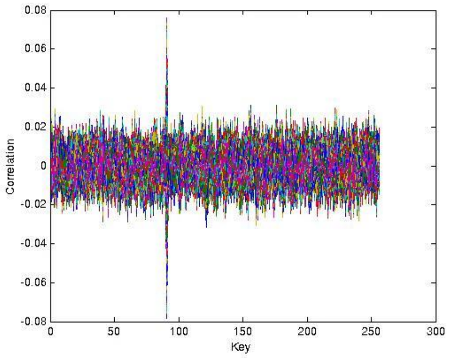


Fig. 7. Successful CPA using 20000 Inputs.

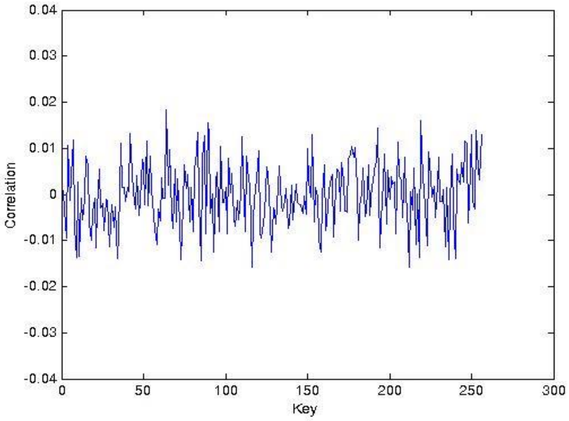


Fig. 8. Correlation Coefficient of an in Correct Sub-Key Assumption.

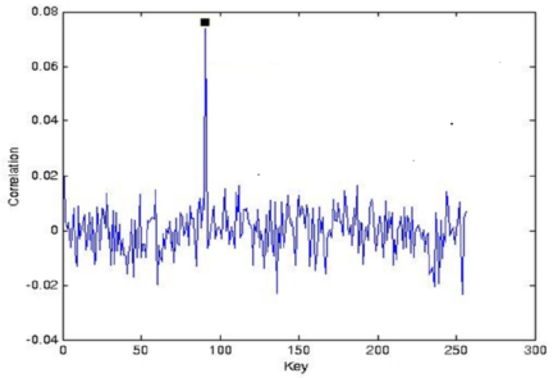


Fig. 9. Correlation Coefficient of a Correct Sub-Key Assumption.

In Fig. 10 the correlation coefficient for all the sub-key assumption, in terms of the number of power measurement was presented. This correlation trace presents the correlation coefficient between the measured power consumption and the predicted power consumption for various numbers of traces.

In fact, we remark that the correct sub-key assumption (plotted in black) can be distinguished by approximately 6000 power traces. This obtained result proves that CPA attack is an efficient power side-channel attack technique to extract the secret key.

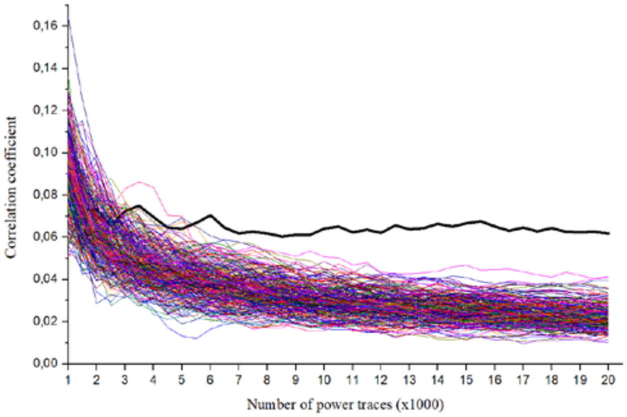


Fig. 10. Correlation Coefficient for all the Sub-Key Assumption.

IV. The Proposed Fault and Side-Channel Resistant 32-Bit AES

Power-Channel Analysis and Fault Attacks have seen a rise in popularity these last years, due to their practical methods to reveal the sensitive data from most cryptosystems which supposed to be cryptanalytically secure.

Our goal is to protect the cryptographic application against such attacks by detecting injected faults in the cryptocore and artificially introducing a noise in order to enhance the attack difficulty and reduce the probability of successful attacks. This section presents the proposed AES hardwarebased countermeasure to resist both Fault injection and Power sidechannel attacks. Fig. 11 shows the 32-bits AES block used as a case study.

The proposed design takes four 32-bit columns for the input data, one by one, executes them independently, and at the end, produces four 32-bit output columns.

A. Proposed Fault Detection Scheme for the AES

To secure the AES implementation against fault injection attacks, we incorporate fault-resilient techniques into the considered cryptographic hardware. Various DFA countermeasure schemes have been proposed to secure AES implementations, which are based on some sort of redundancy: information redundancy, hardware redundancy, or time redundancy in order to detect injected faults.

In this section we present an efficient combined fault detection based countermeasure that applies time and information redundancy to secure the considered AES cryptocore.

In this section we present an efficient combined fault detection based countermeasure that applies time and information redundancy to secure the considered AES cryptocore. The proposed Fault resilient AES implementation uses error detection code based on the (5, 4) CRC [29] to protect linear transformations and temporal redundancy approach for the Subbyte nonlinear transformation.

In fact, the information and hardware redundancy techniques induce more hardware overhead which can degrade cryptographic devices performances [30]. On the other hand The SubBytes operation is the most important non-linear operation in the AES, it occupies 70% of the AES Round area and 60% of the Key_Generator area.

Fig. 12 shows the temporal redundancy-based countermeasure to protect the Subbyte implementation.

As shown in Fig. 11, using same inputs, the Subbytes transformation is calculated twice with the same S-box hardware block and at the end, the result of the main calculation is compared with the result of the recalculation.

The fault detection scheme checks whether the indicator Sbox_FLAG equals zero or not. This Sbox_flag is calculated using a parity tree of exclusive-or gates.

Fig. 13 shows the information redundancy based countermeasure for linear transformations (ShiftRow, Mixcolum and AddRoundKey) with concurrent error detection. The ShiftRow shifts the bytes in each state row by a certain position without changing the parity from its input to its output. Correspondingly, there is no parity modification from the inputs of MixColumn to its outputs.

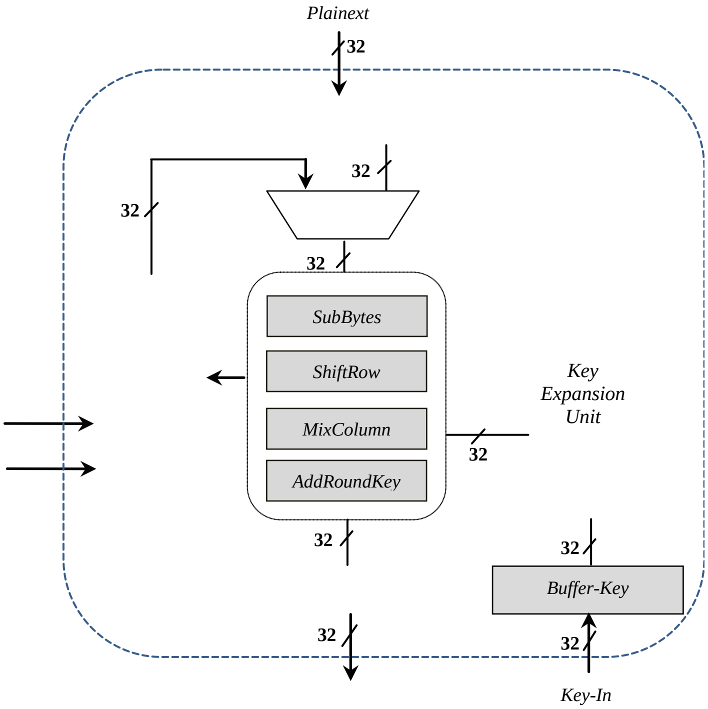


Fig. 11. Proposed 32-Bit Data-Path.

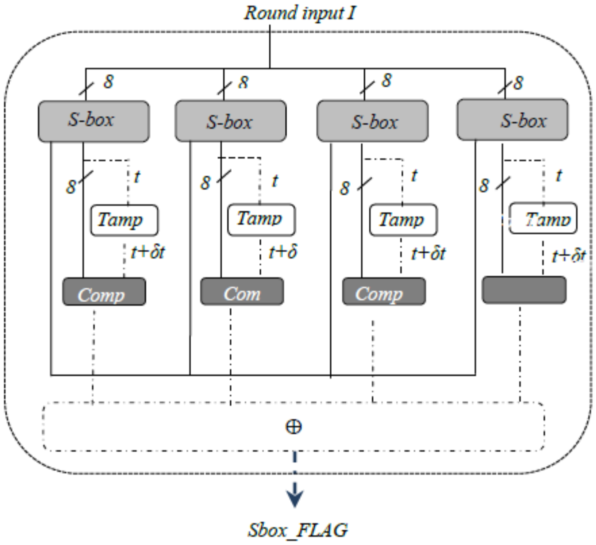


Fig. 12. SubByte Block with Error Check.

To detect injected fault in the Shiftrow and the Mixcolumn operations, we apply the information redundancy technique by using redundant information to protect these transformation.

Let SB(x) the Subbytes output and SR(x) the Shiftrow output as shown in Fig.13, where SB(x) = sb_0+sb_1x +sb_2 x^2 +sb_3 x^3 and SR(x) = sr_0 +sr_1 x +sr_2 x^2 +sr_3 x^3, {sb$i$, sr_i} ∈ GF(2^8). P_{SB} is the Shiftrow's input parity obtained by (12).

$$P_{SB}=P(SB)= \sum_{i=0}^{3} sb_i \tag{12}$$

where sb_i ∈ GF (2^8) is fault detection approach checks whether the flag Shiftrow_fg , obtained by (13), equals zero or not.

$$Shiftrow\_fg = P_{SB} \oplus \sum_{i=0}^{3} sr_i \tag{13}$$

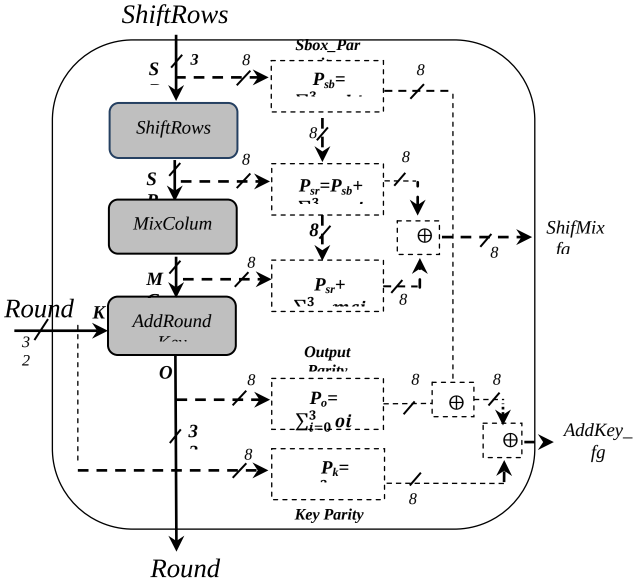


Fig. 13. Concurrent Error Detection Bloc for Linear Operations.

To protect Mixcolum operation, we use the same information redundancy based technique and we compute the flag mixcolumn_fg. The produced Flag will be XOR-red with the ShiftRow_fg in order to produce (ShifMix_fg).

At the Round output, the Shiftrow‗s input parity Psb will not be modified by the Mixcolum and the ShiftRow operations. But it will be changed by the AddRoundkey operation. So, to secure the AddRoundKey operation against fault injection, the key‗s parity PK must be calculated and xor-ed with the parity Psb. The obtained result will be XOR-ed with PO, the output‗s parity round, to produce the *AddKey_fg* flag. This DFA-countermeasure can be used to detect injected faults during the encryption process and produces Flags in order to interrupt the AES process. (see Fig.13).

B. Proposed SCA-Countermeasure for AES Cryptocore

The threats from Power attacks and Fault attacks challenge the integrity and security of cryptosystems. Various countermeasures for these attacks have been extensively studied in the existing literatures.

The author in [31] shows the impact of the countermeasure for one type of attack on the efficiency of another type of attack has been well explored.

Their experimental results show that the parity check code based fault detection technique makes CPA attack more difficult to retrieve the key than the original AES implementation. Based on this study, our Fault attack-resistant AES will affect the key retrieve speed of the CPA attack.

To more improve our fault resilient AES implementation, the hiding technique aims at lowering the Signal Noise Ratio (SNR) during the cryptographic operation by either adding more sources of noise or lessening the strength of the signal power trace that relates to the cryptocore operations. This makes CPA attack much harder as the data leaked has also to be correlated with external interfered key used inside the secured core.

Fig. 14 shows the Power analysis countermeasure for the AES cryptocore. This technique will be applied to the combined Fault resistant architecture presented in previous section. As shown in this figure, a parallel noise was incorporated into the AES cryptocore. The noise generating circuit obfuscates the AES cryptocore power traces by a power trace signal correlated with the plaintext and an interfering random key K_{interf}.

The AES cryptocore execute the AddRoundKey operation using the plaintext and the secret key K. Simultaneously, the noise generating circuit performs the same operation with the same plaintext but with the interfering keyK_{interf}. Two similar Subbytes transformation will takes the two AddRoundKey outputs as input and produces two signals S and S_{interf}.

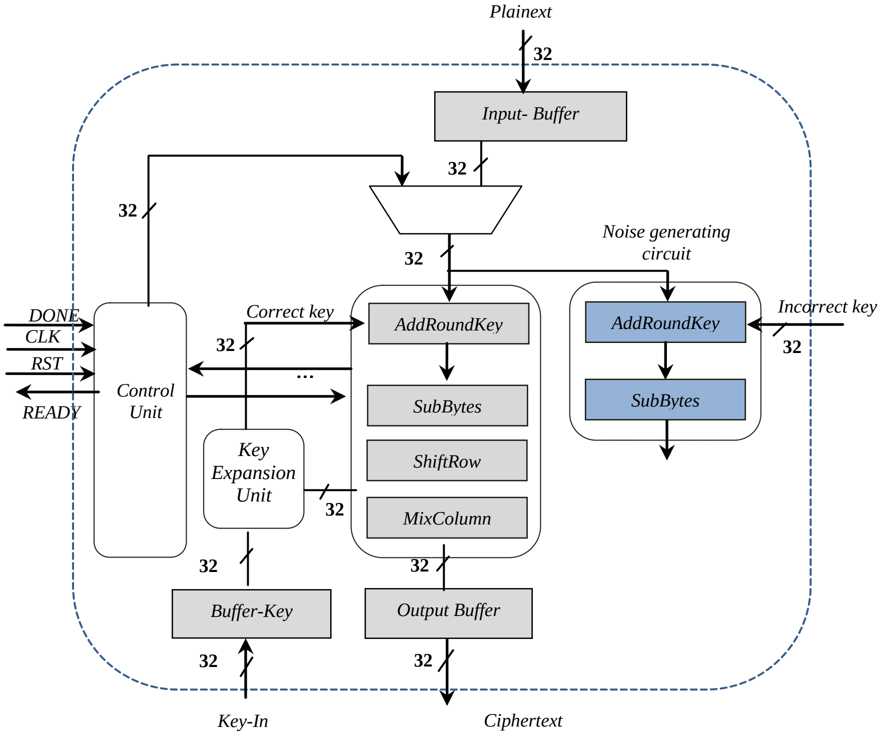


Fig. 14. Power Countermeasure for 32-Bits AES.

This proposed noise injection technique obfuscates the global AES cryptocore power trace by decreasing the correlation and bond between the secret values and the leaked information. In fact, the AES cryptocore power consumption correlates to plaintext and the secret keys couple (K, Kinterf). Furthermore, this added noise cannot be removed by statistical differential method; therefore, even if the power consumption curves was moved precisely and the Sboxes corresponding key successfully recovered, the hacker will still end up in failure because of the interference of the generated noise. This noise injection based countermeasure technique was experimentally proved in [32].

V. Implementation Results and Comparison

The proposed SCA/DFA Countermeasures for the AES design is practically examined using a Xilinx FPGA device, while the FI resistance is evaluated using the extensive fault simulations.We synthesized our implementation with the Xilinx ISE using the XC5VFX70t FPGA platform. The results and comparison with similar reported works are presented in Table II.

The AES-encryption implementation without the proposed countermeasure takes 445 slices for 296.43 MHz. The FPGA implementation result shows that our secured AES-encryption design occupies 14 % more area and 13% less throughputs compared to the original AES-cryptocore.

It can be seen from Table II, our proposed design has the minimum area overhead compared to [33], [34], [35] and [37].

TABLE II. FPGA Implementation of the Fault Resilient AES: Result and Comparison

	Area Overhead (%)	**Time Overhead (%)**	**FC (%) Single-bit**	**FC (%) Random-bit**
Our secured AES	**14**	**-13.5**	**100**	**100**
[33]	**43.33**	**-7.91**	**67.70**	**-**
[34]	**26.9**	**≈ 0**	**100**	**99.996**
[35]	**81**	**≈ - 0.17**	**100**	**93.75**
[36]	**2.3**	**-50**	**100**	**75**
[37]	**14.45**	**-18.71**	**85.958**	**98.54**

The temporal redundancy countermeasure presented in [36] add 2.3% overhead in terms of added hardware overhead and presents, approximately, four times degradation in terms of throughput overhead compared to our proposed countermeasures. Compared to [37] and [33], our secured AES design has the minimum area overhead and time overhead. These results prove that our proposed circuit is relevant to be arranged in many security domains such as embedded services routers, smartcards and emerging technologies using IoTs.

In order to evaluate the fault coverage of our protected AES cryptocore, fault-simulations are performed using the VHDL language. Two type of fault are used in the considered fault coverage simulation: Single-bit faults and Random-bit fault. For the single bit-fault type, we consider that a single-bit fault is inserted into 1-bit in random locations at random clock cycles of random rounds. On the other hand, random-bit fault type considers that faults are injected with random faulty bit number at random locations of random rounds. Fault simulations are performed over 1 000 000 times. As shown in the Table II, our fault simulations show that for single bit and random-bit faults, our protected AES-cryptocore have error coverage of 100 %.

Comparing our design to other similar countermeasures, our obtained fault coverage has the highest protection. Although our secured AES needs more resource overhead compared to some designs, it allows an excellent trade-off between fault-attack coverage, implementation area and throughput, which are relevant to secure embedded systems with resource constraint. In our future work, our power-baser cannel attack must be enhanced using new approaches based on deep-learning models.

VI. Conclusion

Recently, cryptographic embedded platforms used for trusted execution environment have been proven to be vulnerable to the power SCA and fault attacks. In this paper, we present a detailed fault and power based Hacking techniques to demonstrate how the fault injection or power analysis can be exploited to reveal the AES secret key. In the proposed case study, we explain the principal techniques that can threat the security of the AES design by using DFA and CPA attacks. This study was conducted in order to propose an adequate low-overhead hardware countermeasure that secures critical 32-bit AES cypto-core against both fault injection and power-based side-channel attacks. The proposed countermeasure gathers a combined fault resistance approach using parity testing for linear operations and time redundancy for the non-linear SubBytes operation with an artificially introduced noise provided by a correlated power noise block. The proposed countermeasure can be used for the encryption and decryption designs in order to enhance attack difficulty and reduce the probability of successful attacks. The proposed combined countermeasure has low overhead and achieves a 100% fault coverage during the considered AES process.

References

[1] A. Singh, M. Kar, S. K. Mathew, A. Rajan, V. De, and S. Mukhopadhyay, “Improved Power/EM Side-Channel Attack Resistance of 128-Bit AES Engines With Random Fast Voltage Dithering,” IEEE J. Solid-State Circuits, vol. 54, no. 2, pp. 569–583, Feb. 2019, doi: 10.1109/JSSC.2018.2875112.

[2] D. Das and S. Sen, “Electromagnetic and Power Side-Channel Analysis: Advanced Attacks and Low-Overhead Generic Countermeasures through White-Box Approach,” Cryptography, vol. 4, no. 4, p. 30, Oct. 2020, doi: 10.3390/cryptography4040030.

[3] C. Reinbrecht, A. Susin, L. Bossuet, G. Sigl, and J. Sepúlveda, “Timing attack on NoC-based systems: Prime+Probe attack and NoC-based protection,” Microprocess. Microsyst., vol. 52, pp. 556–565, 2017, doi: https://doi.org/10.1016/j.micpro.2016.12.010.

[4] Eng. Mustafa M. Shiple, Prof. Dr. Iman S. Ashour and Prof. Dr. Abdelhady A. Ammar, “Attacking Misaligned Power Tracks Using Fourth-Order Cumulant” International Journal of Advanced Computer Science and Applications (IJACSA), 4(12), 2013. http://dx.doi.org/10.14569/IJACSA.2013.041202

[5] H. S. Lim, J. H. Lee, and D. G. Han, “Novel fault injection attack without artificial trigger,” Appl. Sci., vol. 10, no. 11, 2020, doi: 10.3390/app10113849.

[6] R. Wang, X. Meng, Y. Li, and J. Wang, “Towards Optimized DFA Attacks on AES under Multibyte Random Fault Model,” Secur. Commun. Networks, vol. 2018, pp. 1–9, Aug. 2018, doi: 10.1155/2018/2870475.

[7] S. Patranabis, J. Breier, D. Mukhopadhyay, and S. Bhasin, “Side-Channel Assisted Fault Analysis,” 2018, pp. 59–77.

[8] S. Patranabis, J. Breier, D. Mukhopadhyay, and S. Bhasin, “One Plus One is More than Two: A Practical Combination of Power and Fault Analysis Attacks on PRESENT and PRESENT-Like Block Ciphers,” in 2017 Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC), Sep. 2017, pp. 25–32, doi: 10.1109/FDTC.2017.11.

[9] D. D. Hwang et al., “AES-Based Security Coprocessor IC in 0.18um CMOS With Resistance to Differential Power Analysis Side-Channel Attacks,” IEEE J. Solid-State Circuits, vol. 41, no. 4, pp. 781–792, Apr. 2006, doi: 10.1109/JSSC.2006.870913.

[10] M.-L. Akkar, R. Bevan, P. Dischamp, and D. Moyart, “Power Analysis, What Is Now Possible...,” 2000, pp. 489–502.

[11] A. Poschmann, A. Moradi, K. Khoo, C.-W. Lim, H. Wang, and S. Ling, “Side-Channel Resistant Crypto for Less than 2,300 GE,” J. Cryptol., vol. 24, no. 2, pp. 322–345, Apr. 2011, doi: 10.1007/s00145-010-9086-6.

[12] C. Tokunaga and D. Blaauw, “Securing Encryption Systems With a Switched Capacitor Current Equalizer,” IEEE J. Solid-State Circuits, vol. 45, no. 1, pp. 23–31, 2010, doi: 10.1109/JSSC.2009.2034081.

[13] M. Kar, A. Singh, S. K. Mathew, A. Rajan, V. De, and S. Mukhopadhyay, “Reducing Power Side-Channel Information Leakage of AES Engines Using Fully Integrated Inductive Voltage Regulator,” IEEE J. Solid-State Circuits, vol. 53, no. 8, pp. 2399–2414, Aug. 2018, doi: 10.1109/JSSC.2018.2822691.

[14] A. Singh et al., “Enhanced Power and Electromagnetic SCA Resistance of Encryption Engines via a Security-Aware Integrated All-Digital LDO,” IEEE J. Solid-State Circuits, vol. 55, no. 2, pp. 478–493, Feb. 2020, doi: 10.1109/JSSC.2019.2945944.

[15] N. Benhadjyoussef, M. Karmani, and M. MacHhout, “The Secured AES designs against Fault Injection Attacks: A comparative Study,” 2020, doi: 10.1109/ATSIP49331.2020.9231942.

[16] C. Giraud, “DFA on AES BT - Advanced Encryption Standard – AES,” 2005, pp. 27–41.

[17] N. Liao, X. Cui, K. Liao, T. Wang, D. Yu, and X. Cui, “Improving DFA attacks on AES with unknown and random faults,” Sci. China Inf. Sci., vol. 60, no. 4, p. 42401, 2016, doi: 10.1007/s11432-016-0071-7.

[18] Y. Liu, X. Cui, J. Cao, and X. Zhang, “A hybrid fault model for differential fault attack on AES,” in 2017 IEEE 12th International Conference on ASIC (ASICON), 2017, pp. 784–787, doi: 10.1109/ASICON.2017.8252593.

[19] H. Lim, J. Lee, and D.-G. Han, “Novel Fault Injection Attack without Artificial Trigger,” Appl. Sci., vol. 10, no. 11, p. 3849, Jun. 2020, doi: 10.3390/app10113849.

[20] N. Benhadjyoussef, M. Karmani, and H. Mestiri, “Power Analysis for Smartcard’s Authentication-Protocol,” 2019, doi: 10.1109/ASET.2019.8870994.

[21] P. Kocher, J. Jaffe, and B. Jun, “Differential Power Analysis BT - Advances in Cryptology — CRYPTO’ 99,” 1999, pp. 388–397.

[22] E. Brier, C. Clavier, and F. Olivier, “Correlation Power Analysis with a Leakage Model BT - Cryptographic Hardware and Embedded Systems - CHES 2004,” 2004, pp. 16–29.

[23] O.-X. Standaert, E. Peeters, G. Rouvroy, and J.-J. Quisquater, “An Overview of Power Analysis Attacks Against Field Programmable Gate Arrays,” Proc. IEEE, vol. 94, no. 2, pp. 383–394, Feb. 2006, doi: 10.1109/JPROC.2005.862437.

[24] H. Mestiri, N. Benhadjyoussef, M. MacHhout, and R. Tourki, “An FPGA implementation of the AES with fault detection countermeasure,” 2013, doi: 10.1109/CoDIT.2013.6689555.

[25] M. Randolph and W. Diehl, “Power Side-Channel Attack Analysis: A Review of 20 Years of Study for the Layman,” Cryptography, vol. 4, no. 2, p. 15, May 2020, doi: 10.3390/cryptography4020015.

[26] H. Liu, G. Qian, S. Goto, and Y. Tsunoo, “Correlation Power Analysis Based on Switching Glitch Model BT - Information Security Applications,” 2011, pp. 191–205.

[27] 2001. Advanced encryption standard (AES). Natl. Inst. Stand. Technol. 8– Fips-197, “Fips-197, 2001. Advanced encryption standard (AES). Natl. Inst. Stand. Technol. 8– 12,” 2011, [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf.

[28] “DPA Contest v2,” 2009/2010. dpacontest.org /v2.

[29] H. Yen and B.-F. Wu, “Simple error detection methods for hardware implementation of Advanced Encryption Standard,” IEEE Trans. Comput., vol. 55, no. 6, pp. 720–731, 2006, doi: 10.1109/TC.2006.90.

[30] N. Benhadjyoussef, M. Karmani, M. Machhout, and B. Hamdi, "A Hybrid-Countermeasure based Fault-Resistant AES Implementation," J. Circuits, Syst. Comput., 2019, doi: 10.1142/S0218126620500449.

[31] H. Pahlevanzadeh, J. Dofe, and Q. Yu, "Assessing CPA resistance of AES with different fault tolerance mechanisms," in 2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC), 2016, pp. 661–666, doi: 10.1109/ASPDAC.2016.7428087.

[32] N. Kamoun, L. Bossuet, and A. Ghazel, "Correlated power noise generator as a low cost DPA countermeasures to secure hardware AES cipher," in 2009 3rd International Conference on Signals, Circuits and Systems (SCS), 2009, pp. 1–6, doi: 10.1109/ICSCS.2009.5412604.

[33] H. Mestiri, N. Benhadjyoussef, and M. Machhout, "Fault attacks resistant AES hardware implementation," 2019, doi: 10.1109/DTSS.2019.8914979.

[34] M. Mozaffari-Kermani and A. Reyhani-Masoleh, "Concurrent Structure-Independent Fault Detection Schemes for the Advanced Encryption Standard," IEEE Trans. Comput., vol. 59, no. 5, pp. 608–622, 2010, doi: 10.1109/TC.2010.33.

[35] J. Chu and M. Benaissa, "Error detecting AES using polynomial residue number systems," Microprocess. Microsyst., vol. 37, no. 2, pp. 228–234, 2013, doi: https://doi.org/10.1016/j.micpro.2012.05.010.

[36] J. Rajendran, H. Borad, S. Mantravadi, and R. Karri, "SLICED: Slide-based concurrent error detection technique for symmetric block ciphers," in 2010 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST), 2010, pp. 70–75, doi: 10.1109/HST.2010.5513109.

[37] H. Mestiri, F. Kahri, B. Bouallegue, and M. Machhout, "A high-speed AES design resistant to fault injection attacks," Microprocess. Microsyst., vol. 41, pp. 47–55, 2016, doi: https://doi.org/10.1016/j.micpro.2015.12.002.

Drone Security: Issues and Challenges

Rizwan Majeed[1], Nurul Azma Abdullah[2], Muhammad Faheem Mushtaq[3], Rafaqut Kazmi[4]
Faculty of Computer Science and Information Technology[1, 2]
Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja 86400, Johor, Malaysia
Department of Computer Science[3]
The Islamia University of Bahawalpur, 63100 Bahawalpur, Pakistan
Department of Software Engineering[4]
The Islamia University of Bahawalpur, 63100 Bahawalpur, Pakistan

***Abstract*—Recent advancements in drone technology are opening new opportunities and applications in various fields of life especially in the form of small drones. However, these advancements are also causing new challenges in terms of security, adaptability, and consistency. This research discusses the drone technology, area of usages, citizen multi-objective uses, drones security, protection, and secrecy apprehensions, drone current intimidations and susceptibilities, existing approaches for drone cyber-security methods, security threats to drones and data sources for current literature review. Small drones are proving to be a new opportunity for the civil and military industries. The small drones are suffering from architectural issues and the definition of security and safety issues. The rapid growth of the Internet of Things (IoT) opens new dimensions for drone technology but posing new threats as well. The tiny flying intelligent devices are challenging for the security and privacy of data. The design of these small drones is yet not matured to fulfill the domain requirements. The basic design issues also need security mechanisms, privacy mechanisms and data transformations.**

Keywords—Drone technology; security; internet of things; threats; privacy

I. Introduction

In this modern atomic world, drone technology is mainly used for military purposes and defensive areas. Drone technology is rapidly growing for defense purposes. These microdevices are flying in the air 200 feet above the ground. This range of height varies from device to device and purpose to purpose. This range can be in feet, meters, and kilometers. Flight time of these intelligent devices also varies from device to device [1][2]. The frequency variations and their properties are discussed in Table I.

TABLE I. Frequency Variations and their Properties

Parameters	**2GHz**	**5GHz**
Frequency band	Low speed	High Speed
Cost	Cheap	Costly
Range	Extended range	Undersized range
Effect of noise	Noisy	Less noisy
interference	Prone to interference	Less prone to interference
Physical barriers	Overcome physical barriers	Unable to overcome physical barriers.
Performances	Disturb Wi-Fi speed	Don't disturb Wi-Fi speed

A. Rules and Guidelines

Many countries have strict rules and regulations for drone usage in civilian areas [16-18]. The rules and regulations issued advice for drone owners to use drones under government license. Drones must not enter private legacies without the permission of the owner. In Lebanon, strict rules are followed for drone usage. No person can fly drones in private areas, without the permission of the owner. To fly drones in private areas, the owner of the drone must get the permission certification for this act. Without this, one can face serious prosecutions. Following are some rules which must be followed by each drone owner in the USA.

- A drone must not reach 400 feet in height.
- A drone must not enter sensitive areas i.e. aircraft and airfields.
- A drone must be operated with safety and care to prevent face examination.
- A drone must not enter civilian areas.
- A drone must not be operated in gatherings and functions

Such rules are strictly followed by drone owners for safe drone flights. Heavy fines and punishments can be faced by the owners of the drone if such rules are violated.

B. Drone Structural Design

A UAV structure comprises three components. These components include Unmanned Aircraft, Ground Controller, and Communication link [3,19]. An explanation of these components is given below.

- The Flight Controller is the CPU of a drone.
- Ground station controller provides communication between the craft and controller.
- Data communication link provides information communication between ground controller and drone.

In [3] drones are categorized depending on their distance from the ground controller. These categories are given below.

- Line of sight distance drones use radio waves for communication.

- Beyond line of sight distance, drones use satellite communications.

C. Drone Communications Categories

Drones are classified into different categories based on their communication styles. These categories are D2D, D2G, and D2N. This communication is described in Fig. 2.

1) Drone-to-drone: Drone to Drone communication is not yet legalized. Machine learning approaches can be used in such communication in a wireless environment. This is also known as peer-to-peer communication. Such communication is more prone to jamming and DoS attacks [20 – 23].

2) Drone-to-ground location: This type of communication is mainly standardized and based on specific protocols with 2 and 5 GHz frequencies. It can also be operated on Bluetooth and Wi-Fi. Many of such communications are not available in the public sector because of security and authentication issues. Such communication is more prone to eavesdropping and man-in-the-middle attacks as shown in Fig. 1.

3) Drone-to-network: Such communication allows network selection for control and transfer of information. Several Wi-Fi networks can be used at different frequencies in such types of communication.

4) Drone-to-satellite: GPS devices are involved to provide real-time communication in such types of drone communication. Drone communicates with the satellites for location measurement. This communication is more secure and safe as compared to other categories and is used in the military.

D. Unmanned Aerial Vehicles Types

All Unmanned Aerial Vehicles can be called drones. In this section, differences between drones, UAVs and UAS are described in detail. Fig. 7 explains this classification difference in detail.

1) Drones: Drones are remote-based air devices that can be used on earth and in ships for different purposes. Different types of drone categories are described below based on their technology.

a) Multi-Blade Drones: Such drones are based on vertical landing and flying mechanisms. Such drones usually based on fixed operations and are easier to be attacked by external devices. This communication is very accurate but has less mobility because of the huge consumption of energy.

b) Fixed-Blade Drones: These drones consume less energy and are most suitable as compared to multi-blade drones [33,34]. These drones are faster and can handle heavy payloads.

c) Hybrid-Blade Drones: These drones have fixed as well as moving wings. These drone devices are speedy as compared to other types.

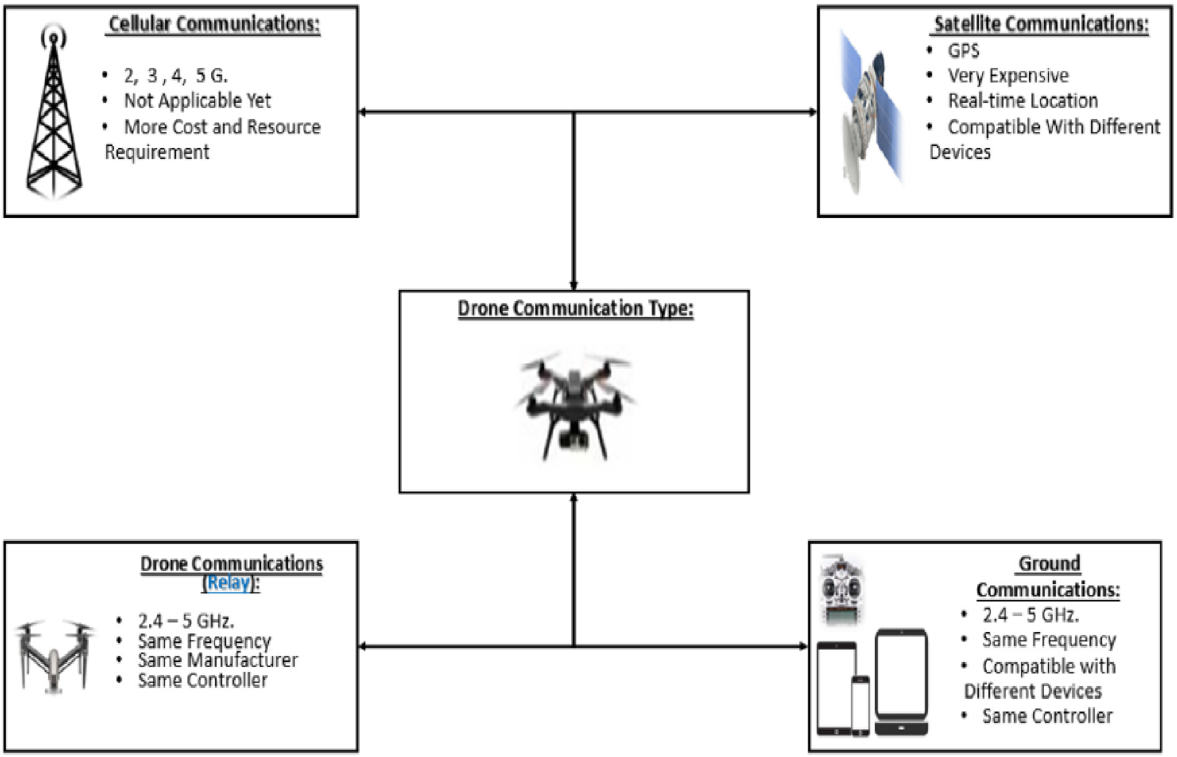


Fig. 1. Drone Communication.

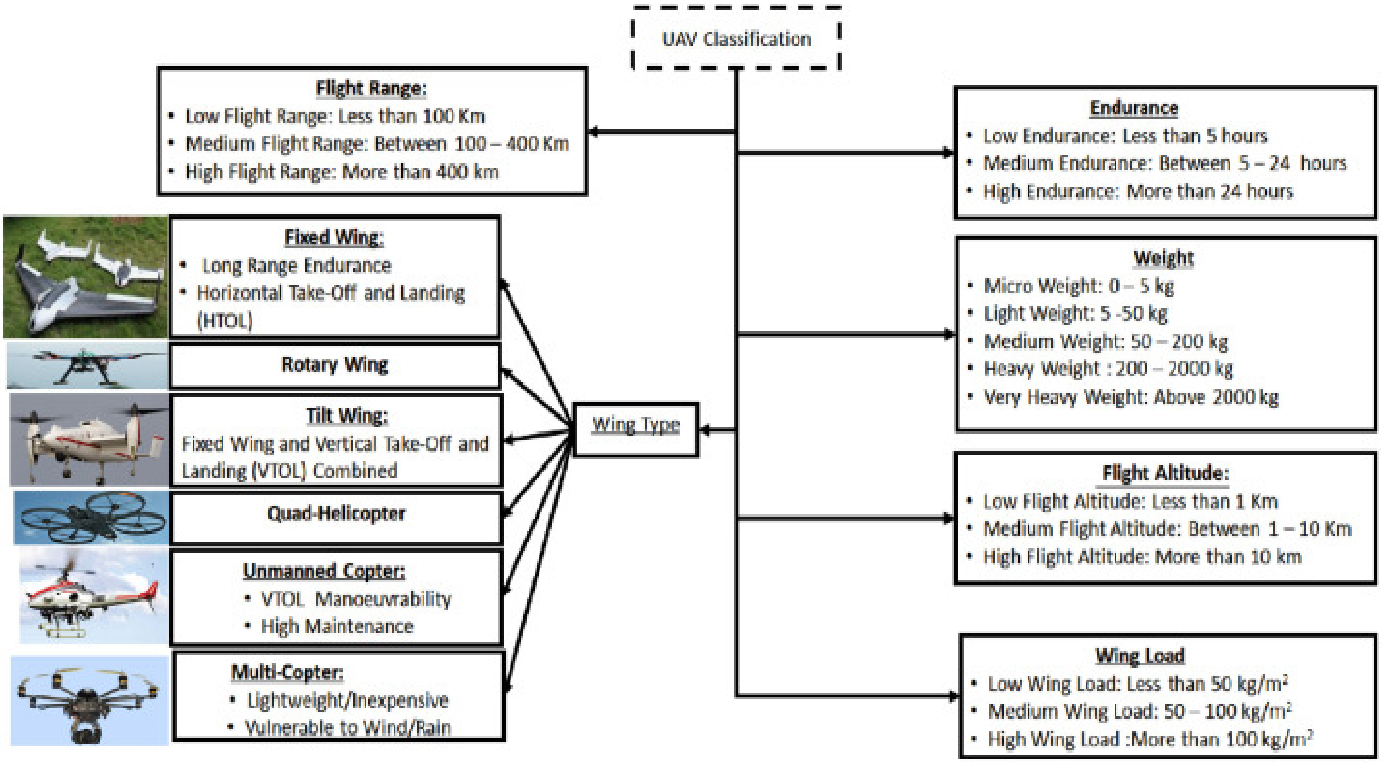

UAV Classification
Flight Range:
• Low Flight Range: Less than 100 Km
• Medium Flight Range: Between 100 – 400 Km
• High Flight Range: More than 400 km
Endurance
• Low Endurance: Less than 5 hours
• Medium Endurance: Between 5 – 24 hours
• High Endurance: More than 24 hours
Fixed Wing:
• Long Range Endurance
• Horizontal Take-Off and Landing (HTOL)
Rotary Wing
Tilt Wing:
Fixed Wing and Vertical Take-Off and Landing (VTOL) Combined
Quad-Helicopter
Unmanned Copter:
• VTOL Manoeuvrability
• High Maintenance
Multi-Copter:
• Lightweight/Inexpensive
• Vulnerable to Wind/Rain
Wing Type
Weight
• Micro Weight: 0 – 5 kg
• Light Weight: 5 -50 kg
• Medium Weight: 50 – 200 kg
• Heavy Weight : 200 – 2000 kg
• Very Heavy Weight: Above 2000 kg
Flight Altitude:
• Low Flight Altitude: Less than 1 Km
• Medium Flight Altitude: Between 1 – 10 Km
• High Flight Altitude: More than 10 km
Wing Load
• Low Wing Load: Less than 50 kg/m²
• Medium Wing Load: 50 – 100 kg/m²
• High Wing Load :More than 100 kg/m²

Fig. 2. UAV Classification.

2) UAVs: These devices consist of a controller which can be a mobile or a computer. These devices can be operated independently and are classified into the following categories. A UAV can fly remotely/autonomously using a controller, mobile phone, computer, or even a tablet [24]. They are characterized into the following types [25]:

- Remote Pilot Control UAVs are based on the controller which is handled by a human on the ground.
- Remote Supervised Control UAVs works independently and can be handled by a human if needed.
- Full Autonomous Control drones work independently without any human intervention.

3) UAS: UAS is a combination of UAV and drone. A UAV is also a UAS that is controlled by a controller [38].

4) RPA: Remotely Piloted Aircraft works independently for a long time without any intervention. It is used in complex missions and flights [37].

II. Areas of Usage

In the current era of technology, drone demand is increasing day by day. Drones help in delivering foods, items, and goods. Drones are also called flying riders. The main objective of using drones is to reach those areas where human access is limited. Drones usage is present in many areas of life [13][33]. This usage is discussed below in Fig. 3.

A. Citizen Multi-Objective Uses

Drone technology is used by civilians in many areas of daily life. These domains consist of life-saving, disaster management, etc. main uses of drones are discussed below.

a) Camerawork: Drones use is increasing in photography and video shooting in those areas where the human approach is difficult [34].

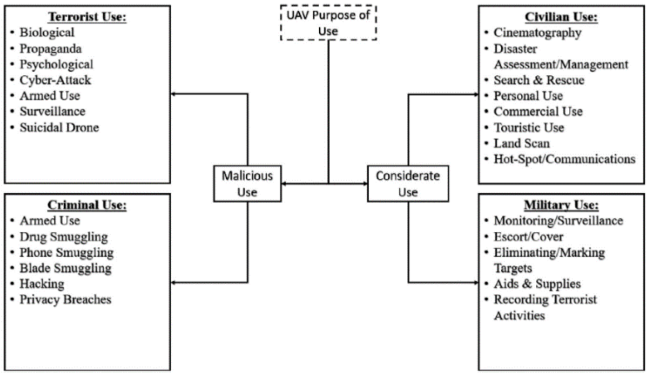


Fig. 3. Drone Uses in different Areas of Life.

b) Disaster Management: Disaster monitoring and controlling is also possible using UAVs and UAS since the 2005 major disaster. It helps in evaluating the situation of destruction and damage.

c) Exploration and Rescue: UAVs are also used to explore and save the lives of people and other livings beings.

d) Tourism: UAVs help in exploring different areas in tourism and traveling trips [35], which improves tourism revenue.

e) Commercials: Drones and UAVs are used to shoot commercials and Ads with HD cameras in minimum time. This saves time and cost for overall shooting.

f) Emergency Supervision: Extremist attacks and disasters can be handled and monitored using UAVs and drones [36].

g) Quick Reaction: Drones and UAVs as a first aid kit in many situations [37]. Drones were also used in the Covid-19 pandemic to provide disinfectant spray. Drones help in delivering foods and other items in this situation [38, 39 and 40].

h) Ecological Management: Drones provide different ecological measurement tasks including population calculation etc. it can also be used to calculate environmental parameters. It is also used to monitor crop parameters, air quality analysis, and weather measurement.

i) Subsurface/Oceanic Devotions: Underwater drones help in monitoring the underwater environment and operations [41]. Drones can provide live recording and monitoring.

III. Drones Security, Protection, and Secrecy Apprehensions

Drone technology provides several advantages and benefits for human beings. It helps in day-to-day activities as well as the military and monitoring of weather. However, several privacy and safety concerns are associated with their advantages. Security and privacy breaches should be addressed properly. Recording and image capturing by drones must be done keeping in mind the privacy and confidentiality of people's concerns [42]. For sanctuary and risk examination of drones, several studies are present in which risk associated with it is considered and discussed.

It is important to maintain secrecy, reliability, obtainability, verification, and non-denial possessions above message-passing networks are fulfilled. This can be achieved through AAA procedures and progressions:

- Authorization is achieved by providing access to the control unit of the drone/UAV.
- Verification can be obtained by using multi-level authentication using knowledge {specific key), identity verification and biometric verification.

Security threats are associated with drones which can be physical or cyber-attacks. It is important to limit drone usage in civilian areas and properties. The negative use of drones is also increasing day by day. This usage also creates problems for the citizens and civilians. Drone owners use Bluetooth or Wi-Fi channels to control their drones in restricted areas. This can leads to financial loss. Drones are used to breach Wi-Fi connections and Bluetooth signals. Such breach causes so many privacy and security concerns for peoples. Fig. 4 shows the important security threats to drones. Solutions to prevent these threats are also discussed. A summary of present and upcoming safety apprehensions is given below.

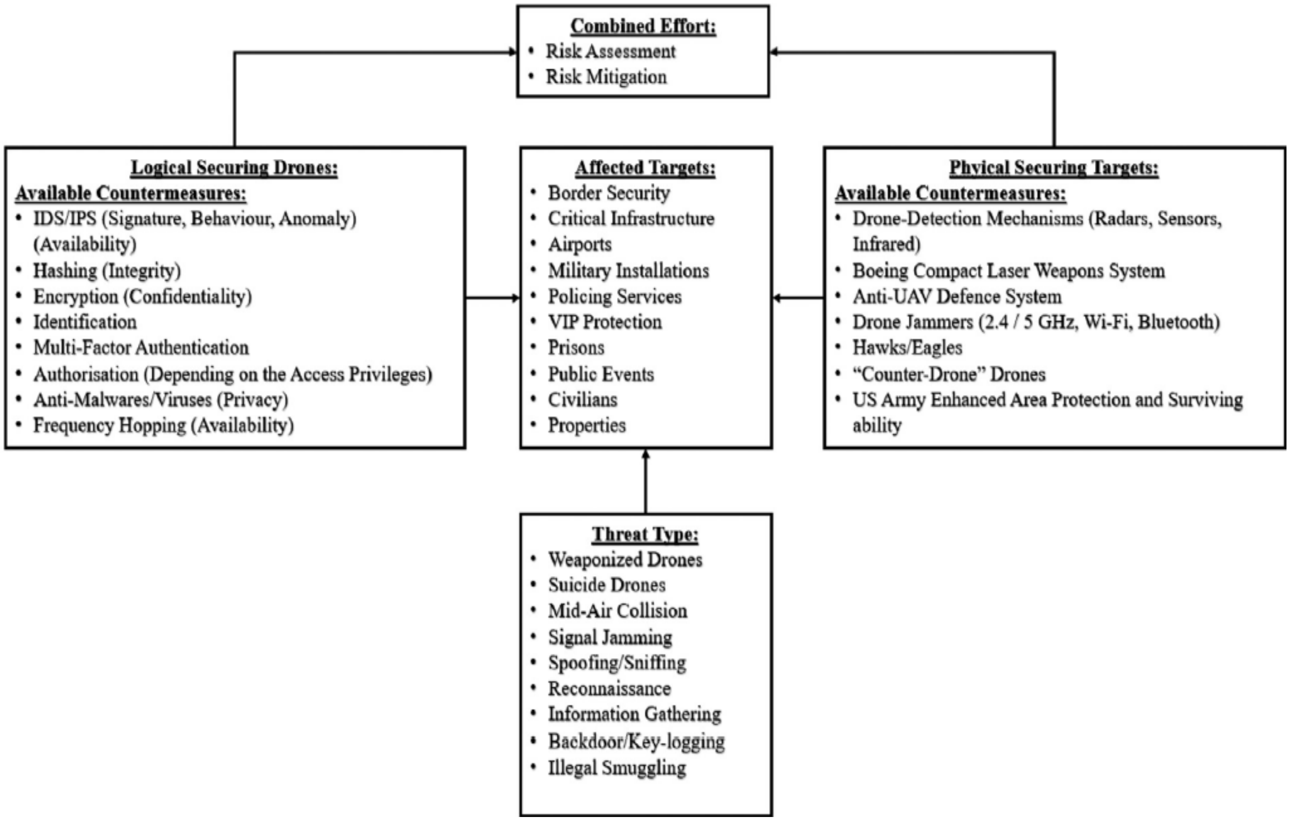

Combined Effort:
• Risk Assessment
• Risk Mitigation
Logical Securing Drones:
Available Countermeasures:
• IDS/IPS (Signature, Behaviour, Anomaly) (Availability)
• Hashing (Integrity)
• Encryption (Confidentiality)
• Identification
• Multi-Factor Authentication
• Authorisation (Depending on the Access Privileges)
• Anti-Malwares/Viruses (Privacy)
• Frequency Hopping (Availability)
Affected Targets:
• Border Security
• Critical Infrastructure
• Airports
• Military Installations
• Policing Services
• VIP Protection
• Prisons
• Public Events
• Civilians
• Properties
Phyical Securing Targets:
Available Countermeasures:
• Drone-Detection Mechanisms (Radars, Sensors, Infrared)
• Boeing Compact Laser Weapons System
• Anti-UAV Defence System
• Drone Jammers (2.4 / 5 GHz, Wi-Fi, Bluetooth)
• Hawks/Eagles
• "Counter-Drone" Drones
• US Army Enhanced Area Protection and Surviving ability
Threat Type:
• Weaponized Drones
• Suicide Drones
• Mid-Air Collision
• Signal Jamming
• Spoofing/Sniffing
• Reconnaissance
• Information Gathering
• Backdoor/Key-logging
• Illegal Smuggling

Fig. 4. Drone Threats Taxonomy.

A. Safety Apprehensions

A drone is tiny, lightweight, and has high mobility characteristics. It can be used to monitor criminal activities which are done at a high level of privacy. It can also be used by criminals to perform their illegal activities.

Drones can be equipped with dangerous objects to perform criminal acts. Such acts can create damage to civilians. It is a matter of concern to overcome such extremist activities for the wellbeing of peoples. Several terrorist groups can associate armed objects with a drone to perform their illegals activities.

Security doesn't constantly deliver protection. There are chances of damage done by the civilians in civilian areas which can result in financial loss [43]. The following list provides the safety concerns in detail.

- Minimum safety features in architecture can lead to control drone usage. This can result in damage and loss [44].
- Minimum mechanical and operative ethics include smashing avoidance techniques which can lead to drone's incapability to identify airliners [45].
- Absence of Administrative knowledge: especially it mainly occurs when people have less knowledge of safety features [46].

B. Confidentiality Apprehensions

Privacy is also the main factor to be considered for people. Drones must be kept out of those areas which are private. One must know the level of privacy of people before capturing or entering a private legacy. Three types of privacy threats are discussed below.

- Flying drones over someone's property is considered a major issue because of the risks associated with this act. Because such data can be used by scammers for negative purposes.
- Monitoring somebody's location must be avoided without their permission [47].
- Monitoring someone's acts and doings is also another unethical act which is also a matter of concern [48].

IV. Drones Current Intimidations & Susceptibilities

Several security threats are associated with drones and UAVs. Many design and architecture issues of drones cause such threats and vulnerabilities. Data and information protection must be addressed in drones to overcome these issues [49].

- Susceptible to Spoofing: several weak points are present which are related to architecture and control of UAVs and drones which may result in spoofing of information shared by the drone. Spoofing can be of several types [50, 51, 52, 53] . GPS spoofing is the easiest way to capture information and do modifications to it.
- Susceptible to Viruses: there are chances of malware and virus threats in the information shared by drones via a communication device which can be a mobile or a computer. This communication is also insecure because of poor wireless connection.
- Susceptible to data interruption: drones are more susceptible to data interruption and interception. Which may result in various threats associated with data vulnerability [54]. Such type of data interruption also causes malicious data insertion from drone to controller device [55].
- Susceptible to handling: drone devices are tiny objects which have building programming and control instructions, these instructions are more prone to manipulation which may change their operation and cause serious destructions [56].
- Susceptible to machine-driven problems:
- Many technical problems may occur during flight operations which may result in damage and destruction of data. Such failures may include unstable control connection [57,58].
- Susceptible to functioning problems: such issues include poor knowledge of control and command of a drone. Such lack of functioning information causes accidents and damage [59]. Many times, the drone also gets crushed in such incidents [60].
- Susceptible to environmental problems: Environmental factors include, wind speed, rain, heat, humidity result in problems for drone flight [61, 62].
- Susceptible to signal congestion: drones can also be spoofed using a signal loss mechanism which may transfer the control to a third party. Such situations also cause problems. This problem occurs by using device microcontrollers i.e. Arduino and raspberry-pi [63].

V. Existing Approaches for Drone Cyber-Security Methods

Major security methods for drone cybersecurity are categorized into the following types. Such classification is based on the attacker's aim and purpose. The following section discusses the existing approaches to prevent drone security issues.

A. Drone Network Security

Several security problems occur during drone flights and communication with a base station. To overcome such problems, and intrusion detection method was identified which can recognize illegal activities. Intrusion detection methods capture network flow and detect abnormal activities. Several intrusion detection methods are present which are used to analyze anomalies. These methods include rule-based, signature-based, and anomaly-based intrusion detection methods.

B. Drone Information Safety

Drone communication must be transformed into packet data to prevent load on the communication network. Such packaging allows safe communication to some extent. However such packaging also produces many problems. In one study [64], cipher security is discussed which can protect data from attackers.

C. Scientific Resolutions

Scientific methods are present in the drone domain which works by analyzing the network flow using forensic methods of monitoring. Such monitoring provides identification and detection of illegal access and capturing.

VI. Security Threats to Drones

Drone security has many layers and types according to its size, use, and controlling mechanism. The drone control in many cases using Wi-Fi with IEEE 802.11 communication protocol [65]. The typical designs of drones using the communication network are Wi-Fi networks with its ground stations. These networks are vulnerable to security breaches. Professional drones may be hijacked due to no proper encryption on their chips [66]. The second hijacking mechanism known to the research community was the man-in-middle attacks which is possible up to 2 KM only. The bottleneck so far emerged was with no encryption drones may be hijacked by individuals [67].

The novice trend in drone security is the Internet of Drones (IoD). The concept is equally popular in defense and industry drones [68]. The wide range of applications of IoD is in civilian and military drones simultaneously. The basic problem with drones was that they will design without security mechanisms in mind. There were fundamental security and privacy issues in drone technology regarding its design. The major issues identified in the domain of IoD security are privacy leakage, data confidentiality, data protection, data flexibility, data accessibility, and data encryption and decryption strategies [69].

In many studies in the last few years, various security and data privacy threats are identified by various researchers. The identified cybersecurity attacks are divided into four categories such as protocol-based attacks, sensors-based attacks, compromised components attacks, and jammers attacks. An account of such identified possible threats under these four cybersecurity attacks, found in the literature review, is given in Table II.

Table II shows that the majority of the available work in cybersecurity and data privacy of industrial drones is just the identification of possible threats. The solution to these threats does not exist. An attempt was made to use encryption to secure data transmission of a drone to a base station by using a Key Encryption algorithm for secure packet delivery [11].

Small drones and their usage are gaining the attention of the research community in recent years. These drones are popular because they have fewer wingspan and lightweight as well. The security and privacy of individuals and governments under threat due to these small drones as well [74]. Some other studies are also highlighting common challenges and threats to drone security [14-19].

Tian presented an efficient privacy-preserving authentication framework for edge-assisted internet of drones that was capable of ensuring the privacy of the drones Network [20]. Similarly, Hell presented a drone system for the security and surveillance purpose of a factory [21]. This system was capable of monitoring a defined area of a factory for security purposes. A similar application was also presented by Tosato in 2019, where he introduced an autonomous application of a swarm of drones for sensing industrial gas [22]. These types of drones are getting popular these days to monitor and surveillance the industrial area or an agriculture farm for the sake of security management.

TABLE II. Typical Cybersecurity and Data Privacy Threat to Smart Drones

	Common Cybersecurity Threats	**Threats Identified Citations**	**Countermeasures Citations**
Protocol-based Attacks	Security of Communication Link	[70], [71], [72], [73]	[72]
	Data Confidentiality Protection	[1]	
	Replay Arrack	[34], [35]	[36]
	Privacy Leakage	[1], [23]	
	De-authentication Attack	[8], [10],	
Sensors based Attacked	GPS Spoofing/Jamming Attack	[4], [37], [38]	[39], [40]
	Motion Sensors Spoofing	[41]	[42]
	UAV Spoofing/Jamming Attack	[4]	
Compromised Component	IoT Security Threats	[4], [5], [18]	
	Control/Data Interception	[4], [6],	
Jammers	Denial of Service	[4], [8], [10],	
	Stop Packet Delivery	[11]	[11]

The contributions to the identification of cybersecurity threats to drones are the healthy research area in the recent past. In 2016, Vattapparamban discussed a study on the application of drones for smart cities where he also discussed basic issues of Cybersecurity and privacy [23]. A few of these issues are also highlighted in Table I. A similar study to identify the possible security attacks along with the limitations of drone systems with a set of recommendations was also presented in 2020 by Yaacoub [24]. There are various studies [25], [26], and [27] that closely study the problems and challenges in drone security and its applications in business, commerce, etc. A few of them also proposes the use of blockchain for secure data delivery using 5G and IoT enables drones [27]. However, this system heavily relies on manual identification of threats type and intensity. There is a need for a smart and intelligent system for drone security that can analyze data of security breach attempts and attacks and adopt a proactive measure to ensure the security of the drones.

There are a few other survey-based studies [28], [29], [30], and [31] to identify common challenges and solutions to security threats and issues to drones used for industrial and commercial purposes. A few tried to address the problem of device authentication with lightweight authentication using key agreement [28] and key-enabling technique [30] for safe drones. The application of IoT drones in agriculture is also getting very common and recent contributions are discussed [9][32-33].

The problem of hacking and hijacking of drones and UAVs is a common threat to commercial drones that are specifically studied in [34], [35], [36], and [37]. The countermeasures problem of hacking and hijacking of drones and other UAV machines is proposed in [36] and [37]. Another common issue of drones is capturing of the drones [38] and UAV machines are GPS spoofing that also needs a secure and authentic solution. A few other studies on the hijacking of drones and intercepting control of a drone are also discussed in [39], [40], [41], and [42].

A. Gap Analysis of Drone Security using Machine Learning

There are a few basic types of machine learning techniques such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, deep learning, etc. During the literature survey, it was found in the recent past that, various attempts have been made to use machine learning solutions to handle cybersecurity attacks for mobile networks [45], wireless sensor networks [46], cloud computing [48], IoT systems [53][7][12], etc. An account of previous attempts to use machine learning for the security of various types of wireless networks is given in Table III.

However, there is no previous work is found to use machine-learning-based cybersecurity solutions for drone security threats. Additionally, we propose to use a machine-learning-based security solution with Blockchain to improve authentication and access control mechanisms in drone security.

In the detailed survey of literature from 2010 to 2020 in the area of drone and UAVs security, safety, and privacy concerns, more than 30 contributions were found in the form of research papers mainly published in IEEE and ACM journals. Majority of these papers highlight challenges and problems in the area of cybersecurity, device interception, data privacy, GPS spoofing, IoT spoofing, hijacking drones, and many similar cybersecurity threats in recent times. However, the majority of the literature is just highlighting identified key threats and issues to the security of drones but a majority of them are not giving solutions and countermeasures to handle security threats to drones. Only, in [27], blockchain is proposed for secure data delivery using 5G and IoT enables drones. However, this system heavily relies on manual identification of threats type and intensity. Other attempts are also relying on key-based authentication of devices that are not authentic of security specifically in the domain of IoT based drones. There is a clear research gap to make drones secure and safe from major cybersecurity threats to make drones useful for commercial and industrial purposes.

TABLE III. APPLICATION OF MACHINE LEARNING-BASED SOLUTION FOR CYBERSECURITY

Sr. No	Attacks	Security Technique	Machine Learning Solution
1	Jamming	Secure Offloading	Q-learning [45], [46] DQN [47]
2	Denial of Service	Secure Offloading	Neural Networks [48] Multivariate correlation analysis [49] Q-learning [50]
3	Malware	Access Control	Q/Dyna-Q/PDS [51] K-nearest neighbors [52] Random Forest [52]
4	Intrusion	Access Control	Naive Bayes [53] Support vector machine [53] Neural network [54] K-NN [55]
5	Spoofing	Authentication	SVM [56] DNN [57] Dyna-Q [58] Q-learning [58]
6	Traffic blockage	Authentication	Q-learning [59]

There is a need for a smart and intelligent system for drone security that can analyze data of security breach attempts and attacks and adopt a proactive measure to ensure the security of the drones. Previously, machine learning-based cybersecurity solutions are proposed for mobile networks, wireless sensor networks but not proposed for drone security. Additionally, we propose to use a machine-learning-based security solution with Blockchain to improve authentication and access control mechanisms in drone security.

VII. Data Source for Current Literature Review

Major data sources are explored for literature study of proposed system. Instead of searching through internets, direct databases are accessed to get relevant studies. However some other sources are also searched to get all possible relevant papers. These papers are filtered and only relevant papers are studied for this review. Manual selection is performed to exclude irrelevant papers. Table IV shows the data sources for current study.

A. Workflow of Current Literature Study

Literature study is carried out indifferent steps. Different actions are performed at each stage to get accurate and relevant proposed solutions. In the first step, literature review is performed with IoT, smart devices and security challenges for these devices. These security challenges are categorized and discussed accordingly. In next step, solutions for these threats are discussed and security issues are identified. Limitations for existing solutions are identified in these papers. A new solution is proposed to overcome these limitations. Fig. 5 shows the workflow of current literature study.

B. Year-Wise Publications

The following graph shows the year-wise publications list. In 2011 relevant studies were 9 which are considered in the proposed study shown in Fig. 6. This number increases each year. In 2020, relevant studies are increased which include IoT and drone security and solutions. Fig. 7 shows the graph in which drone security issues, solutions, threats, uses, frameworks, and algorithms are described for the current selection of papers.

TABLE IV. FOUNDATION OF DATA FOR CURRENT STUDY

Data Foundation	Web link
IEEE Xplore	http://ieeexplore.ieee.org/Xplore/
ACM Digital Library	http://dl.acm.org/
Springer	http://www.springerlink.com/
Elsevier	Elsevier http://www.elsevier.com/
Science Direct	https://www.sciencedirect.com/
Other Sources	Conferences, books and webpages

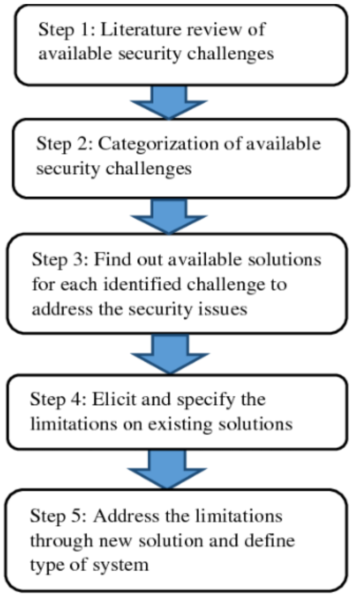

Step 1: Literature review of available security challenges
Step 2: Categorization of available security challenges
Step 3: Find out available solutions for each identified challenge to address the security issues
Step 4: Elicit and specify the limitations on existing solutions
Step 5: Address the limitations through new solution and define type of system

Fig. 5. Workflow for Literature Review.

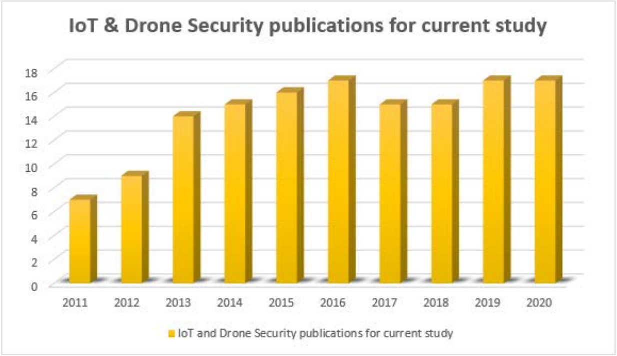


Fig. 6. IoT and Drone Security Publications for a Current Study Concerning the Year.

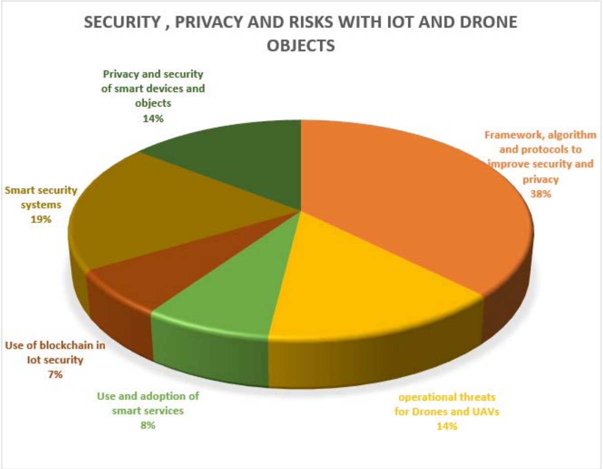


Fig. 7. Security, Privacy, and Risks Associated with IoT and Drone Objects.

VIII. Conclusion and Future Work

Now-a-days, IoT is the most widely used network among all networks. It is a network of communicating devices, which are interacting with each other. It is widely used for home automation, health care, and remote devices. Among all these applications, IoT is wonderfully used in civilian departments. It is widely utilized to interconnect medical and industrial devices. These devices provide easy and effective facilities to people. The machine learning based solutions for drone security is also used to handle cybersecurity attacks for mobile networks, wireless sensor networks, cloud computing, IoT systems, etc. IoT devices provide far-off fitness, manufacturing, and emergency alert method. This research can be further enhance using machine learning based security solution with Blockchain to improve authentication and access control mechanisms in drone security.

References

[1] E. Biddlecombe, “UN predicts 'internet of things',” July 6, 2009.

[2] D. Butler, “2020 computing: Everything, everywhere,” Nature, vol. 440, no. 7083, pp. 402-409, 2006.

[3] S. Dodson, “The net shapes up to get physical,” Guardian, 2008.

[4] N. Gershenfeld, R. Krikorian, and D. Cohen, “The Internet of things,” Scientific American, October, 2004.

[5] R. Lombreglia, “The Internet of things,” Boston Globe, pp. 76–83, 2005.

[6] A. Reinhardt, “A machine-to-machine Internet of things,” 2004.

[7] E. A. Kosmatos, N. D. Tselikas, and A. C. Boucouvalas, “Integrating RFIDs and smart objects into a unified Internet of things architecture,” Advances in Internet of Things, vol. 1, pp. 5-12, 2011.

[8] M. A. Shahid, U. Akram, M. M. A. Shahid, A. Samad, M. F. Mushtaq and R. Majeed, “A Systematic Approach Towards Compromising Remote Site HTTPS Traffic Using Open Source Tools” IEEE 23rd International Multitopic Conference (INMIC), 2020.

[9] H. Choi, W.-C. Lee, Y. Aafer, F. Fei, Z. Tu, X. Zhang, D. Xu, and X. Xinyan, “Detecting attacks against robotic vehicles: A control invariant approach,” in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 801–816, 2018.

[10] G. Biamino, “Semantic model for socially aware objects, Advances in Internet of things,” vol. 2, pp. 47-55, 2012.

[11] I. Ungurean, N. C. Gaitan, and V. G. Gaitan, “An IoT architecture for things from industrial environment,” in Proc. 10th Int. Comm., 2014, pp. 1–4.

[12] F. Li, M. Voegler, M. Claessens, and S. Dustdar, “Efficient and scalable IoT service delivery on cloud,” Proc. in IEEE 6th Int. Conf. Cloud, pp. 740–747, 2013.

[13] K.D. Atherton, “The faa says there will be 7 million drones flying over america by 2020”, Popular Sci., 2016.

[14] E. Vattapparamban, I. Güvenç, A. I. Yurekli, K. Akkaya, S. Ulua gac, “Drones for smart cities: issues in cybersecurity, privacy, and public safety”, IEEE International Wireless Communications and Mobile computing Conference (IWCMC), pp. 216–221, 2016.

[15] K. Dalamagkidis, K.P. Valavanis and L.A. Piegl, “Aviation history and unmanned flight”, On integrating unmanned aircraft systems into the national airspace system, Springer, pp. 11–42, 2012.

[16] M. Juul, Civil drones in the European union, Eur. Parliament. Res. Serv. (ed.). Eur. Union, 2015.

[17] R. Stopforth, “Drone licenses-neccesities and requirements”, II Ponte, Vol. 73, no. 1, 149–156 2017.

[18] V.S. Campos, “European union policies and civil drones”, Ethics and Civil Drones, Springer, Cham, pp. 35–41, 2018.

[19] D.M. Marshall, R.K. Barnhart, S.B. Hottman, E. Shappee and M.T. Most, “Introduction to unmanned aircraft systems”, CRC Press, 2016.

[20] J. Dinger, H. Hartenstein, "Defending the Sybil attack in P2P networks: taxonomy, challenges, and a proposal for self-registration," First Int. Conf. on Availability, Reliability and Security (ARES'06), 2006.

[21] M. F. Mushtaq, S. Jamel, and M. M. Deris, “Triangular Coordinate Extraction (TCE) for Hybrid Cubes,” J. Eng. Appl. Sci., vol. 12, no. 8, pp. 2164–2169, 2017.

[22] H. Rowaihy, W. Enck, P. McDaniel and T. La Porta, “Limiting sybil attacks in structured P2P networks, 26th IEEE International Conference on Computer Communications, pp. 2596–2600. 2007.

[23] N. Naoumov and K. Ross, “Exploiting P2P systems for DDoS attacks”, Proc. of the 1st international conference on Scalable information systems, ACM, 2006.

[24] J. Irizarry, M. Gheisari and B.N. Walker, “Usability assessment of drone technology as safety inspection tools”, J. Inf. Technol. Construct. (ITcon), vol. 17, no. 12, 194–212, 2012.

[25] R. Altawy and A.M. Youssef, “Security, privacy, and safety aspects of civilian drones: a survey”, ACM Trans. Cyber-Phys. Syst. Vol. 1, no. 2, 2017.

[26] F. Barfield, “Autonomous collision avoidance: the technical requirements”, Proceedings of the IEEE National Aerospace and Electronics Conference, pp. 808–813, 2002.

[27] R. Sharma and D. Ghose, “Collision avoidance between UAV clusters using swarm intelligence techniques”, Int. J. Syst. Sci., vol. 40, no. 5, pp. 521–538, 2009.

[28] A. Hamza, U. Akram, A. Samad, S. N. Khosa, R. Fatima and M. F. Mushtaq, “Unmaned Aerial Vehicles Threats and Defence Solutions”, IEEE 23rd International Multitopic Conference (INMIC), 2020.

[29] S. Ueno and T. Higuchi, “Collision avoidance law using information amount”, Numerical Analysis-Theory and Application, InTech, 2011.

[30] J. Israelsen, M. Beall, D. Bareiss, D. Stuart, E. Keeney and J. Berg, “Automatic collision avoidance for manually tele-operated unmanned aerial vehicles”, IEEE International Conference on Robotics and Automation (ICRA), pp. 6638–6643, 2014.

[31] E. Yanmaz, R. Kuschnig, M. Quaritsch, C. Bettstetter and B. Rinner, “On path planning strategies for networked unmanned aerial vehicles”, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 212–216, 2011.

[32] L. H. Hernandez, A. Tsourdos, H. S. Shin and A. Waldock, “Multi-objective UAV routing”, IEEE International Conference on Unmanned Aircraft Systems (ICUAS), pp. 534–542, 2014.

[33] E. Vattapparamban, I. Guvenc, A. I. Yurekli, K. Akkaya and S. Ulua gac, “Drones for smart cities: issues in cybersecurity, privacy, and public safety”, IEEE International Wireless Communications and Mobile computing Conference (IWCMC), pp. 216–221, 2016.

[34] A. Rango, A. Laliberte, C. Steele, J.E. Herrick, B. Bestelmeyer, T. Schmugge, A. Roanhorse and V. Jenkins, “Using unmanned aerial vehicles for rangelands: current applications and future potentials”, Environ. Pract. 8, 3, pp. 159–168, 2006.

[35] N. Jumaat, B. Ahmad and H.S. Dutsenwai, “Land cover change mapping using high resolution satellites and unmanned aerial vehicle”, IOP Conference Series: Earth and Environmental Science, 2018.

[36] J.-P.A. Yaacoub, M. Noura, H.N. Noura, O. Salman, E. Yaacoub, R. Couturier and A. Chehab, “Securing internet of medical things systems: limitations, issues and recommendations”, Future Generat. Comput. Syst., 2019.

[37] C.A. Thiels, J.M. Aho, S.P. Zietlow and D.H. Jenkins, “Use of unmanned aerial vehicles for medical product transport”, Air Med. J. 34 (2) (2015) 104–108.

[38] M. Lipsitch, D.L. Swerdlow and L. Finelli, “Defining the epidemiology of covid-19-studies needed”, N. Engl. J. Med., 2020.

[39] F. Jiang, L. Deng, L. Zhang, Y. Cai, C.W. Cheung and Z. Xia, “Review of the clinical characteristics of coronavirus disease 2019 (covid-19)”, J. Gen. Intern. Med., pp. 1–5, 2020.

[40] M.N.K. Boulos and E.M. Geraghty, “Geographical tracking and mapping of coronavirus disease covid-19/severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic and Associated Events around the World: How 21St Century GIS Technologies Are Supporting

the Global Fight against Outbreaks and Epidemics", Int J Health Geogr., 2020.

[41] L.K. Johnson, A.W. Dorn, S. Webb, S. Kreps, W. Krieger, E. Schwarz, S. Shpiro, P.F. Walsh and J.J. Wirtz , "An ins special forum: intelligence and drones/eyes in the sky for peacekeeping: the emergence of UAVs in un operations/the democratic deficit on drones/the german approach to drone warfare/pursuing peace: the strategic limits of drone warfare/seeing but unseen: intelligence drones in israel/drone paramilitary operations against suspected global terrorists: us and australian perspectives/the 'terminator conundrum' and the future of drone warfare", Int. Natl. Secur., vol. 32, no. 4, pp. 411–440, 2017.

[42] A. Cavoukian, "Privacy and drones: Unmanned aerial vehicles, Information and Privacy Commissioner of Ontario", Canada Ontario, 2012.

[43] R.L. Finn and D. Wright, "Unmanned aircraft systems: surveillance, ethics and privacy in civil applications", Comput. Law Secur., vol. 28, no. 2, pp. 184–194, 2012.

[44] H. Du and M.A. Heldeweg, "Responsible design of drones and drone services: Legal perspective synthetic report", 2017.

[45] K. Wackwitz and H. Boedecker, "Safety risk assessment for UAV operation, Drone Industry Insights", Safe Airspace Integration Project, Part One, Hamburg, Germany, 2015.

[46] E.B. Carr, "Unmanned aerial vehicles: examining the safety, security, privacy and regulatory issues of integration into us airspace", Natl. Centre Policy Anal. (NCPA), 2014.

[47] R.L. Finn, D. Wright and M. Friedewald, "Seven types of privacy", European data protection: coming of age, Springer, 2013.

[48] R. Clarke, "The regulation of civilian drones' impacts on behavioural privacy", Comput. Law Secur. Rev., vol. 30, no. 3, pp. 286–305, 2014.

[49] R. Majeed, N. A. Abdullah, I. Ashraf, Y. B. Zikria, M. F. Mushtaq and M. Umer, "An Intelligent, Secure, and Smart Home Automation System", Scientific Programming, 2020.

[50] Y. Zeng, R. Zhang and T.J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges", arXiv preprint arXiv:1602.03602, 2016.

[51] D. Rudinskas, Z. Goraj and J. Stankunas, "Security analysis of UAV radio communication system", Aviation, vol. 13, no. 4, pp. 116–121, 2009.

[52] A.J. Kerns, D.P. Shepard, J.A. Bhatti and T.E. Humphreys, "Unmanned aircraft capture and control via GPS spoofing", J. Field Rob., vol. 31, no. 4, 617–636, 2014.

[53] S. H. Seo, B. H. Lee, S. H. Im, and G. I. Jee, "Effect of spoofing on unmanned aerial vehicle using counterfeited GPS signal", J. Positioni. Navigat. Timing, vol. 4, no. 2, pp. 57–65, 2015.

[54] X. Lin, R. Wiren, S. Euler, A. Sadam, H. L. Maattanen, S.D. Muruganathan, S. Gao, Y. P .E. Wang, J. Kauppi and Z. Zou, "Mobile networks connected drones: field trials, simulations, and design insights", arXiv Preprint arXiv:1801.10508, 2018.

[55] A. Abdallah, M.Z. Ali, J. Misic and V.B. Misi, "Efficient security scheme for disaster surveillance UAV communication networks", Information, vol. 10, no. 2, 2019.

[56] P. Ramon Soria, R. Bevec, B. Arrue, A. Ude and A. Ollero, "Extracting objects for aerial manipulation on UAVs using low cost stereo sensors", Sensors, vol. 16, no. 5, 2016.

[57] S.J. Kim, G.J. Lim and J. Cho, "Drone flight scheduling under uncertainty on battery duration and air temperature", Comput. Ind. Eng., 291–302, 2018.

[58] C. M. Tseng, C.K. Chau, K. Elbassioni and M. Khonji, "Autonomous recharging and flight mission planning for battery-operated autonomous drones", arXiv preprint arXiv:1703.10049, 2017.

[59] H. Du and M.A. Heldeweg, "Responsible design of drones and drone services: Legal perspective synthetic report", 2017.

[60] M. F. Mushtaq, S. Jamel, K. M. Mohamad, S. K. A. Khalid, and M. M. Deris, "Key Generation Technique based on Triangular Coordinate Extraction for Hybrid Cubes," J. Telecommun. Electron. Comput. Eng., vol. 9, no. 3–4, pp. 195–200, 2017.

[61] M. Erdelj and E. Natalizio, "Drones, smartphones and sensors to face natural disasters", DroNet'18: Proceedings of the 4th ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications, pp. 75–86, 2018.

[62] P. Velagapudi, S. Owens, P. Scerri, M. Lewis and K. Sycara, "Environmental factors affecting situation awareness in unmanned aerial vehicles", AIAA Infotech. Aerospace Conference and AIAA Unmanned. Unlimited Conference, 2009.

[63] O. Westerlund and R. Asif, "Drone hacking with raspberry-pi 3 and wifi pineapple: Security and privacy threats for the internet-of-things", 2019 IEEE 1st International Conference on Unmanned Vehicle Systems-Oman (UVS), pp. 1–10, 2019.

[64] D. He, S. Chan and M. Guizani, "Drone-assisted public safety networks: the security aspect", IEEE Commun. Mag., vol. 55, no. 8, 218–223, 2017.

[65] A. Nayyar, B. L. Nguyen, and N. G. Nguyen. "The Internet of Drone Things (IoDT): Future Envision of Smart Drones", First International Conference on Sustainable Technologies for Computational Intelligence, Springer. pp. 563-580, 2020

[66] Z. Yin, Q. Song, G. Han, and M. Zhu, "Unmanned optical warning system for drones", Global Intelligence Industry Conference (GIIC 2018), 2018.

[67] R. Koslowski and M. Schulzke, "Drones along borders: border security UAVs in the United States and the European Union", International Studies Perspectives, vol. 19, pp. 305-324, 2018.

[68] M. O. Ozmen and A. A. Yavuz, "Dronecrypt-an efficient cryptographic framework for small aerial drones", in MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM), pp. 1-6, 2018.

[69] M. O. Ozmen, R. Behnia, and A. A. Yavuz, "IoD-Crypt: A Lightweight Cryptographic Framework for Internet of Drones", arXiv preprint arXiv:1904.06829, 2019.

[70] E. Bertino, "Data Security and Privacy in the IoT," in EDBT, pp. 1-3, 2016.

[71] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, "A Comprehensive Survey on the Cryptographic Encryption Algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 11, pp. 333–344, 2017.

[72] M. O. Ozmen, R. Behnia, and A. A. Yavuz, "IoD-Crypt: A Lightweight Cryptographic Framework for Internet of Drones", arXiv preprint arXiv:1904.06829, 2019.

[73] D. Hussain, M. A. Khan, S. Abbas, R. A. Naqvi, M. F. Mushtaq, A. Rehman and A. Nadeem, Enabling Smart Cities with Cognition Based Intelligent Route Decision in Vehicles Empowered with Deep Extreme Learning Machine, Computers, Materials & Continua, 2020.

[74] Z. Lv, "The security of Internet of drones", Computer Communications, vol. 148, pp. 208-214, 2019.

Blockchain Technology in Education System

A Survey Examining Potential Uses of Blockchain in Saudi Arabia Education

Afnan H. Alsaadi[1], Doaa M. Bamasoud[2]
Dept. of IS, College of Computing and Information Technology
University of Bisha, Bisha
Saudi Arabia

***Abstract*—The aim of this paper is to review the blockchain technology and its benefits in relations to education system. Blockchain technology is widely researched and highly evaluated and appraised for its unique infrastructure. In general, blockchain researched for its association with Bitcoin and cryptocurrency advantages. In this survey the plan is to conduct a full review of previous literatures focused on blockchain in education systems. Provide overall reviews of blockchain concepts and architecture behind the technology and to examine verification software that used by the technology to improve security and immutability. Brief discussion on the consensus algorithms and hashing function and how these operate and difference type of blockchain will be discussed. In this survey, the existing technology used in Saudi Arabia will be reviewed. In-depth research conducted for over 70 papers, of which 35 noted in this survey. Blockchain emerging promise a real time democracy and justice to all users over the world. Educational Industries said to revolutionize its communication system and accessibility and extend their market globally by widening their admissions and providing secure cost-effective, transparent and immutable communications across their educational platforms.**

Keywords*—Blockchain; certifications; authentication; decentralized-education; transparency; immutability; smart contracts; learning accessibility; fraud prevention; sustainability; ledger; consensus

I. Introduction

Electronic peer- to - peer without any third party is the solution to end double spending [1]. A digital signature is the part of the solution [1]. The authors in [1] invoked hashing by timestamped proof, “the network timestamps transactions by chain them into an ongoing chain of hashing-based proof forming a record cannot be changed”. Additionally, added “the longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest pool-of CPU power; keeping long chains to outpace attackers. Blockchain technology first emerged as cryptocurrency technology only, but its popularity increased as the technology addressed other classical problems for the centralized system.

Mainly described as a ledger for digital currency and used for cryptocurrency. This decentralized cryptographic system provided a platform for users to transfers money without necessarily relying on any centralized trusted establishment such as payment services or banks [2].

Blockchain technology becomes popular among researchers and technology developers and cooperation e.g., IBM, Sony and Samsung, immediately after disappearance of its creator [1]. The attention, however, advanced the research of blockchain and brings more critical analysis and examination. Blockchain is not a ledger, it is a journal. Carlson in [3] considered blockchain idea only buried into bitcoin currency and not addressed technology in-depth. Added that “the author has weighed the costs and benefits of transacting with virtual currencies, considered the sustainability of the virtual currency, and contemplated the application of existing regulatory schemes to virtual currency” [3].

This suggesting that blockchain structure is misunderstood and not clearly defined by the authors. However, this is also followed by another Blockchain critics, the authors in [4] suggesting that blockchain technology has a potential to diminish government’s authorities by end their role of supervising people’s money and subverting critical regulation by cutting out middleman such as large operations and corporations. A full analysis conducted by [5] analyzed the code of law as valuable contribution to blockchain. A critical analysis improves the quality of the research into blockchain technology. On another hand, the strong arguments indicted the usefulness of the technology to be taking seriously by researchers. In [4] the authors added the architectural features of blockchain are the center of the attentions. Blockchain designed in such a way is that it evades the effectiveness of regulation and enforcement [5]. The suggestion made by [4] quoted by [5] mainly focused on the novelty of blockchain as the main point of revolution, rather than using all characteristic structures of blockchain which are not revolution. The points below quoted directly from researcher [5] summarized the basis of blockchain structure suggested to underline both opportunities and threats in the discussion and the reviews:

1) *Blockchains are disintermediated* and transnational networks, often relying on open-source software protocols.

2) *Resilient and tamper-resistant*, due to their distributed nature, the consensus mechanisms employed, and the use of hashing.

3) *Blockchains are transparent*—in the sense that transaction data is authenticated and visible—and the data they contain is non-repudiable (due to the use of public-private key cryptography).

4) *Characterized by pseudonymity*—as they allow transacting parties to participate in the system without disclosing their identity.

5) Blockchains have incentives and cost structures, e.g., block rewards and mining fees that incentivize and compensate parties maintaining a blockchain-based network.

6) Blockchain are unique, and characteristic is the deployment of consensus mechanisms to coordinate social activity towards an agreement on the situation within the system.

7) Blockchains enable a specific type of "autonomy": they facilitate the execution of software code that is entirely independent of any one party.

(Global Survey for blockchain carried by Deloitte 2020).

Moving forward with blockchain, this technology despite its young age, it has been the center of attention among researchers, analysts, law assimilators and centralized system beneficiaries (middleman). A survey conducted by [6] linked the attention is due to the features presented by blockchain, such as decentralized nature, persistency, anonymity, and auditability properties. To be fully conversant with blockchain concept the followings must be clearly defined and understood [7]:

- Asymmetric key cryptography.
- Consensus algorithms.
- Hashing.
- Consensus.

Blockchain is part of the solution to critical problems of ownership, transparency and can be used in many digital communications as solution to our everyday arising challenges [8].

Existing technology provided education system with numerous advancements of learning environment and productivity, however the authors in [9] cited further area of improvements such smart learning track, where the blockchain technology predicted to play significant role.

A block is an entity (data structure) hold information that present past record and current transactions and keep permanent records of transactions. Each block created through process called mining and each new block connected with its predecessor with cryptographically secured reference shown in Fig. 1. The validation process called the hash function.

How hash function design in blockchain is rather critical process to the success of secured blockchain procedure. For a cryptographic hash function to be considered secure; it must represent a certain characteristics or properties, these features determined the hash function suitability for cryptocurrencies such as Bitcoin and other secured transactions. Thus, the cryptographic hash function is an integral part of the blockchain innovation. It is the main characteristic that gives security capabilities to the processed transactions and making it immutable. Hash function designed to protect data stored in the block [10] and then data in a blockchain distributed across the network of multiple nodes or computers. Thus, blocks data becomes very secure and impossible to be altered retroactively without having to alter the entire chain.

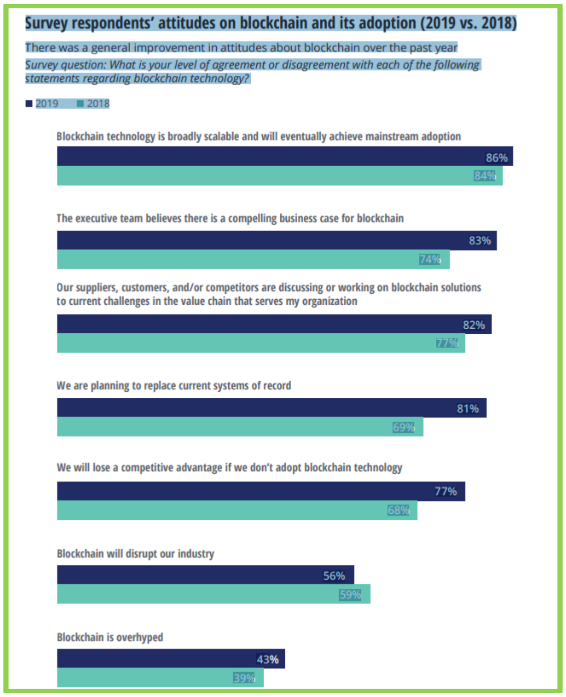


Fig. 1. Blockchain Adoption.

These authentications proofed by mathematical algorithms [11] which deem secure and protected from others. These can be achieved in many ways, but most commonly the use of encryptions and can only be decrypted by legitimate person.

Passing this process, the system integrity can be guaranteed [12] that information has not been modified by unauthorized party. Through this process, system recognizes manipulation and detects it and prevents data corruption. This is how the use of cryptographic hash functions in blockchain works.

There are two significant properties make hash function satisfying [8]:

- Should be fast to compute and strong, and.
- Minimize duplication and data cannot be extracted from its original hash.

The authors stated that: "There are three requirements for practical applications of a hash function for message authentication and digital signatures", the fourth requirement also known as pre-image resistance or one-way property.

A. Blockchain Types

As shown in Fig. 15, there are three different forms of Blockchain [6].

- Public: Everyone can check the transaction and verify it.
- Consortium: It refers to a node with authority that can be selected in advance, typically has business-to-business relationships, the data in blockchain can be open or private, and can be classified as partially decentralized like Hyperledger.

- Private: This applies to the fact that not every node in the blockchain will participate, with one or more restrictions. The interpretation of firm authority for data management access.

B. How Transaction Complete in Blockchain

The transaction complete securely between two parties without exchange of identification or third party using public key. Communications are transparent among all networks. Nodes validate transactions and group them in Block [10].

The created block identified by its hash: “a cryptographically unique value calculated on the contents of the block and includes a reference to the hash from the previous block, so that, blocks are linked” [10]. This chain of blocks is thus a record of transactions or a public accounting book (ledger), shared by all the nodes in the network.

Mining is one of the key concepts behind the Bitcoin protocol said [11]. All transaction collected and grouped into block that ready to be linked to its predecessors.

In order to add new block to blockchains, a signature linking the transaction to previous block must be elected [12]. This process involved finding nonce value which is required to satisfy the secure hashing algorithm SHA256. In this surveyor we only briefly explained the functions rather than diving into equations or calculation.

Blockchain technology provides a structural ledger database for storing transactional records known as blocks and links them with numerous databases known as the chain [13]. These immutable records are then signed cryptographically using a distributed consensus or validation protocol [14]. The following processing staging are essential tools to secure blockchain transaction:

1) Education system: technology strength and transformation: Its evidently education system has been evolving and transforming over the past centuries. Institutions and universities provide diverse methods of learning to reach out learners. One of these methods is the distance learning, which dated back since 1873 where first initiated by Professor Ana Eliot Ticknor and then followed by the University of Queensland in Australia in 1911 and the University of South Africa 1946. These humble initiations shaped the distance learning of today. The only difference between now and then was the methods of delivery. Traditionally, learning delivered on campus, certifications received in persons or by post, transcription and copy of credential kept with registrars, but this paradigm have been shifting gradually to online educations and communications: All communications currently handled online. From entry applications, payment system, and learning, handling coursework, and submitting assignment and communicating with tutors all completed through universities platforms.

The paradigm shifting in educations often appraised for its efficiency, increased productivity and improved learning outcomes. A research conducted by Bill& Melinda Foundation on positive distance learning and how technology advanced learned outcome.

The report conducted by [15] expressed positive outcome for learner at the time, it was a revolution concept to create such “learning environment that student in rural areas had expanded learning opportunities in a variety of subject.” [15]. The Impact of digital technology is countless and how this technology improves learning outcomes is incalculable.

A research conducted in relation to digitalization. “In terms of education this means we should create Knowledge that is accessible virtually with the focus on the student”. This means students are build and developed from all areas and these facilities aid their learning.

Rather those days where students she/he travel distances or encountering lock of access to documents or physical visit to library and time restrictions to information.

The authors in [16] added virtual access is achieved through Internet or Intranets, facilities such as “e-mail, web notices, discussion forums and video conferencing allow a student to access information without visiting the physical location of delivery”. A typical interactive e-learning system will have these characteristics and thus demonstrates the paradigm shift” [16].

2) Education system: technology weakness and volatility: Despite the enormous benefits of technology in education and the positive contribution achieved, technology comes with some downside risks and must be appropriately evaluate [17]. In relation to how technology apprised, and debates the author, highlighted that the impact of technology in education should appropriately addressed both strength and weakness. Unforeseen its potential risks will hinder the quality of the any review [17].

A recent public awareness publishing; highlighted these facts: Data published by [18] highlighted critical information into how criminal may damage universities infrastructure-based knowledge and steal significant data that can be harmful to countries’ economy.

“In 2018 researcher discovered logging pages for 300 fake websites and login pages for 76 universities across 14 countries”. “And between 2013 and 2017, accounts of more than 100,000 professors worldwide stolen, and led to the loss of more than 30 terabytes of academic data and intellectual property”. [18]. As we moving to cloud space and more online present via working groups or communicating online, there is urgency to seek more security and proof systems to convey information and communication through.

3) Education: system venerability (centralized cloud space): Cyberspace attacks can be permanently damaging to system and can be irreversible. The authors in [18] Highlighted reasons behind such attacks that could be and not limited to the followings:

- For financial gain, criminals can sale critical data and research information and breakthrough to hostile enemies or competitors to publish dedicated researcher findings and claim ownerships.

- Steal critical information on national infrastructures research for knowledge advancement, substage strategic planning and future projects.
- Malicious attackers can damage important files and documents and can navigate through to gain full access to intellectual properties.
- Talents at universities can be targeted and their activities monitories by infiltrator.
- Undermine national security and economic prosperity.

4) Education: blockchain the transformation promise: Blockchain Technology its innovative disruptive system with prominent potentials, fully authenticated and transparent, economically viable and with direct and indirect benefits to support sustainable educational system for all.

Recent survey carried by [19] on usability and impact of blockchain adoption including positive impact and barriers of use. Lack of knowledge of the technology and comparison with the existing system showed impactful results. The following Fig. 1 to 6 shows Survey results:

(Global Survey for blockchain Deloitte 2020).

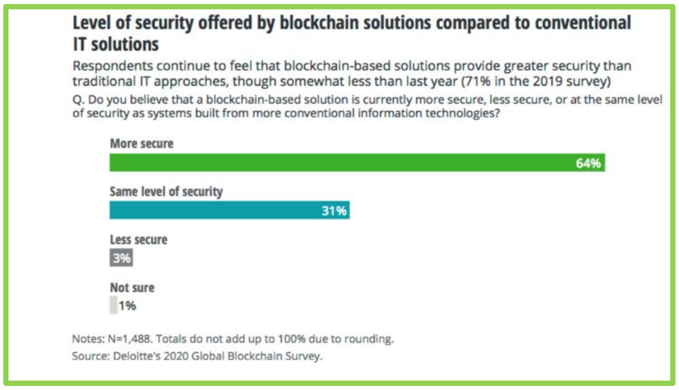


Fig. 2. The Level of Security vs IT Solution.

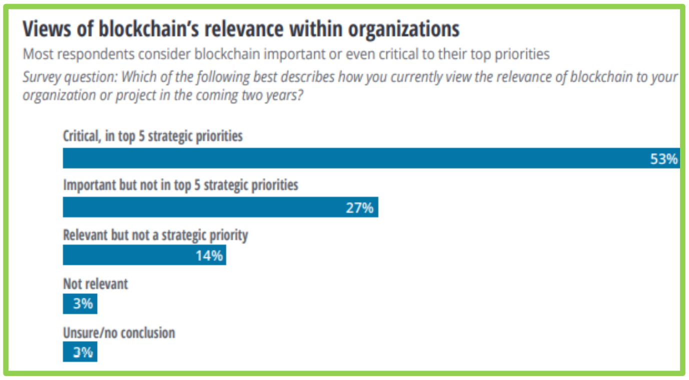


Fig. 3. Blockchain Significant Advantages over Existing System.

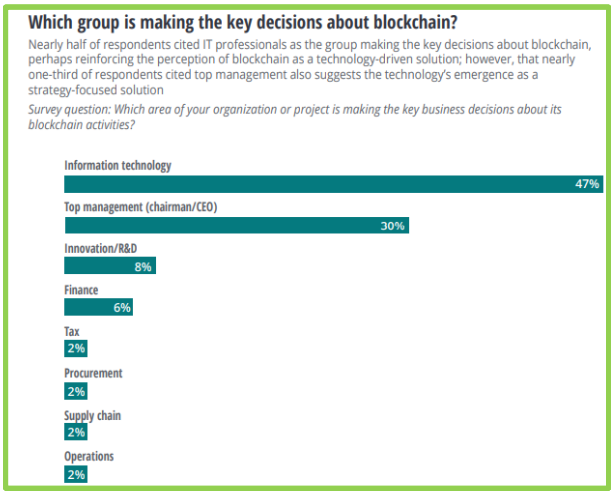


Fig. 4. Global Survey for Blockchain Carried by Deloitte 2020.

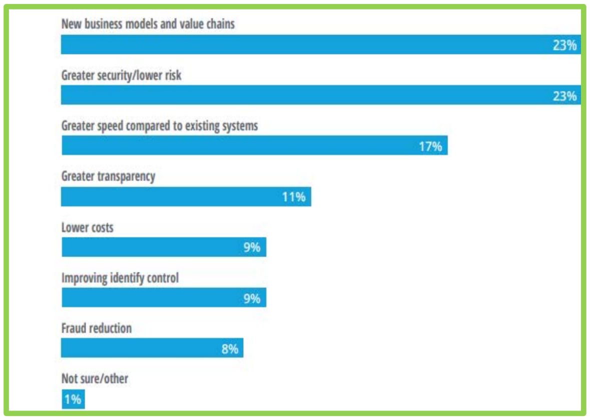


Fig. 5. Decision Maker on Organization.

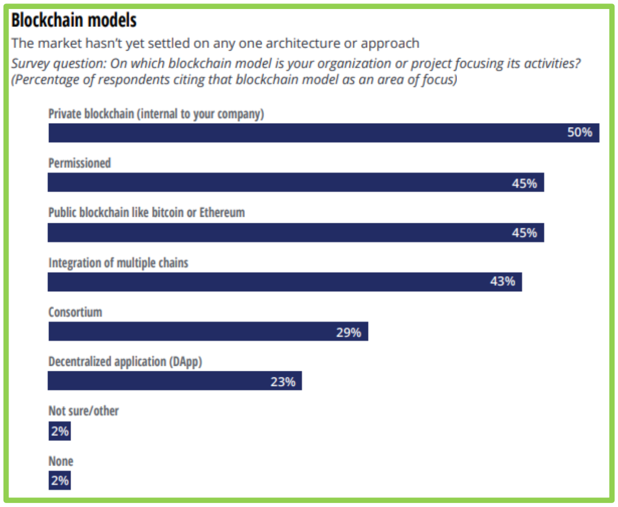


Fig. 6. FBlockchain Model.

5) Education: blockchain technology: The first case reviewed is a Cambridge, Massachusetts education model: “MIT has formed the Digital Credentials Consortium an international network of universities to develop a shared system for digital academic credentials [20]. Public ledger technology enables us to develop new tools, standards, and strategies to store and manage digital academic credentials. Certificates can be cryptographically signed and tamper-proof. They can represent or recognize many different types of achievements [21]. DCC was founded in 2018 by leading universities with expertise in the design and verification system.

Academia of DCC raised questions about the overall system of trusted verification and authentication of learning and credentials; and there was a clear concern about the way credential issuing and transcription managed, after consultation, consortium strongly agreed that there is an urgency to renew system. Benefits for learners:

- Maintain a compelling verifiable record of student lifelong learning achievements to share with employers in a trusted manner.
- A safe and secure digitally transferred credential with no extra fees.

Available secure stored data from multiple educational institutions and credential. A team from learning machine and MIT media lab creates “blockers” an open standard platform to issuing and verifying blockchain certificates. DCC highlighted future project planning [21], “we are exploring how recent advances in public key infrastructures, public ledgers, and blockchains can be used to rethink the way we recognize and transact with academic achievements”.

Digital Credential Consortium is a powerful organization with unique strength of intellectual talents. “The infrastructure for design and governance of academic credentials has been set a sight and credentials now transformed into tokens of social and human capital that can create new opportunities for participation in educational industry” [21].

“Real-life use case: More than 600 of 2018 MIT graduates chose to receive a digital version of their diplomas on Blockcerts’ blockchain. Consequently, the students' academic records will be stored forever, and future employers can immediately verify them” [20].

Organization and world’s fined universities and institutions are benefiting from blockchain technology and building searcher on how to completely adopt this technology. In this survey, several universities will be reviewed:

a) Curriculum: Smart curriculum design is one of the most researched subjects these days. Mainly associated with blockchain technology and the usefulness of its features. In [22], the QualiChain platform that will offer blockchain-enabled verification of education and other credentials as well as data analytics and decision support for process optimization [22].

Permanence of blockchain record, smart contracts these features seen as attractive tool for education commissioners [23]. These features make blockchain technology-based products or services significantly different from previous internet-based commercial developments [23]. The authors in [23] addressed the lack of knowledge about the social advantages and potential for blockchain in education by stakeholders. The authors highlighted the benefits blockchain in comparison with existing system: shown in Fig. 7.

- Self-sovereignty, i.e. for users to identify themselves while at the same time maintaining control over the storage and management of their personal data.
- Trust, i.e. for a technical infrastructure that gives people enough confidence in its operations to carry through with transactions such as payments or the issue of certificates.
- Transparency & Provenance, i.e., for users to conduct transactions in knowledge that each party has the capacity to enter that transaction.
- Immutability, i.e. for records to be written and stored permanently, without the possibility of modification.
- Disintermediation, i.e. the removal of the need for a central controlling authority to manage transactions or keep records.
- Collaboration, i.e. the ability of parties to transact directly with each other without the need for mediating third parties.

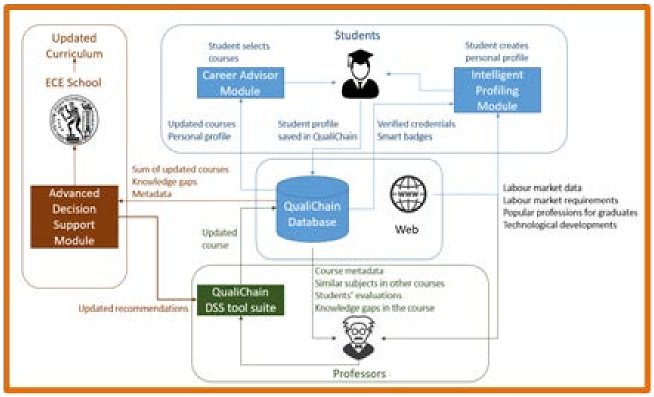


Fig. 7. Education Curriculum.

b) Certification: The Hyperledger Fabric is an open source blockchain platform (refer to Fig. 8 and 9).

Hyperledger Fabric Architecture and permits for faster blockchain network. The application connects with blackchin via REST web, APIs and by using HTTP and JSON data format [25].

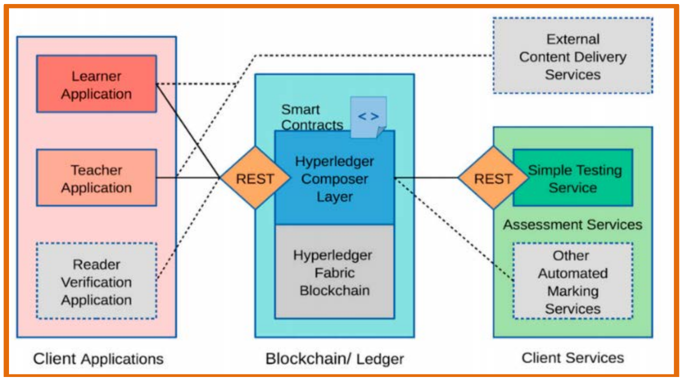


Fig. 8. Hyperledger Structure.

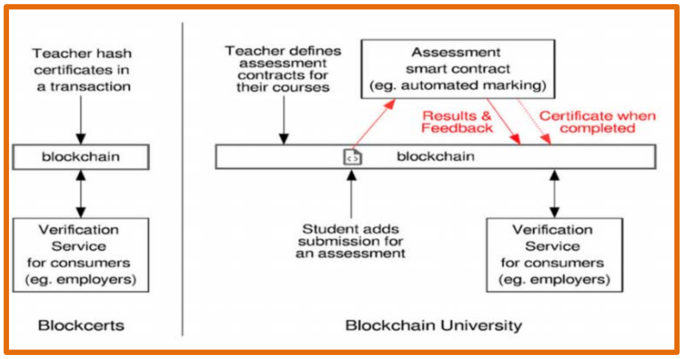

Teacher hash certificates in a transaction
blockchain
Verification Service for consumers (eg. employers)
Blockcerts
Teacher defines assessment contracts for their courses
Assessment smart contract (eg. automated marking)
Results & Feedback
Certificate when completed
blockchain
Student adds submission for an assessment
Verification Service for consumers (eg. employers)
Blockchain University

Fig. 9. Blockcert (Blockchain Certificates) for University.

c) Smart Contract: Smart contract covers ranges of facilities in crypto-educations system, from enrolling into course to coursework evidenced based, timeline tracking, helps both tutors and learners' opportunities to manage datelines, fast enquiry and verifications, track of students' learning progress etc.

Transaction processed in transparent and scalable manner to insure full transparent [24].

- Summarized five Timestamped logs.
- Cryptographic &Timestamped Logs.
- Cryptographic hash function.
- Timestamped append-only Logs.
- Block headers & Merkle Trees.

d) Immutability: Technology profoundly changed educational system but "the way we issue and managed academic credentials, which represent learning outcomes and achievements, has not yet taken advantage of the possibilities of digital technology". The blockchain can resolve these issues and provides data access to universities in permissions-less or permissioned (private permissioned -loop) such as peers control access described by [25]. Peers control access model discussed and adopted worldwide by large organization and private consortiums. Is less complex to use and tailored to specific application.

"The adoption of cloud computing is in line with the objective of saving resources/saving costs and improving the interaction between students and the educators. "The future of How Blockchain deliver its activities in absent of administrators and how tasks, sequences and process occurred in a simple structured and intelligent way and protect itself without external forces.

Although blockchain structured as database or system platforms blockchain is also machine learning, deep learning and autonomous systems simply does everything and can do everything. Large corporation making the utmost benefits of blockchain, Sony, IBM, IEEE, etc.

e) Cybersecurity: Security is one of essential unique feature of blockchain, hence each step carried with security in mind. These three-encryption control mechanisms are indispensable for blocks encryptions:

- The system randomly generates a 32-bit whole number known as (nonce) this number only use once, which is then added to a hashed to encrypt. These steps are essential to protect data from cypher attached.
- A block header hash is then generated.
- Hash is a 256-bit cipher merged with the nonce. Data kept within the block.

The hash function key is one of the most important elements in blockchain structure and Merkle tree. These sets of structure's differentiate blockchain uniqueness [14], shown in Fig. 14.

f) Sustainability: Sustainability is the future business. The subject of digitalization and blockchain adoption is not completed without in depth instigation into sustainability, and business continued and its efficiency and consumption. Given the governing authorities and strategies taken by a given government, compliance is needed to carefully address incompatible innovation and instigate threats to digital sustainability and assessment, in particular the role of politics. [26]. Shown in Fig. 10, adaptive governance can be defined as the adjustment of regulatory rules and practices to incorporate new data and to balance the risks and benefits of a given activity.

g) Consensus: It is important part of validation by all participants in this mechanism Blockchain enhances trust across a business network". With blockchain trust comes natural within the network community because it provides cryptographic proof over a set of transactions.

"Because the transaction can't be tampered with and are signed by the relevant counterparties, any corruption is readily apparent". Blockchain is sustainable and its existence will be continued since it's an independent and not own by any vendor.

Organizations like the IEEE developing courses in many areas of education to sell to corporate and other societies. Blockchain strengthen existing standards. "As a record-keeping strategy, blockchains pose a novel way to track whether participants are complying with standards that they publicly claim". [4]. IEEE taking full advantageous of blockchain smart processing tools and tight security.

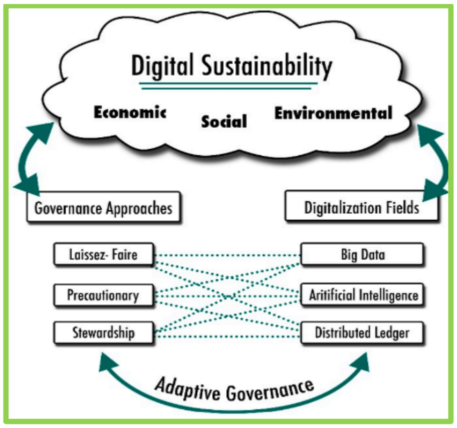


Fig. 10. Digital Sustainability.

h) Consensus process: Blockchain and our smart credential platform in the education sector will explore areas of automatic recognition and attribution or transfer of credits, verifying accreditors lifelong records of learning, student grants and funding also can leverage sovereign identities for verifying students and even educators" [27]. Sony creating trusted platform using blockchain. "We will be demonstrating how blockchain will become the future of maintaining and managing transcripts and high security data in education" [27].

IBM [14] and Linux Foundation also in partnership with Sony developed 'Hyperledger'. These initiatives enables multiple institutions to add transcription and student's achievement on a ledger records for a lifetime. "Blockchain technology makes it possible to associate these types of data with individuals" [20]. The authors added "This will enable people to handle their own academic progress record, and to have control over how to improve their learning cycle, for example. I believe this will change education as we know it and make it substantially more efficient," [20]. As a global leader in this field there is no doubt his input will shape the world of digital education. IBM hyper ledger offer both public and private tailored access to its end-users.

Blockchain technology spreading fast and among large and reputable universities and institutions, this not to underestimates the challenges emerging from researchers challenging blockchain, efficient and scalability of the system, [28]. In a permission-based system, decision taking to for an example, how to ignore malicious nodes and mining processing executions computation. There also concern about designing of smart contract based blockchain application.

Many enterprises integrate blockchain with their systems for the benefits of the blockchain. Despite its strength, blockchain has some challenges in security, privacy, scalability, and other issues [28]. Although blockchain breakthrough solves major issues, certain areas within blockchain deserve further attention and scrutiny. There should be more critical analysis and evaluations, to ensure earl rectification of upcoming problems. Blockchain is a transformation that changed the world for better, despite the slow progress of these technologies beside educational system.

Other technologies emerging from blockchain can bring some social justices to humanity. The past decades technology is controlled by only half of the world and the rest are poor, deprived and disfranchised. Half of the world still does not connect to basic elements of technology for example electricity, education and healthcare, the main reason for these is those intermediaries' systems of greed.

Promotes transparency, "You build peer-to-peer self-sustaining applications that can stand its ground. From destabilized governments to supply chain vendors, use cases of blockchain are endless. Control over personal information is vital and can be successfully addressed using blockchain" [29]. Blockchain story is not going away, and its strength and capability is growing fast, "A billion dollars in venture capital has flowed to more than 120 blockchain-related start-ups, thirty of the world's largest banks have joined a consortium to design and build blockchain solutions, Nasdaq is piloting a blockchain-powered private market exchange and Microsoft has launched cloud-based blockchain-as-a-service" [30].

6) Digital education system: Saudi Government boosted its digital Education system and made enormous transformation by implementing a set of EdTech infrastructure 2017. Future Gate is a country-wide, large-scale initiative of Tatweer education and part of KSA vision for 2030 developments is to focus on digital literacy skills of 21st century.

7) Current education system in Saudi Arabia: Saudi Arabia student's outbound policy decreases and now reverses the equation by investing in their educational system and opening new opportunities of international student studying in the Saudi Arabia. Recent reports by WENR [32] showed slow on Saudi's outbounds students decreases and inbound increases both by national and international students' seeking education in Saudi Arabia. "The expansion of Saudi Arabia's higher education system and the construction of more universities in recent years has helped accommodate a rising inflow of students from other countries, resulting in a high inbound student mobility ratio of 4.6 percent between 2008 and 2017" [33].

Since 2019 Saudi Arabia open new venues of shared researchers and collaborations with international foreign universities to stimulate students' exchange. This is seen as positive steps forward to Saudi openness and commitment to educations. Fig. 11 and 12 show statistic carried by WENR.

Now Saudi Arabia authorized foreign universities to set up Campuses and there is strategy to slow outbound students' mobility. This certainly reduces outbound costs and strengthens inbound prosperity. This is also an opportunity for Saudi to adopt blockchain technology for its international exchange and by slowing outbounds students and bring in international universities can provide cost effectives digital education at home and make double win saving and attracting inbounds students by using Blockchain Technology to facilitate and improve education system [32].

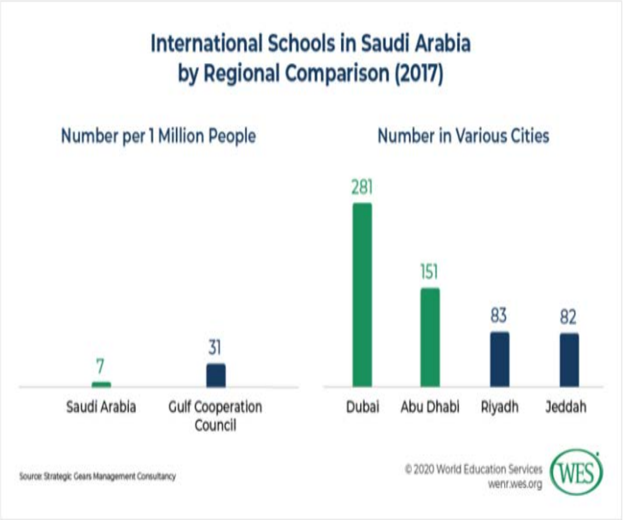


Fig. 11. Statistics of Outbound Saudi Arabian Students.

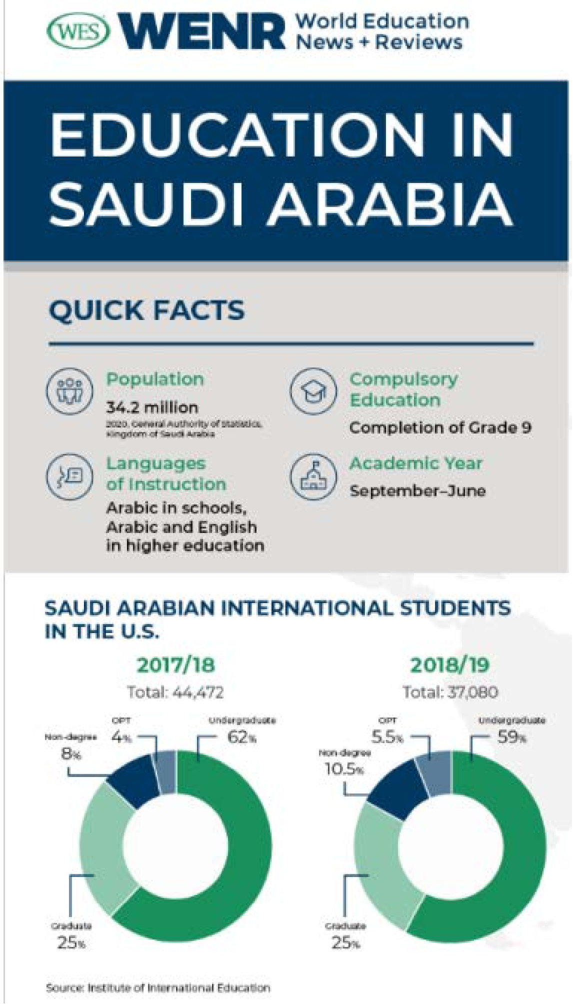


Fig. 12. Saudi Arabian International Students in The U.S.

Blockchain Potential Benefits and Impact on Saudi Arabia Education, as shown in Fig. 13:

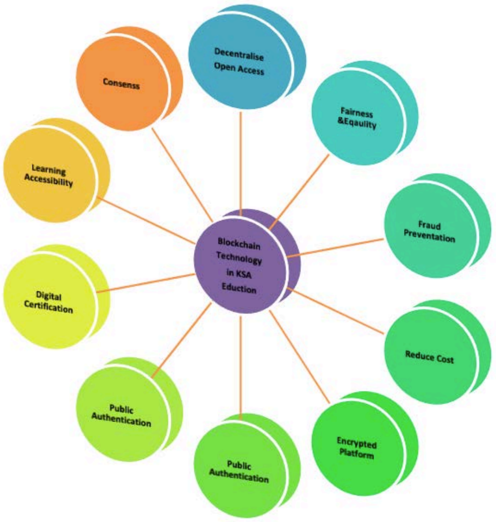


Fig. 13. Proposal Blockchain Potential Benefits to Saudi Education.

II. SWOT Analysis of Using Blockchain Technology

A. Strength of using Blockchain Technology

1) Blockchain is decentralized data setting is incorruptible and consistently reliable online database registry of various digital transactions where participants can modify the data by following a process of approval procedures: this called consensus and, in this process, all nodes in the network or majority must agree to transactions.

*2) Blockch*ain is an encrypted system that uses different styles of encryption and hash to store data in protected databases. The system distributes these data records over various nodes and forms a consensus on the position of the data they contain that makes difficult to corrupt.

*3) Blockch*ain data and information not stored in local system or cloud space but in multiple nodes tight secured. Main principle of decentralization, information not kept in one space and data is transparent [21].

4) Smart contracts are computer protocol that facilitated and enforce negotiation. Can be partially or fully and self-executed and self-enforced. In Ethereum these contracts have the ability to communicate within their internal storage and analyses, evaluate and send message to trigger execution. Smart contracts offer valuable advantages: such as enforcement, management, and payment, and performance, these aspects must be satisfied in fast and effective manners without third party authorization [21].

B. Weaknesses associated with the use of Blockchain Technology

1) Existing centralized vendor managed system both traditional and cloud-based structures offer services at sum costs and provides in return, computing services that includes software, storage, servers, database, networking, analytics reports, intelligence reports and flexible utilization of resources [18]. Thus, makes competition to newly emerged blockchain harder or less favorable to businesses comfortable with these facilities.

2) Blockchain has some challenges in security, privacy, scalability [33].

3) Blockchain is here and to stay and to influence major actors globally, it has been recognized and gradual implementation across industry spreading fast [35] (refer to Fig. 16 blockchain mind map).

4) Old system offers history of services and builds trusts among users and vendors common interests. Offer transformation and rapid innovation to its end-users.

5) Traditional and newly formed cloud space, structure of database stored locally and on cloud storage with simple procedures access. In centralized system (data stored in a company's data centers). Offer tailored and optional transparency and stored data can be visible to public or private as requested.

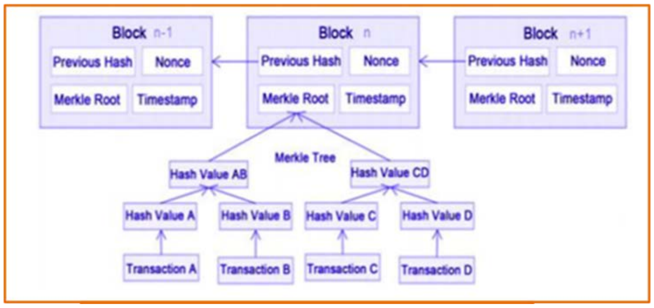


Fig. 14. Hash Function.

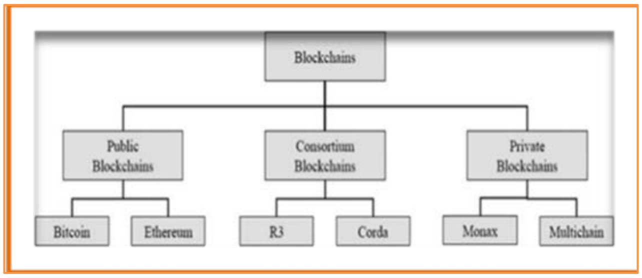


Fig. 15. Types of Blockchain.

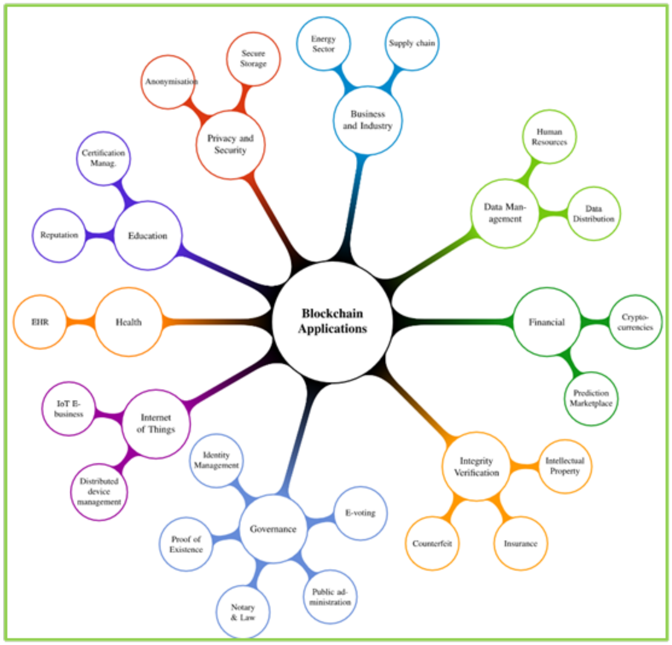


Fig. 16. Mind Map Abstraction of different Types of Blockchain Applications.

C. *Opportunity Associated with Blockchain Technology*

1) Recoded ledger and immutability.

2) Tight cryptography (using hash function).

3) Decentralizes and open access.

4) Availability.

5) Strength distributed resilience and control.

6) Decentralized network and open source.

7) Asset's provenance.

8) Dynamic environment.

9) By removing intermediaries and delivering a dependable shared view of permissioned data, blockchain could:

*10)*Reduce costs effectively: (e.g. Administrations, university running costs, Third party and students).

*11)*Speed up settlement (e.g. faster validation, digital enrolment verifications, and certification).

*12)*Increase resilience (e.g. no single point of failure).

*13)*Improve transparency (e.g. easier to monitor).

D. *Threat*

Threats and risks by [31] analysts seem glomming but effective as well. Ledger conflicts competition, Lack of ledger interoperability Lack of inter-ledger governance and limitation of smart contract coding programming.

III. Conclusion

To conclude this study, blockchain has reached its maturity in education system globally, despite the speed at which educational establishments are moving towards complete change of system and full blockchain technology adaptation. The fact Blockchain still available for adaptations, those who start first will benefit more. The scale of emerging educational platform using blockchain Technology growing fast and making use of all benefits ranging from smart contracts system of payments to Id identification and token and students learning authentications, thus the story of blockchain is growing and staying.

World Class education Institutes are using the technology to their advantageous, for instance the powerful educational Institutes such as Cambridge, MIT, are teaming up to create Blockcerts, an open standard platform for creating and issuing verifying certificate's. In 2018 MIT 600 graduates accepted digital version of their diplomas on Blockcerts blockchain, other universities and educational. A consortium consisted of top world university called (DCC) Digital Credential Consortium issuing both credential and transcripts already in place. During this study research encounter several papers discussing the weakness of this technology and focus was on the scalability issues and shortage of skills rather than major structural issue with blockchain frameworks. These voices raised concerns in relation to scalability and proof-work, others focus on privacy and closed loop system. Blockchain has fruitful future and benefits to all industry including major benefits to education system. This study has its limitation; however a better review in future will cover most areas in more prices and systemic approach. Blockchain offers real promise to all industries without exceptional. This surveyor focused on blockchain functionality and potential for its education system. Most research focused on the main issues blockchain differentiates itself from the reset of the other technology, which is the decentralization and immutability and transparency and more advancement into certification platform such as Hyperledger's. There should be further in-depth research mainly focused in how these platform developed and functionality i.e. prototyping, product testing and survey on privacy vs openness of decentralization. Majority of research developed built on previous literatures and limited further opinions or new development which seems closed loops as if all researchers working from one point of view. There is no underestimation of the prefund efforts made by the blockchain community of research. New development need and research guidance and direction also missing.

References

[1] Nakamoto, S., & Bitcoin, A. (2008). A peer-to-peer electronic cash system. Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf, 4.

[2] Sabry, S. S., Kaittan, N. M., & Majeed, I. (2019). The road to the blockchain technology: Concept and types. Periodicals of Engineering and Natural Sciences (PEN), 7(4), 1821-1832.

[3] Kristofer J Carlson, October 2018, https://www.researchgate.net/publication/328581315_The_Nakamoto_Blockchain.

[4] De Filippi, P. D. F. (2018). Blockchain and the law: The rule of code. Harvard University Press.

[5] Quintais, J., Bodó, B., Giannopoulou, A., & Ferrari, V. (2019). Blockchain and the law: A critical evaluation. Pedro Quintais, B. Bodó, A. Giannopoulou, & A. Ferrari (2019). Blockchain and the Law: A Critical Evaluation. Stanford Journal of Blockchain Law & Policy (2), 1.

[6] Mohanta, B. K., Jena, D., Panda, S. S., & Sobhanayak, S. (2019). Blockchain technology: A survey on applications and security privacy challenges. Internet of Things, 8, 100107.

[7] Sabry, S. S., Kaittan, N. M., & Majeed, I. (2019). The road to the blockchain technology: Concept and types. Periodicals of Engineering and Natural Sciences (PEN), 7(4), 1821-1832.

[8] Sharmila, K., Kamalakkannan, S., Devi, M. R., & Shanthi, M. C. (2019). A comprehensive study on blockchain with its components, taxonomy and consensus.

[9] Hashmani, M. A., Junejo, A. Z., Alabdulatif, A. A., & Adil, S. H. (2020, October). Blockchain in Education–Track ability and Traceability. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 40-44). IEEE.

[10] Pina, A. R. B., Torlà, C. B., Quintero, L. C., & Segura, J. A. (2017). Blockchain en Educación: introducción y crítica al estado de la cuestión. Edutec. Revista Electrónica de Tecnología Educativa, (61), a363-a363.

[11] O'Dwyer, K. J., & Malone, D. (2014). Bitcoin mining and its energy footprint.

[12] Sobti, R., & Geetha, G. (2012). Cryptographic hash functions: a review. International Journal of Computer Science Issues (IJCSI), 9(2), 461.

[13] Morgen Peck, Freelance Technology Writer, A White Paper on "Reinforcing the links of Blockchain " in IEEE Spectrum Magazine special edition "Blockchain World", November 2017.

[14] Doaa Mohey El-Din M. Hussein, Mohamed Hamed N. Taha, Nour Eldeen M. Khalifa Faculty of Computers and Information Cairo University Egypt, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 8, 2018

[15] Jeffrey T. Fouts, 2000, Bill and Milenda Gate Foundation.

[16] Wikramanayake, G. N. (2005). Impact of digital technology on education.

[17] Alhumaid, K. (2019). Four Ways Technology Has Negatively Changed Education. Journal of Educational and Social Research, 9(4), 10-10.

[18] The cyber threat to Universities. (2021). Retrieved 2 April 2021, from https://www.ncsc.gov.uk/report/the-cyber-threat-to-universities.

[19] Deloitte blockchain Global Surevy, 2019-2020. from https://www2.deloitte.com/content/dam/Deloitte/se/Documents/risk/DI_2019-global-blockchain-survey.pdf.

[20] Masaaki IsozuP resident, Sony Global Education, Inc. https://www.sony.com/en/SonyInfo/sony_ai/blockchain.html.

[21] Nadeem bhati, Hyperledger, 20 May 2019, https://www.hyperledger.org/blog/2019/05/20/developer-showcase-series-nadeem-bhati-high-school-technology-services.

[22] Kontzinos, C., Markaki, O., Kokkinakos, P., Karakolis, V., Skalidakis, S., & Psarras, J. (2019). University process optimisation through smart curriculum design and blockchain-based student accreditation. In Proceedings of 18th International Conference on WWW/Internet.

[23] Grech, A., & Camilleri, A. F. (2017). Blockchain in education. Luxembourg: Publications Office of the European Union.

[24] Herian, R. (2018, December). Legal recognition of Blockchain registries and Smart contracts. EU Blockchain Observatory and Forum.

[25] Ullah, N., Mugahed Al-Rahmi, W., Alzahrani, A. I., Alfarraj, O., & Alblehai, F. M. (2021). Blockchain Technology Adoption in Smart Learning Environments. Sustainability, 13(4), 1801.

[26] Van Oudheusden, M. (2014). Where are the politics in responsible innovation? European governance, technology assessments, and beyond. Journal of Responsible Innovation, 1(1), 67-86.

[27] Sony Global Education, 2017, Creating a Trusted Experience with Blockchain, https://blockchain.sonyged.com

[28] Gupta, M. (2020). Blockchain for Dummies (3rd ed.). Hoboken: John Wiley & Sons, Inc.

[29] Alajmi, Q, Sadiq, A. Kamaludin, A., & Al-Sharafi, M. A. (2017, May). E-learning models: The effectiveness of the cloud-based E-learning model over the traditional E-learning model. In 2017 8th International Conference on Information Technology (ICIT) (pp. 12-16). IEEE.

[30] Building the digital credential infrastructure for the future. (2021). Retrieved 1 April 2021, from http://philippschmidt.org/articles/2020-01-White-paper-building.

[31] David Furlonger, Ray Valdes, Published: 03 March 2017, https://blockcointoday.com/wp-content/uploads/2018/04/Practical-Blockchain_-A-Gartner-Trend-Insight-Report.pdf.

[32] Sidiqa Allahmorad, and Sahel Zreik, World Education Services April 9, 2020,

[33] Sanka, A. I., Irfan, M., Huang, I., & Cheung, R. C. (2021). A survey of breakthrough in blockchain technology: Adoptions, applications, challenges and future research. Computer Communications.

[34] Schatsky, D., & Muraskin, C. (2015). Beyond bitcoin. Blockchain is Coming to Disrupt Your Industry.

[35] Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: current status, classification and open issues. Telematics and informatics, 36, 55-81.

Supporting Multi-interface Entities in Software-Defined Wireless Networks

Jun-Hyuk Park[1], Wonyong Yoon[2]
Department of Electronics Engineering
Dong-A University, Busan
South Korea

***Abstract*—Software-Defined Networking (SDN) has gained growing momentum from its earlier application for wired networks (e.g., data center networks) to its application for wireless and mobile networks. In addition, state-of-the-art wireless and mobile networks (cellular networks and mesh networks) have been enhanced through the integration of multiple radio access technologies or multiple interfaces. This paper considers how to evolve multi-interface wireless mobile networks according to a future SDN-based paradigm, and deals with the technical problems therein. It presents the design and implementation of mechanisms that support SDN-based control of two types of multi-interface wireless mobile networks: one with multi-interface user devices and the other with multi-interface switching entities. As a methodology of demonstrating the feasibility of the proposed solution, a novel testing suite incorporating a real SDN controller and a standardized network simulator is designed and built. The functional verification and performance of the proposed solution is demonstrated in a virtual network topology but with the orchestration of the real SDN controller. The results show the multi-interface wireless entities can exploit the multi-radio and multi-channel wireless resources with the help of SDN approach.**

Keywords*—*Software-defined networking; multi-interface; multi-interface switch; flow-precision mobility; flow-precision routing

I. Introduction

Software-Defined Networking (SDN) has recently drawn significant attention from the networking communities owing to its potential to introduce innovative networking mechanisms by decoupling the control plane from the data plane of networks and centrally controlling the resulting networks [1]. The recent high interest in SDN is greatly attributed to Google's B4, which has nicely applied the SDN principle to real operating wide-area networks connecting Google data centers globally [2]. B4 enables near 100% link utilization between data centers using SDN-based centralized traffic engineering and OpenFlow-based switch management [1]. In addition to works on SDN-based data center networks, other SDN applications to wired networks include applying SDN for inter-domain routing [3], and adopting multiple controllers for wide-area wired network orchestration [4]. Gupta and co-workers proposed an SDN-based Internet Exchange Point (IXP) and solved various challenges in building real deployable software defined exchange points [3].

The application of SDN to wireless networks has been recently and extensively studied, for example, wireless local area networks (WLANs), wireless mesh networks, and cellular networks like Long Term Evolution (LTE) [5], [6]. SDN-based approaches to WLANs, both enterprise WLANs [7]-[9] and community WLANs [10], have been proposed for various purposes of network virtualization, radio resource management, flow-based quality of service (QoS), and mobility support. The SDN-based control of wireless mesh networks has been examined from the perspective of load balancing and mobility management [11], [12]. LTE cellular networks have been recently examined for an SDN rebase from the perspective of both the core network part [13]-[20] and radio access network part [21]- [23].

Apart from SDN-driven research, it is noticed that multi-RAT(Radio Access Technology) wireless networks have been a promising approach to resolve the ever-increasing bandwidth demand as various radio access technologies become available with different communication ranges, and radio interfaces become more and more affordable in terms of the unit price [24]-[26]. In multi-RAT wireless networks, either devices (e.g., smartphones) or switching entities (e.g., routers) can be equipped with multiple radio interfaces to provide and utilize an enhanced network capacity. A typical example of the former case is multi-RAT LTE/WLAN heterogeneous networks where user equipment (UE) such as smartphones have LTE and WLAN radio interfaces available for radio access. A typical example of the latter case is multi-RAT, multi-channel wireless mesh networks where wireless routers are equipped with multiple WLAN radio interfaces simultaneously operating on different WLAN channels for capacity enlargement [25]. It is envisaged that these multi-radio wireless mobile networks can be orchestrated based on the SDN paradigm with a centralized holistic view on multi-radio resources more effectively than traditional distributed local control. For example, congestion on LTE access networks can be alleviated by offloading LTE/WLAN multi-radio devices to WLAN access points (APs) based on the centralized decision of the SDN controller, thereby leading to more enhanced multi-radio resource utilization and load balancing.

This trend toward softwarizing networks and enhancing multi-interface availability motivates the authors to examine, to the best of the authors' knowledge, the first in-depth investigation into the feasibility of SDN-based centralized control of multi-RAT wireless mobile networks as a next-generation evolution. In so doing, technical issues induced by multi-radio inherency are studied, which can be further classified into handling (i) multi-RAT devices and (ii) multi-

This work was supported by the Dong-A University research fund. W. Yoon is the corresponding author. The authors thank S. G. Lee.

RAT switching entities in the SDN paradigm. Mechanisms are proposed to resolve the issues of these two categories and implement them into an integrated suite of a real SDN controller and standard-compliant software switches in a simulation space. The first issue is considered in the context of emerging LTE/WLAN heterogeneous multi-radio networks, and thus, how to support LTE/WLAN multi-radio devices using an SDN controller and what benefits an SDN-based approach will bring. The second issue is considered in the context of multi-radio, multi-channel wireless mesh networks, and thus, how to support multi-radio routers using an SDN controller and its effectiveness. Although a case for SDN-based heterogeneous networked environments has been made [27] [42] [43], our investigation tackles detailed technical issues of resolving the limitation of the current open SDN standard in supporting multi-interface wireless network entities and incorporates procedural mechanisms.

The rest of the paper is organized as follows. Section II summarizes previous related works. Section III discusses the problem of supporting devices with multi-RAT interfaces in SDN-based wireless networks and proposes a novel solution to enable it. Section IV discusses the problem of supporting switching entities with multi-RAT interfaces in SDN-based wireless networks and proposes a novel solution to support it. Section V demonstrates the performance evaluation of the proposed solutions. Section VI provides some concluding remarks regarding this research.

II. Related Work

There are previous works applying SDN approaches to WLANs, both enterprise WLANs [7]-[9] and community WLANs [10]. Suresh and co-workers proposed Odin as an open SDN framework for an enterprise WLAN, where the Odin master on an SDN controller communicates with Odin agents at the access points to support seamless mobility, load balancing, hidden terminal mitigation, and other management functions [8]. Schulz-Zander and others proposed a two-tiered control architecture, called AeroFlux, where global control servers deal with global mobility management and load balancing, and near-sighted control servers manage the per-client or per-flow Wi-Fi transmission settings [9]. Yiakoumis and co-workers studied SDN-based home WLANs and proposed a mechanism called BeHop for virtual slicing of a physical WLAN, a personal AP abstraction, and infrastructure management such as for the channel and power [10]. Lee and others proposed an extension of SDN to mobile phones called meSDN [7]. By incorporating an OpenFlow-based local controller in a mobile device, they enabled application-aware uplink flow QoS, a network fault diagnosis, and WLAN virtualization.

Applying the SDN paradigm to multi-hop wireless mesh networks with Wi-Fi radio interfaces has also been studied [11], [12], [32]. Detti and co-workers proposed wmSDN, which is an SDN-based multi-hop wireless mesh network, mainly for the purpose of load balancing between multiple WMN gateways [12]. Dely and others proposed an OpenFlow-based architecture for flexible routing in wireless mesh networks [11] [32]. They demonstrated that the simplification of client mobility between mesh access points is feasible in conformance to OpenFlow messaging triggered by the handover decision of a monitoring server. Although both works introduced an SDN implementation of wireless mesh networks, they did not consider multi-radio capable routers in terms of topology management and routing. Peng et al. proposed Access Point handoff in SDN-based wireless networks by exploiting dual network interfaces [40].

Cellular networks, particularly LTE networks, have recently drawn significant interest from the research community and industries. Some works have studied the issues and solutions for an SDN-based core network (wired part) [13]-[20], and others have focused more on the issues and directions for an SDN-based radio access network (wireless part) [6], [22]. Nagaraj and co-workers propose offloading flows from the evolved packet core (EPC) to other IP transport networks based on flow classification [13]. Mahoodi and others considered SDN-based EPC and studied the case of an intra-LTE handover procedure [14]. Heinonen and co-workers proposed a mechanism for dynamically switching the user plane tunnels in SDN-based cellular core networks [15]. Moradi and others looked into the scalability issue of SDN-based cellular WANs and proposed a three-level control of the switch regions and a recursive topology discovery mechanism [16]. Jin and co-workers proposed a scalable and flexible LTE core network architecture that can program forwarding to middle boxes based on service policies [17]. Pentikousis and others studied applying SDN to mobile core networks, and proposed a flow-based forwarding architecture for the benefit of the operators [18]. Basta and co-workers examined possible scenarios for the placement of EPC functions with a consideration of network function virtualization (NFV) [19]. Hampel and others proposed using SDN to program the encapsulation and decapsulation on top of IP to enable flow-based policy enforcement and mobility [20]. Gudipati and co-workers considered applying the SDN concept to LTE radio access networks, and proposed abstracting multiple eNBs as a single big base station and centrally managing radio resources together for load balancing and utility optimization [21], [22]. There are works that seek joint coordination of radio access networks and transport core networks [33], [34]. Bernardos and co-workers considered accommodating various wireless mobile networks in the SDN realm, and proposed a high-level architecture and the required northbound and southbound interfaces [33]. Tan and others proposed a wireless SDN architecture with a distributed user plane and a centralized function-oriented control plane, and demonstrated end-to-end QoS management and a content-aware data broadcast in converged LTE and Wi-Fi networks [34]. Most recently, Abdulghaffar [39] proposed SDN-based 5G core networks.

Most relevant to our work, some works have considered SDN-based control of heterogeneous wireless mobile networks, including works by Mendonca and co-workers [27], CROWD [35] [36], OpenRoads [37], and SoftMobile [23]. Mendonca and others first examined the possibility of applying an SDN approach in heterogeneous networks consisting of wired networks, infrastructure-based wireless networks, and infrastructure-less wireless networks (e.g., mobile ad hoc networks) [27]. They considered typical use cases such as an SDN-based gateway assignment and data offloading. Ali-

Ahmad and co-workers dealt with very dense networks of multiple radio technologies, and proposed a two-tier controller architecture for scalability in the control of dense networks [35], [36]. Yap and others proposed an OpenFlow-based wireless architecture for orchestrating and virtualizing heterogeneous wireless networks, and demonstrated a case study of WiMAX to/from a Wi-Fi handover [37]. Chen and co-workers considered complex control plane issues in multi-RAT heterogeneous mobile networks, and proposed a holistic control framework for providing a global network view (spectrum, connection, and interference map, among others) and abstraction along with coordination algorithms (for example, joint scheduling and interference cancellation) [23]. Mafakheri and co-workers took software-defined radio access network approach for LTE and WLAN coordination [38]. Alshaer et al. considered the problem of providing application QoS in SDN-based heterogeneous dense wireless networks [41].

III. Supporting Multi-Interface Device

As the first type of multi-radio wireless mobile networks, the authors consider the LTE/WLAN multi-RAT wireless network illustrated in Fig. 1, where user devices (e.g., smartphones) are equipped with both an LTE radio interface and a WLAN radio interface. In 3GPP terminology, these are called UE. An entire LTE network in the data plane consists of an access network part, i.e., evolved node B (eNB), and core network parts, i.e., the serving gateway (S-GW) and packet data network gateway (P-GW). Although 3GPP has specified a tight integration of an LTE network and WLAN networks with P-GW as an anchor, most mobile network operators resist deploying it owing to the high cost of backhauling WLAN networks to an LTE core network part [26], [28]. Thus, the current norm is a loose integration or non-seamless integration of both networks, as depicted in Fig. 1. In this network, multi-radio devices typically make a radio access technology (RAT) selection using their connection manager software with only local logic. Furthermore, once selected initially for a flow, a RAT cannot be changed to another RAT dynamically. For these reasons, the current network has difficulties in achieving an optimal multi-radio resource utilization and providing quality of experience (QoE) dynamically [26]. It is envisaged that the SDN-based integration of LTE and WLAN networks can resolve the above issues without causing a severe backhaul cost.

Tailoring multi-radio wireless mobile networks to the SDN rebase basically involves making OpenFlow channels between an SDN controller and each switching entity, as indicated by the dotted lines in Fig. 1 and 2. Note that we assume OpenFlow to be used as an SDN standard [29]. Issues in the design of such SDN-based multi-radio networks arise in two facets: supporting multi-radio devices and supporting multi-radio switching entities.

In Fig. 1, as per the current SDN standard, each interface of a multi-radio device is separately known to attach a port of a corresponding switching entity to an SDN controller. For this reason, the SDN controller cannot distinguish a LTE/WLAN multi-RAT device from two single-radio devices with LTE and Wi-Fi, respectively. Identifying a multi-radio device and serving it with a combination of adequate radios with the help of the SDN controller is not possible in the current SDN standard, and hence, some new enhancements are needed to resolve this issue.

An SDN controller’s awareness of multi-radio devices is a key enabler to the effective utilization of multi-radio access technology (RAT) resources. For example, flow-precision inter-RAT mobility for a multi-radio device can be supported in the SDN context to ensure load balancing for a holistic network perspective. For this purpose, the authors propose mechanisms for integrated device identification and inter-RAT flow mobility control.

Fig. 2 illustrates the mechanisms in the form of a message sequence chart and explains each step. SDN controllers should incorporate device management module with multi-RAT device support: When a device attaches to a RAT base station through a radio interface, the controller keeps track of an interface mapping from the radio interface (with a unique MAC address) to a {switch, port} tuple. When a device has multi-RAT attachments concurrently, the controller keeps track of a list of multiple interface mappings per device.

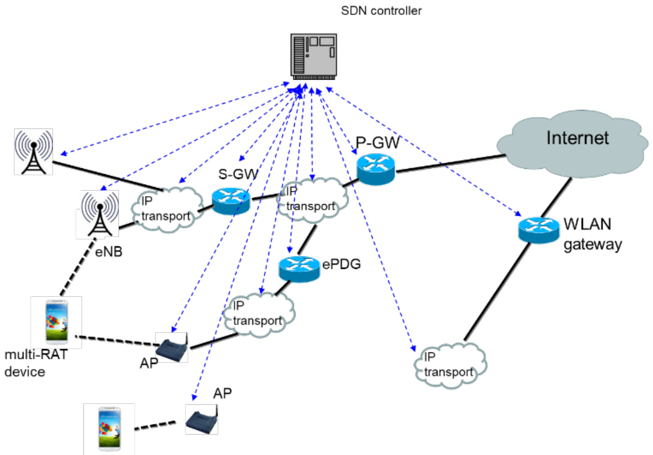


Fig. 1. An LTE/WLAN Network with Multi-RAT Devices. Its SDN Rebase is indicated by the Dotted Lines.

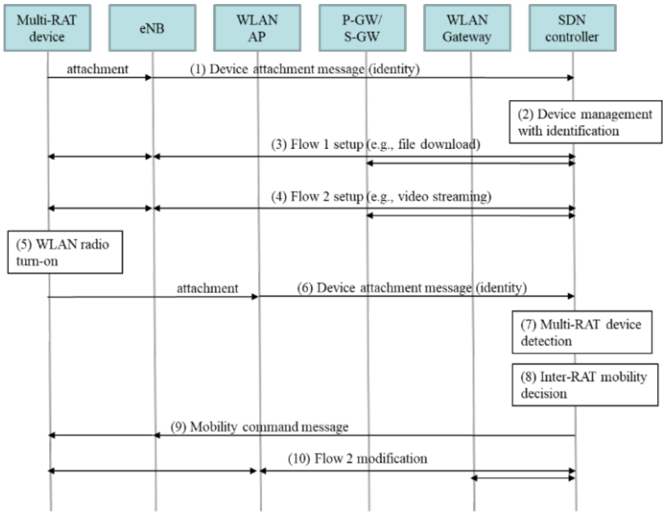

Multi-RAT device
eNB
WLAN AP
P-GW/ S-GW
WLAN Gateway
SDN controller
attachment
(1) Device attachment message (identity)
(2) Device management with identification
(3) Flow 1 setup (e.g., file download)
(4) Flow 2 setup (e.g., video streaming)
(5) WLAN radio turn-on
attachment
(6) Device attachment message (identity)
(7) Multi-RAT device detection
(8) Inter-RAT mobility decision
(9) Mobility command message
(10) Flow 2 modification

Fig. 2. A Message Sequence Chart for Handling a Multi-RAT Device.

(Step 1) A multi-radio device attaches to RAT 1, which is LTE, and a switching entity corresponding to eNB then notifies the SDN controller of the attachment and identification of the device. For the purpose of identification, an international mobile equipment identity (IMEI) can be used. (Step 2) The controller becomes aware of the device through LTE. (Steps 3 and 4) Two flows are set up based on the OpenFlow standard. (Step 5) The multi-radio device turns on its RAT 2 interface, which is a Wi-Fi interface, and discovers a WLAN AP. (Step 6) The device attaches to a WLAN AP and the AP in turn notifies the SDN controller of the attachment and identification of the device on WLAN. For the purpose of identification, the same IMEI can be used. (Step 7) Through the identification, the controllers handle the device as multi-radio capable. (Step 8) The controller makes a decision on the mobility of flow 2 from RAT 1 to RAT 2. (Step 9) A mobility command is sent out to the RAT 1 base station and device. (Step 10) A new flow table entry for flow 2 on RAT 2 is configured using an OpenFlow Flow Modify message.

IV. Supporting Multi-Interface Switch

As the second type of multi-interface wireless networks, the authors consider the multi-radio multi-channel wireless mesh network illustrated in Fig. 3, where mesh routers are equipped with multiple IEEE 802.11-based radio interfaces operating on separate physical channels in parallel. In the figure, three different colors are used to indicate three non-overlapping WLAN channels. How to map each interface to a channel is a key concern for resource utilization, and static, dynamic, and hybrid channel assignments have been extensively studied. The current state-of-the-art method is a distributed approach that can rely only on local information for a channel assignment [25]. The authors believe that an SDN-based holistic coordination of multiple (homogeneous) radio interfaces of each wireless mesh router and their operating channels can result in better utilization throughout the entire network.

In Fig. 3, each interface of a multi-radio switch (mesh router) should operate on different channels to fully exploit the available Wi-Fi channels. However, as per the current SDN standard, information regarding on which physical channel a port (i.e., a wireless interface) operates is lacking, and thus, the optimization of multi-RAT resource utilization is hard to achieve. Furthermore, a port for a wireless interface of a wireless switch should actually be used for both incoming packets and outgoing packets because it uses the wireless broadcast medium. However, the SDN standard only specifies that an incoming port should be different from an outgoing port, and thus, we need some mechanisms for supporting wireless ports in the SDN paradigm.

For a multi-interface switch, a radio port is first configured for each RAT interface (e.g., IEEE 802.11 radio interface) and its operating channel is initially configured. Because a radio interface uses the wireless broadcast medium, forwarding rules can be properly specified only by distinguishing which neighbor is the intended receiver. For this purpose, for a radio port, virtual radio ports should be defined for each neighbor relation. A wireless interference and channel reconfiguration may impact wireless neighbor links and hence trigger the creation or deletion of virtual radio ports. The neighbor switches of each switch can be discovered using the peer management protocol (PMP) of the IEEE 802.11s mesh standard. Thus, the switches themselves can determine whether a virtual radio port should be set for each discovered neighbor.

Fig. 4 illustrates mechanisms for handling multi-radio switches and multi-radio multi-channel aware routing in the form of a message sequence chart, and explain each step. (Step 1) A wireless port is set up, and a radio port status message including its operating channel is sent to the controller. (Step 2) Depending on how many neighbors the wireless port has, virtual radio ports are set up and known to the controller. (Step 3) Another wireless port is set up, and a radio port status message is transmitted. (Step 4) virtual radio ports for this port are set up and known to the controller. (Step 5) The controller is aware of a multi-radio switch, as well as its radio and virtual radio ports. (Step 6) A new flow 1 is created. (Step 7) The controller sets up a route for flow 1 from switch 1 to switch 2 and switch 4. (Step 8) Another flow 2 is created. (Step 9) The controller's multi-radio, multi-channel aware routing module determines whether to support the new flow 2 on a route from switch 1 to switch 3 and switch 4. (Step 10) The interfaces on the route use channel 2, which is less utilized than channel 1, and hence, can benefit from the multi-channel enhanced network capacity.

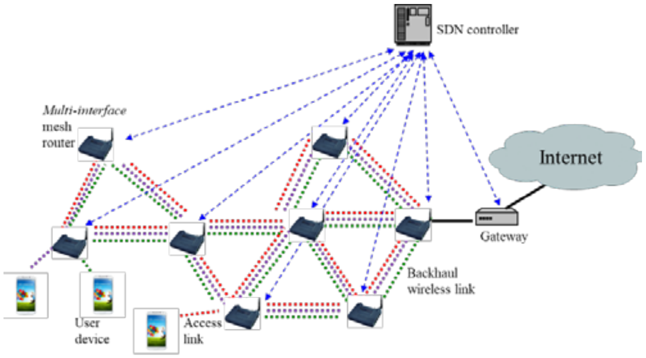


Fig. 3. A Wireless Mesh Network with Multi-interface Mesh Routers. Its SDN Rebase is indicated by the Dotted Lines.

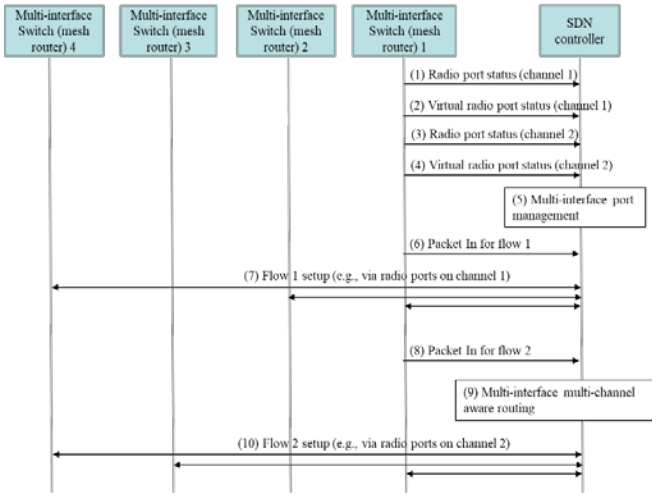

Multi-interface Switch (mesh router) 4
Multi-interface Switch (mesh router) 3
Multi-interface Switch (mesh router) 2
Multi-interface Switch (mesh router) 1
SDN controller
(1) Radio port status (channel 1)
(2) Virtual radio port status (channel 1)
(3) Radio port status (channel 2)
(4) Virtual radio port status (channel 2)
(5) Multi-interface port management
(6) Packet In for flow 1
(7) Flow 1 setup (e.g., via radio ports on channel 1)
(8) Packet In for flow 2
(9) Multi-interface multi-channel aware routing
(10) Flow 2 setup (e.g., via radio ports on channel 2)

Fig. 4. A Message Sequence Chart for Handling a Multi-radio Switch (Mesh router). A Routing Procedure Exploiting Multi-channel Diversity is also included.

Switch management module with multi-radio switch support: The controller keeps track of each switch's radio ports and the mappings from each radio port to a list of virtual radio ports. The data structure for the radio ports and virtual radio ports includes operating physical channels for multi-channel awareness.

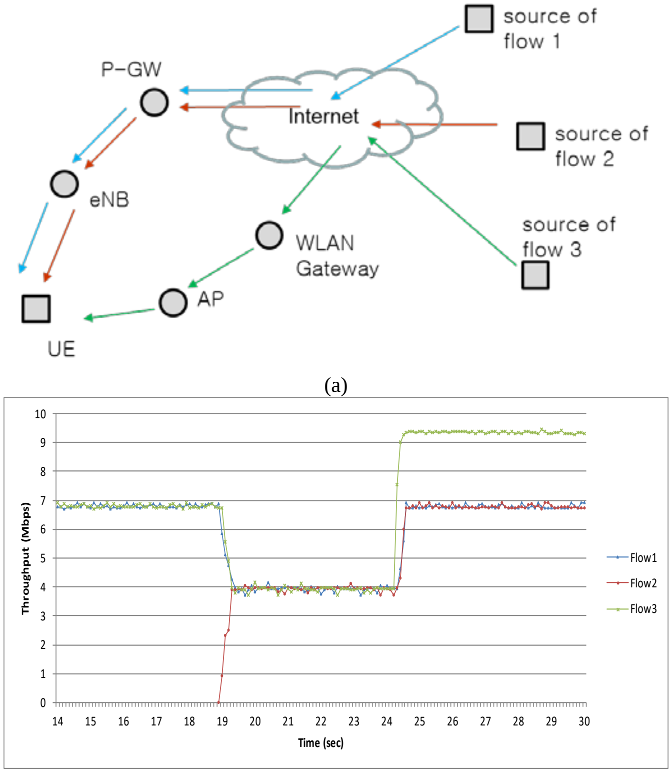
Forwarding module: When a new flow arises on multiple radio switches, a route for the flow can be determined based on a multi-radio, multi-channel aware algorithm to exploit channel diversity. Without multi-radio, multi-channel awareness, traditional shortest-path routing algorithms such as the well-known Dijkstra algorithm will only consider the link costs. For a multi-radio device, an existing flow on one radio can be examined to move to another radio based on an inter-RAT mobility algorithm with a holistic view of the entire multi-radio network.

V. Performance Evaluation

The authors built an integrated testing suite for SDN controller implementation [30] [31] and software switch implementation using a network emulation scheme. It is verified that the controller and switches interoperate with each other in conformance with the OpenFlow specifications. For the implementation of the proposed mechanisms for multi-radio wireless mobile networks, we use Vendor messages to carry the additional proposed information on multi-radio devices and multi-radio switches. In this way, legacy single-radio devices and switches can still be supported without modification. The modules in the switch software are coded in C/C++ languages. The authors designed and implemented the proposed mechanisms using a multi-radio switch and device support in a widely used network simulator space (ns-3), and made them interoperate with a real SDN controller. Hence, the inherent benefits of ns-3 (e.g., module reusability and extensibility) can be exploited. Our experience in this paper shows that researchers and developers can prototype and test new ideas and mechanisms for OpenFlow switches used in software in a timely and flexible manner.

The authors present two exemplary performance results to demonstrate the benefits of the proposed mechanisms. The tests were conducted using simple but effective topologies to highlight the benefits. We first examine the performance of the proposed SDN-based LTE/WLAN flow mobility procedure for a multi-radio device. We use a small topology in Fig. 5 for the ease of analysis. Initially, two flows (flows 1 and 3) exist over the LTE radio access technologies for a multi-radio device, and later, another flow (flow 2) over LTE is created. At time 24, the multi-radio device turns on its Wi-Fi radio interface, and according to the proposed mechanisms, the SDN controller instructs inter-RAT flow mobility for flow 3. The topology in Fig. 5(a) shows the flows after the flow mobility is performed. The traffic load on LTE is distributed to Wi-Fi, and hence, the network utilization is improved and individual flows can enjoy higher rates. This result highlights the strength of SDN-based holistic orchestration of heterogeneous multi-radio resources and flows.

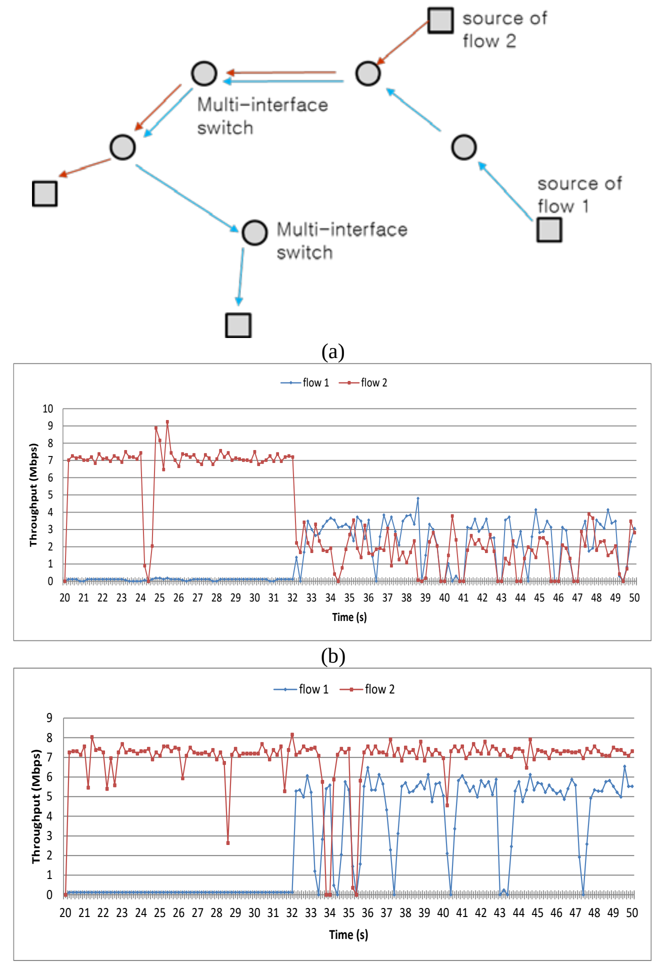
Next, we examine the performance of the proposed WMN flow-precision multi-channel routing. In Fig. 6(a), a wireless mesh network topology is illustrated. Each node is equipped with two Wi-Fi radio interfaces, and flow 1 (blue) and flow 2 (red) are considered for the simulation. First, Fig. 6(b) shows the throughput for when both flow 1 and flow 2 use the same channel, i.e., the SDN controller does not have a notion of multi-channel routing. Initially, flow 2 monopolizes the channel, and at time 30, flow 1 is created and uses the same channel. We observed that the channel is shared between the two flows. Next, Fig. 6(c) shows the throughput for multi-channel aware routing. The scenario is the same, but when flow 1 is created, the SDN controller can conduct multi-channel aware routing, and thus two orthogonal channels can be utilized by the two flows, respectively. We observed that the throughput is significantly better for multi-channel aware routing by the SDN controller. Although we use a simple multi-channel aware routing algorithm in this evaluation, we anticipate that more sophisticated algorithms can solve a very complex channel assignment and routing problems more effectively.



(a)

(b)

Fig. 5. An Example LTE/WLAN Topology for Evaluation of Inter-RAT flow Mobility: (a) LTE/WLAN Topology and (b) Throughput with inter-RAT Flow Mobility.



(a)

(b)

(c)

Fig. 6. An Example Multi-radio Mesh network Topology for Evaluation of Multi-channel Flow Routing: (a) Wireless Mesh network Topology, (b) Throughput with Single-channel Routing, and (c) Throughput with Multi-Channel aware Routing.

VI. Conclusion

As the emergence of multi-radio (heterogeneous or homogeneous radio) wireless mobile networks increases, the mechanism of centrally coordinating all available multi-radio network capabilities in a holistic manner has become essential. This paper presented the first in-depth study on how to incorporate multi-radio devices and multi-radio switches within the SDN paradigm. We examined the possibility of SDN-based control of two promising multi-interface wireless mobile networks, i.e., LTE/WLAN heterogeneous networks and multi-radio multi-channel wireless mesh networks, and proposed their control mechanisms. Through the integration of a real SDN controller and standard-compliant simulated switching entities, it is demonstrated the functional verification and performance of the proposed mechanisms. Future work may include the implementation and experimentation of the proposed solution in real network settings.

References

[1] N. McKeown et al., “OpenFlow: Enabling Innovation in Campus Networks,” ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, 2008, pp. 69-74.

[2] S. Jain et al., “B4: Experience with a Globally-Deployed Software Defined WAN,” ACM SIGCOMM'13, 2013, pp. 3-14.

[3] A. Gupta et al., “SDX: A Software Defined Internet Exchange,” ACM SIGCOMM’14, August 2014, pp. 551-562.

[4] R. Ahmed and R. Boutaba, “Design Considerations for Managing Wide Area Software Defined Networks,” IEEE Communications Magazine, vol. 52, no. 7, July 2014, pp. 116-123.

[5] N.A. Jagadeesan and B. Krishnamachari, “Software-Defined Networking Paradigms in Wireless Networks: A Survey,” ACM Computing Surveys, vol. 47, no. 2, January 2015.

[6] A. Weissberger, “Analysis of Open Network Foundation new 5G SD-RAN™ Project,” IEEE Commnications Society Technology Blog, August 2020.

[7] J. Lee et al., “meSDN: Mobile Extension of SDN,” ACM MCS’14, June 2014, pp. 7-14.

[8] L. Suresh et al., “Towards Programmable Enterprise WLANs with Odin,” ACM HotSDN’12, August 2012, pp. 115-120.

[9] J. Schulz-Zander; N. Sarrar; and S. Schmid, “Towards a Scalable and Near-Sighted Control Plane Architecture for WiFi SDNs,” ACM HotSDN '14, 2014, pp. 217-218.

[10] Y. Yiakoumis et al., “SDN for Dense Home Networks,” Open Networking Summit’14, March 2014, pp. 1-2.

[11] P. Dely, A. Kassler, and N. Bayer, “OpenFlow for Wireless Mesh Networks,” IEEE WiMAN’11, August 2011.

[12] A. Detti et al., ”Wireless Mesh Software Defined Networks (wmSDN),” Workshop on Community Networks and Bottom-up-Broadband (CNBuB 2013), October 2013.

[13] K. Nagaraj and S. Katti, “ProCel: Smart Traffic handling for a Scalable Software EPC,” ACM HotSDN’14, August 2014, pp. 43-48.

[14] T. Mahoodi and S. Seetharaman, “On Using a SDN-based Control Plane in 5G Mobile Networks,” WWRF32, May 2014.

[15] J. Heinonen et al., “Dynamic Tunnel Switching for SDN-Based Cellular Core Networks,” Workshop on All Things Cellular, August 2014, pp. 27-32.

[16] M. Moradi et al., “SoftMoW: Recursive and Reconfigurable Cellular WAN Architecture,” ACM CoNEXT'14, December 2014, pp. 377-389.

[17] X. Jin et al., “SoftCell: Scalable and Flexible Cellular Core Network Architecture,” ACM CoNEXT'13, December 2013, pp. 163-174.

[18] K. Pentikousis, Y. Wang, and W. Hu, “Mobileflow: Toward Software-Defined Mobile Networks,” IEEE Communications Magazine, vol. 51, no. 7, July 2013, pp. 44-53.

[19] A. Basta et al., “A Virtual SDN-enabled LTE EPC Architecture: a Case Study for S-/P-Gateways functions,” SDN4FNS’13, 2013.

[20] G. Hampel, M. Steiner, and T. Bu, “Applying Software-Defined Networking to the Telecom Domain,” IEEE Global Internet Symposium, April 2013, pp. 133-138.

[21] A. Gudipati et al., “SoftRAN: Software Defined Radio Access Network,” ACM HotSDN’13, August 2013, pp. 25-30.

[22] A. Gudipati, L.E. Li, and S. Katti, “RadioVisor: A Slicing Plane for Radio Access Networks,” ACM HotSDN’14, August 2014, pp. 237-238.

[23] T. Chen et al., “SoftMobile: Control Evolution for Future Heterogeneous Mobile Networks,” IEEE Wireless Communications, vol. 21, no. 6, December 2014, pp. 70-78.

[24] S. Song et al., "Coverage and Economy Modeling of HetNet under Base Station on-off Model," to be published in ETRI Journal, 2015.

[25] W. Yoon and N.H. Vaidya, “A Link Layer Protocol and Link-State Routing Protocol Suite for Multi-Channel Ad Hoc Networks,” Wireless Communications and Mobile Computing, vol. 12, no. 1, January 2012, pp. 85-98.

[26] R. Mahindra et al., “A Practical Traffic Management System for Integrated LTE-WiFi Networks,” ACM MobiCom’14, September 2014, pp. 189-200.

[27] M. Mendonca, K. Obraczka, and T. Turletti, “The Case for Software–Defined Networking in Heterogeneous Networked Environments,” ACM CoNEXT Student’12, December 2012, pp. 59-60.

[28] W. Yoon and B. Jang, “Enhanced Non-Seamless WLAN Offload for LTE and WLAN Networks,” IEEE Communications Letters, vol. 17, no. 10, October 2013, pp. 1960-1963.

[29] Open Networking Foundation, “OpenFlow Switch Specification,” Version 1.5.1, March 2015.

[30] B. Lee et al., “IRIS: the OpenFlow-based Recursive SDN Controller,” International Conference on Advanced Communication Technology (ICACT), 2014.

[31] S.H. Park et al., “RAON: Recursive Abstraction of OpenFlow Networks,” European Workshop on Software Defined Networks, 2014.

[32] P. Dely et al., “A Software-Defined Networking Approach for Handover Management with Real-Time Video in WLANs,” Journal of Modern Transportation, vol. 21, no. 1, 2013, pp. 58-65.

[33] C.J. Bernardos et al., “An Architecture for Software Defined Wireless Networking,” IEEE Wireless Communications, vol. 21, no. 3, June 2014, pp. 52-61.

[34] W. Tan et al., "SDN-enabled Converged Networks," IEEE Wireless Communications, vol. 21, no. 6, December 2014, pp. 79-85.

[35] H. Ali-Ahmad et al., “CROWD: An SDN Approach for DenseNets,” EWSDN'13, 2013, pp. 25-31.

[36] H. Ali-Ahmad et al., “An SDN-based Network Architecture for Extremely Dense Wireless Networks,” SDN4FNS’13, 2013.

[37] K.K. Yap et al., “Blueprint for Introducing Innovation into Wireless Mobile Networks,” ACM VISA’10, September 2010.

[38] Mafakheri et al., “LTE/Wi-Fi Coordination in Unlicensed Bands: An SD-RAN Approach,” IEEE NetSoft, June 2019.

[39] A. Abdulghaffar et al., “Modeling and Evaluation of Software Defined Networking Based 5G Core Network Architecture,” IEEE Access, vol. 9, January 2021.

[40] D. Peng et al., “A Dual-NIC Mutual Backup Solution of Access Point Handoff in SDN-based Mobile Networks,” ICCC, December 2020.

[41] Alshaer et al., “Software-Defined Networking-Enabled Heterogeneous Wireless Networks and Applications Convergence,” IEEE Access, vol. 8, April 2020.

[42] W. Huang et al, “QoE based SDN heterogeneous LTE and WLAN multi-radio networks for multi-user access,” IEEE WCNC, April 2018.

[43] P. Engelhard et al, “Software-Defined Networking in an Industrial Multi-Radio Access Technology Environment,” ACM SOSR, March 2018.

A Machine Learning based Analytical Approach for Envisaging Bugs

Dr. Anjali Munde
Amity College of Commerce and Finance
Amity University Uttar Pradesh
Noida, India

***Abstract*—A software imperfection is a shortcoming, virus, defect, mistake, breakdown or glitch in software that initiates it to establish an unsuitable or unanticipated result. The foremost hazardous components connected with a software imperfection that is not identified at an initial stage of software expansion are time, characteristic, expenditure, determination and wastage of resources. Faults appear in any stage of software expansion. Thriving software businesses emphasize on software excellence, predominantly in the early stage of the software advancement. In succession to disable this setback, investigators have formulated various bug estimation methodologies till now. Though, emerging vigorous bug estimation prototype is a demanding assignment and several practices have been anticipated in the text. This paper exhibits a software fault estimation prototype grounded on Machine Learning (ML) Algorithms. The simulation in the paper directs to envisage the existence or non-existence of a fault, employing machine learning classification models. Five supervised ML algorithms are utilized to envisage upcoming software defects established on historical information. The classifiers are Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF). The assessment procedure indicated that ML algorithms can be manipulated efficiently with high accuracy rate. Moreover, an association measure is employed to evaluate the propositioned extrapolation model with other methods. The accumulated conclusions indicated that the ML methodology has an improved functioning.**

Keywords*—Software bug prediction; prediction model; data mining; machine learning; Naïve Bayes (NB); support vector machine (SVM); k-nearest neighbors (KNN); decision tree (DT); random forest (RF); python programming

I. Introduction

From the time of establishment of software expansion, defect restoration is studied as the most monotonous tasks, primarily for its in-built vagueness. Furthermore, the procedure of repairing bugs is gradual. The procedure of bug-restoration has a chief involvement in the software advancement. In order to lessen the concern of fault correction, bug estimation is examined significantly by the investigators. Numerous machine learning directed estimation prototypes are constructed and verified on several arguments.

The continuation of software faults influences considerably on software consistency, feature and upholding expense. Attaining errorless software is laborious, when the software utilized meticulously as largely there are unknown defects. Furthermore, extending software fault estimation prototype which can estimate the imperfect components in an initial stage is an actual test.

Contemporary developments rotate about the information that defects can be envisaged, widely beforehand they are identified. Significant corpuses of preceding fault information are fundamental to be proficient to envisage defects with sufficient precision. Software analytics has initiated continuous opportunities for tapping data analytics and rationalizing to enhance the feature of software. Functional analytics applies the outcomes of the software evaluation as real time data, to create valuable extrapolations.

Software defect estimation is an indispensable action in software expansion as envisaging the defects components earlier to software implementation attains the operator contentment and corrects the complete software functioning. Besides, envisaging the software fault initially increases software alteration to distinctive situations and enlarges the resource consumption.

Several methods are recommended to undertake Software fault estimation obstruction. The utmost comprehended procedures are Machine Learning procedures. Machine learning is effectively employed to build extrapolations in numerous database. Provided the enormous amount of fault database accessible currently, envisaging the occurrence of faults can be completed employing several machine learning procedures.

The application of machine learning to establish an exclusively mechanised technique of determining the act to be acquired by a business when a fault is testified was initially propositioned through Cubranic and Murphy [1]. The technique implemented helps text classification to envisage defect rigorousness. This technique functions accurately on 30% of the defects testified to creators. Sharma, Sharma and Gujral [2] apply feature selection to enhance the precision of the fault estimation prototype.

In this communication, supervised Machine Learning (ML) classifiers are employed to assess the ML potentials in Software fault estimation. The analysis examined Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF) classifier. The considered ML classifiers are directed to three distinctive database acquired from [3] and [4] mechanisms.

Further, the manuscript evaluated among Naïve Bayes (NB) classifier, Support Vector Machine (SVM) classifier, K-

Nearest Neighbors (KNN) classifier, Decision Tree (DT) classifier and Random Forest (RF) classifier and compared then on the basis of distinct assessment quantities for instance accuracy, precision, recall, F-measures and the ROC curves of the classifiers.

The manuscript is structured as per the following sequence. An examination of the associated work in Software fault estimation is exhibited in the Literature Review. An outline of the designated ML algorithms is exhibited in the Proposed Model. The database and the assessment technique is explained in the Evaluation Methodology. Investigational outcomes are depicted in Results accompanied by inferences and future works.

II. Literature Review

Formerly, various efforts in the subjects of fault estimation have been achieved. Peng He et al. performed a practical analysis on software fault estimation with a basic metric set [5]. Investigation has been performed on 34 announcements of 10 open source assignments accessible at PROMISE repository. The outcome signifies the outcome of uppermost-k metrics or minimum metric subset gives satisfactory result in comparison with standard forecasters.

Anuradha Chug et al. [6] employed three supervised and unsupervised learning algorithms for envisaging faults in software. NASA MDP database were administered by utilizing Weka tool. Various quantities such as recall and f-measure were applied to estimate the functioning of classification and clustering algorithms. Through examining distinct classification algorithms Random Forest has the maximum accuracy of MC1 database and gives maximum rate in recall, f-measure and receiver operating characteristic [ROC] curve and it specifies least amount of root mean square errors in all conditions. In an unsupervised algorithm k-means gave the smallest amount of inaccurate clustered examples and it considers least period for envisaging defects. Hammouri, A. et al [7] proposed software defect estimation prototype established on Machine Learning Techniques to envisage impending software defects created on past database and exhibited that Machine Learning techniques can be applied successfully with high precision.

Logan Perreault et al. [8] employed classification algorithm for instance naïve bayes, neural networks, support vector machine, linear regression, K-nearest neighbor to discover and envisage faults. The investigators manipulated NASA and tera PROMISE database. To compute the accomplishment, they tapped accuracy and f1 measure with noticeably distinct metrics.

R. Malhotra in [9] exhibited a valuable methodical evaluation for software fault estimation procedures applying Machine Learning. The article encompassed an evaluation of all the findings concerning the interval of 1991 and 2013, examined the Machine Learning methods for software fault estimation prototypes, and evaluated their functioning, matched among Machine Learning and statistic methods, evaluated among distinct Machine Learning methods and reviewed the power and the limitation of the Machine Learning methods.

Singh and Chug [10] examined widespread Machine Learning algorithms tapped for software fault estimation. The analysis exhibited significant outcomes comprising that the Artificial Neural Network has least inaccuracy amount, but the linear classifier is advanced than auxiliary algorithms in term of fault estimation precision.

Malhotra and Singh [11] indicated that the Area Under Curve is constructive metric and utilised to envisage the defects in initial stages of software expansion and to increase the validity of Machine Learning methods.

This article examines established machine learning techniques Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF). The communication estimates the Machine Learning classifiers by means of different performance quantities. Three known database are employed to assess the Machine Learning classifiers.

Alternatively, maximum cited assignments examined new Machine Learning techniques and distinct database. Few of the earlier investigations primarily concentrated on the metrics that generate the Software fault estimation as feasible as imaginable, though earlier investigations suggested distinctive approaches to estimate software faults in place of Machine Learning methods.

III. Proposed Model

The research directs to examine and assess supervised Machine Learning algorithms. The investigation exhibits the performance correctness and competency of the Machine Learning algorithms in software fault estimation and postulates a comparative study of the designated Machine Learning algorithms.

The supervised machine learning algorithms attempt to create an extrapolating function through deducing associations and needs among the identified feed in and outturn of the categorized training data, thus we can envisage the outturn amounts for recent feed in data created on the resulting extrapolating function. Subsequently are encapsulated explanations of the designated supervised Machine Learning algorithms:

- Naïve Bayes (NB): Naïve Bayes classifier functions on the theory of probability. The notion of naïve bayes classifier is established on the effort of Thomas Bayes (1702-1761) of Bayes Theorem for conditional probability. Naïve Baye's Classifier performs on the notion of baye's theorem through a naïve theory that an existence of a specific feature in a class is entirely discrete to the existence of additional features.

- Support Vector Machine (SVM): SVM is most widespread supervised machine learning technique which is equivalently tapped for classification and regression, however SVM is typically utilised for classification. The notion of SVM is to obtain a hyperplane that categorizes the training data points in order to obtain marked classes. The feed in of SVM is the training data and it functions the training sample feature to envisage category of test feature.

- K Nearest Neighbour (KNN): KNN algorithm well -known by K-Nearest Neighbours Algorithm is tapped to elucidate the difficulties of classification together with regression. The theory of algorithm is primarily established upon feature comparison in two of them, classification and regression. KNN classifier is distinct from previous probabilistic classifiers as the simulation encompasses a discovering phase of calculating probabilities from a training experiment and employ them for impending estimation of a test experiment. In probability established prototype when the prototype is proficient the training experiment could be dropped and classification is completed by means of the calculated probabilities.

- Decision Tree (DT): DT is a familiar investigation technique utilised in data mining. Decision Tree signifies a hierarchal and extrapolative prototype that utilises the elements examination as branches to access the elements target amount in the leaf. Decision Tree is a tree with decision nodes, that have several branches and leaf nodes that characterise the conclusion.

- Random Forest (RF): Random Forest comprises of a substantial quantity of distinct decision trees that function as an ensemble. Individual tree in the random forest separate out a class estimation. The class that has the highest votes turns out to be prototypes estimation. A big quantity of comparatively disjointed prototypes (trees) functioning as a group will outshine any of the specific prototypes.

IV. Evaluation Methodology

The database used in the study are three different databases, specifically DB1, DB2 and DB3. All databases comprise of two measures; the amount of defects (Bi) and the amount of test workers (Wi) for respective day (Ti) in a section of software launches period. The DB1 database has 46 quantities that were included in the examining procedure exhibited in [4]. DB2, captured from [4], computed a technique where defects for the period of 111 consecutive days of examining the software technique. DB3 includes 109 quantities. DB3 is established in [3], that comprises actual calculated records for a restoration plan of a real time control utilization exhibited in [12]. Tables I to III show DB1, DB2 and DB3, respectively.

TABLE I. The First Software Defects Database

The first software defects database	**DB1**		
	Ti	***Bi***	***Wi***
	1	2	75
	2	0	31
	3	30	63
	4	13	128
	5	13	122
	6	3	27
	7	17	136
	8	2	49
	9	2	26
	10	20	102
	11	13	53
	12	3	26
	13	3	78
	14	4	48
	15	4	75
	16	0	14
	17	0	4
	18	0	14
	19	0	22
	20	0	5
	21	0	9
	22	30	33
	23	15	118
	24	2	8
	25	1	15
	26	7	31
	27	0	1
	28	22	57
	29	2	27
	30	5	35
	31	12	26
	32	14	36
	33	5	28
	34	2	22
	35	0	4
	36	7	8
	37	3	5
	38	0	27
	39	0	6
	40	0	6
	41	0	4
	42	5	0
	43	2	6
	44	3	5
	45	0	8
	46	0	2

TABLE II. The Second Software Defects Database

The second software defects database	**DB2**		
	Ti	***Bi***	***Wi***
	1	5	4
	2	5	4
	3	5	4

The second software defects database	**DB2**		
	Ti	***Bi***	***Wi***
	4	5	4
	5	6	4
	6	8	5
	7	2	5
	8	7	5
	9	4	5
	10	2	5
	11	31	5
	12	4	5
	13	24	5
	14	49	5
	15	14	5
	16	12	5
	17	8	5
	18	9	5
	19	4	5
	20	7	5
	21	6	5
	22	9	5
	23	4	5
	24	4	5
	25	2	5
	26	4	5
	27	3	5
	28	9	6
	29	2	6
	30	5	6
	31	4	6
	32	1	6
	33	4	6
	34	3	6
	35	6	6
	36	13	6
	37	19	8
	38	15	8
	39	7	8
	40	15	8
	41	21	8
	42	8	8
	43	6	8
	44	20	8
	45	10	8
	46	3	8
	47	3	8
	48	8	4
	49	5	4
	50	1	4
	51	2	4
	52	2	4
	53	2	4
	54	7	4
	55	2	4
	56	0	4
	57	2	4
	58	3	4
	59	2	4
	60	7	4
	61	3	4
	62	0	4
	63	1	4
	64	0	4
	65	1	4
	66	0	4
	67	0	4
	68	1	3
	69	1	3
	70	0	3
	71	0	3
	72	1	3
	73	1	4
	74	0	4
	75	0	4
	76	0	4
	77	1	4
	78	2	2
	79	0	2
	80	1	2
	81	0	2
	82	0	2
	83	0	2
	84	0	2
	85	0	2
	86	0	2
	87	2	2
	88	0	2
	89	0	2
	90	0	2
	91	0	2
	92	0	2
	93	0	2
	94	0	2
	95	0	2
	96	1	2
	97	0	2
	98	0	2
	99	0	2

The second software defects database	**DB2**		
	Ti	***Bi***	***Wi***
	100	1	2
	101	0	1
	102	0	1
	103	1	1
	104	2	1
	105	0	1
	106	1	2
	107	0	2
	108	0	1
	109	1	1
	110	0	1
	111	1	1

TABLE III. THE THIRD SOFTWARE DEFECTS DATABASE

The Third Software Defects Database	**DS3**		
	Ti	***Bi***	***Wi***
	1	4	1
	2	0	1
	3	7	1
	4	10	1
	5	13	1
	6	8	1
	7	13	1
	8	4	1
	9	7	1
	10	8	1
	11	1	1
	12	6	1
	13	13	1
	14	7	1
	15	9	1
	16	8	2
	17	5	2
	18	10	2
	19	7	2
	20	11	2
	21	5	2
	22	8	2
	23	13	2
	24	9	2
	25	7	2
	26	7	2
	27	5	2
	28	7	2
	29	6	1
	30	6	1
	31	4	1
	32	12	2
	33	6	2
	34	7	2
	35	8	2
	36	11	2
	37	6	2
	38	9	2
	39	7	2
	40	12	2
	41	12	2
	42	15	2
	43	14	2
	44	7	2
	45	9	2
	46	11	2
	47	5	2
	48	7	2
	49	7	2
	50	14	2
	51	13	2
	52	14	2
	53	11	2
	54	2	1
	55	4	1
	56	4	2
	57	3	2
	58	6	2
	59	6	2
	60	2	2
	61	0	1
	62	0	1
	63	3	1
	64	0	1
	65	4	1
	66	0	1
	67	1	1
	68	2	1
	69	0	2
	70	1	2
	71	2	2
	72	5	2
	73	3	2
	74	2	2
	75	1	2
	76	11	2
	77	1	2
	78	0	2

The Third Software Defects Database	**DS3**		
	Ti	***Bi***	***Wi***
	79	2	2
	80	2	2
	81	4	2
	82	1	2
	83	0	2
	84	4	2
	85	1	1
	86	1	1
	87	0	1
	88	2	3
	89	0	1
	90	0	2
	91	1	1
	92	1	1
	93	0	1
	94	0	2
	95	0	1
	96	0	1
	97	1	2
	98	0	1
	99	1	1
	100	0	1
	101	0	1
	102	0	2
	103	0	1
	104	2	1
	105	0	1
	106	1	2
	107	0	2
	108	2	2
	109	0	2

The database was subjected to pre-treatment through a recommended clustering method. The recommended clustering method indicates the data with class labels. The labels are fixed to categorize the amount of defects into six distinct classes; A, B, C, D, E and F (Table IV).

TABLE IV. AMOUNT OF EVERY CLASS AND QUANTITY OF OCCURENCES

Amount of Every Class and Quantity of Occurrences				
Fault Class	***Number of Faults***	***DB1***	***DB2***	***DB3***
A	0-4	30	77	57
B	5-9	5	22	33
C	10-14	5	4	18
D	15-19	2	3	1
E	20-24	2	3	0
F	More than 25	2	2	0

To estimate the functioning of utilising Machine Learning algorithms in software fault extrapolation, we tapped an array of prominent quantities [13] established on the created confusion matrices. The subsequent subdivisions explain the confusion matrix and the tapped estimation quantities.

a) Confusion Matrix: The confusion matrix is an explicit table employed to determine the functioning of Machine Learning algorithms. Fig. 1 to 6 exhibits an illustration of a standard confusion matrix. Every row of the matrix signifies the occurrences in an actual class, although every column signifies the occurrences in a forecasted class. Confusion matrix recapitulates the outcomes of the examining algorithm and specifies a description of the amount of True Positive (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

Confusion Matrix for the Training Data - Decision Tree.

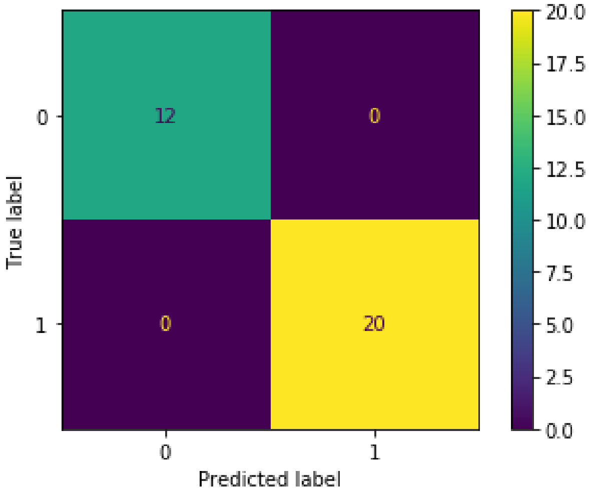


Fig. 1. Confusion Matrix of Training Data for DB1.

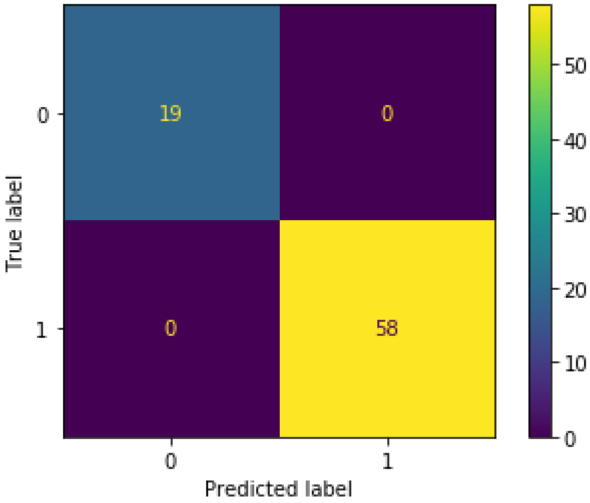


Fig. 2. Confusion Matrix of Training Data for DB2.

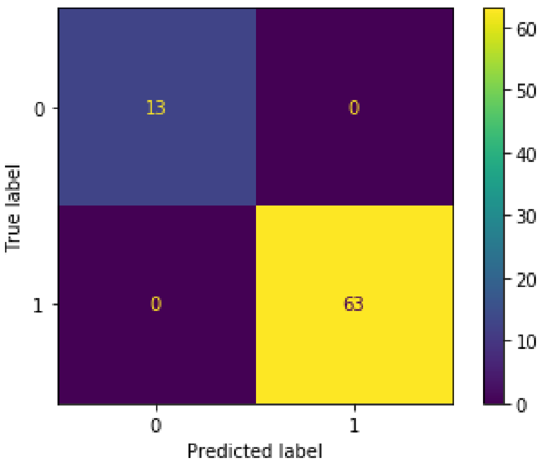


Fig. 3. Confusion Matrix of Training Data for DB3.

Confusion Matrix for the Testing Data - Decision Tree.

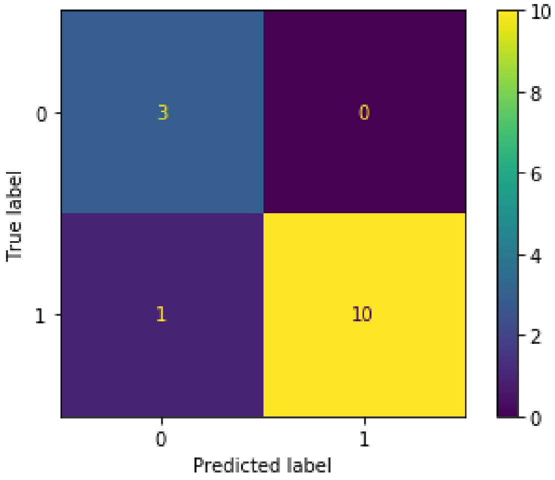


Fig. 4. Confusion Matrix of Testing Data for DB1.

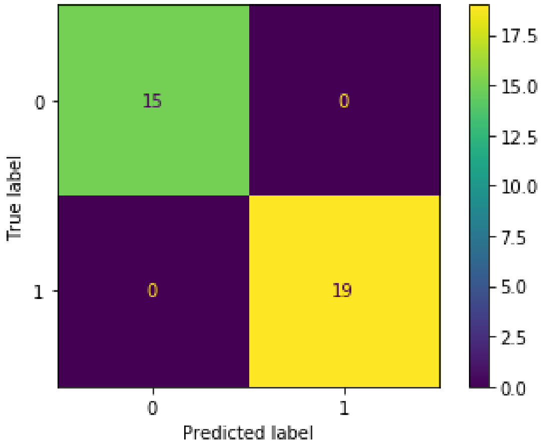


Fig. 5. Confusion Matrix of Testing Data for DB2.

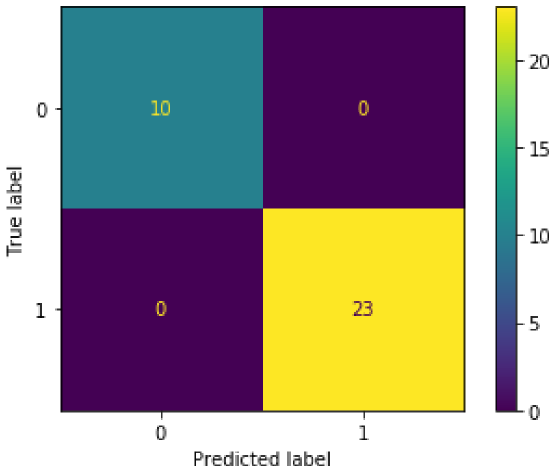


Fig. 6. Confusion Matrix of Testing Data for DB3.

b) Accuracy: Accuracy is the quantity of accurate outcomes between the total amount of inspected occurrences. The highest accuracy is one, while the poorest accuracy is zero. Accuracy could be calculated through the subsequent rule (Table V):

ACC = (TP + TN) / (TP + TN+ FP + FN)

TABLE V. ACCURACY FOR THE DATABASES

Accuracy for the Databases					
Database	***NB***	***SVM***	***KNN***	***DT***	***RF***
DB1	1	0.71	0.79	0.93	0.93
DB2	1	0.85	0.79	1	1
DB3	1	0.70	0.70	1	1
Average	1	0.75	0.76	0.98	0.98

c) Precision: Precision is computed as the amount of true positive extrapolations divided with the total amount of positive extrapolations. The highest precision is one, while the poorest is zero and could be computed through (Table VI):

Precision = TP / (TP + FP)

d) Recall: Recall is computed as the amount of positive extrapolations divided with the total amount of positives. The highest recall is one, while the poorest is zero. Recall is evaluated through the subsequent rule (Table VII):

Recall = TP / (TP + FN)

TABLE VI. PRECISION FOR THE DATABASES

Precision for the Databases					
Database	***NB***	***SVM***	***KNN***	***DT***	***RF***
DB1	1	1	1	1	1
DB2	1	0.85	0.77	1	1
DB3	1	0.70	0.78	1	1
Average	1	0.85	0.85	1	1

TABLE VII. RECALL FOR THE DATABASES

Recall for the Databases					
Database	***NB***	***SVM***	***KNN***	***DT***	***RF***
DB1	1	0.64	0.73	0.91	0.91
DB2	1	0.89	0.89	1	1
DB3	1	1	0.78	1	1
Average	1	0.84	0.80	0.97	0.97

e) F-measure: F-measure is described by way of the weighted harmonic mean of precision and recall. Generally, it is tapped to join the Recall and Precision quantities in one quantity so as to evaluate distinct Machine Learning algorithms among each other. F-measure rule is evaluated through the subsequent rule (Table VIII):

F- measure= (2* Recall * Precision)/(Recall + Precision)

TABLE VIII. F-MEASURE FOR THE DATABASES

F-Measure for the Databases					
Database	***NB***	***SVM***	***KNN***	***DT***	***RF***
DB1	1	0.78	0.84	0.95	0.95
DB2	1	0.87	0.83	1	1
DB3	1	0.82	0.78	1	1
Average	1	0.82	0.82	0.98	0.98

f) Root-Mean-Square Error (RMSE): RMSE is a quantity for assessing the functioning of an extrapolation prototype. The perception is to compute the variation among the envisaged and the definite estimates. If the definite estimate is X and the envisaged estimate is XP then RMSE is computed by the subsequent formula:

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (Xi - XPi)^2}$$

g) Area Under Curve(AUC): AUC exemplifies the probability that the classifier would rank an arbitrarily selected positive instance greater than an arbitrarily selected negative instance. The AUC is established on a chart of the false positive value with the true positive value. The highest value is one signifies that 100% estimation of the model is accurate, while the poorest is zero signifies that 100% estimation of the model is inaccurate. Fig. 7 to 12 exhibits an illustration of the Area Under Curve.

h) Receiver Operating Characteristic (ROC): ROC Curve is an outstanding technique of calculating the functioning of a Classification prototype. The True Positive value is plot with False Positive value for the probabilities of a classifier estimations.

AUC and ROC for the Training Data - Decision Tree

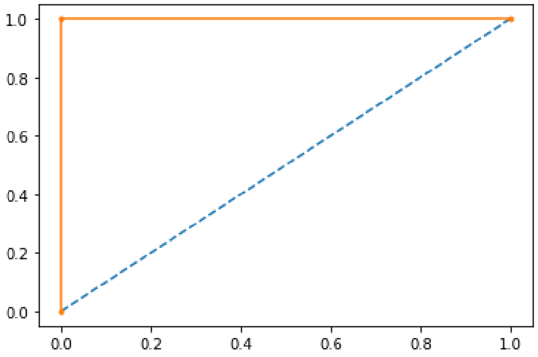


Fig. 7. AUC of Training Data for DB1.

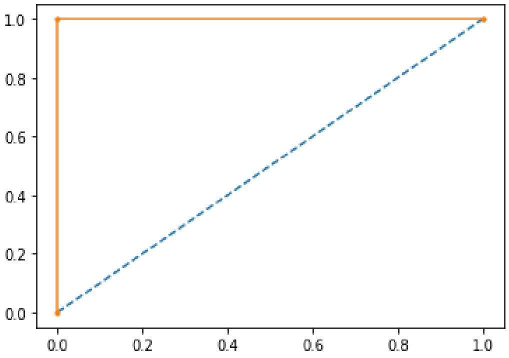


Fig. 8. AUC of Training Data for DB2.

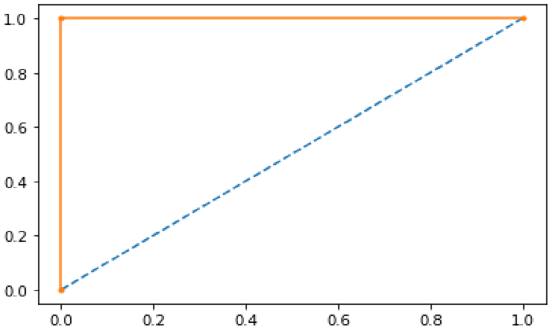


Fig. 9. AUC of Training Data for DB3.

AUC and ROC for the Testing Data - Decision Tree.

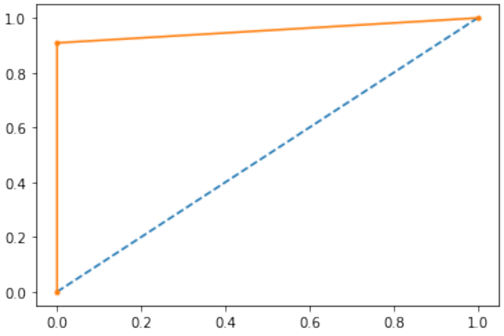


Fig. 10. AUC of Testing Data for DB1.

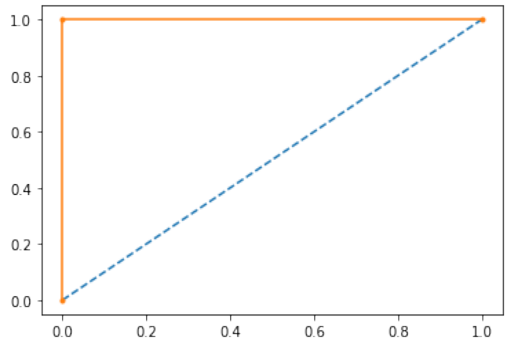


Fig. 11. AUC of Testing Data for DB2.

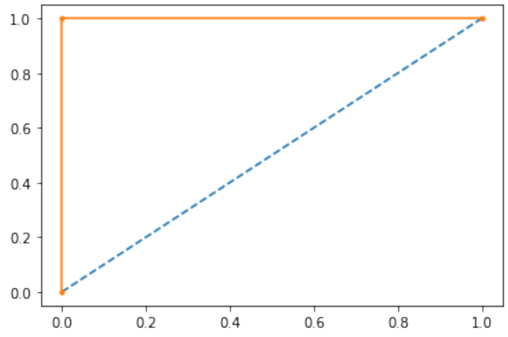


Fig. 12. AUC of Testing Data for DB3.

Conclusively, to assess the Machine Learning algorithms with additional methods, the RMSE value is estimated. The composition in [2] anticipated a Linear Regression (LR) prototype to envisage the increasing amount of software defects utilising past calculated defects. The assessment procedure was performed on the similar database that is used in this investigation. The lesser the RMSE amount, the reliable the model. Table IX exhibits the RMSE values for all the ML Algorithms and LR models.

TABLE IX. RMSE VALUES FOR THE ML ALGORITHMS AND LR MODELS

RMSE Values for the ML Algorithms and LR Models						
Database	***NB***	***SVM***	***KNN***	***DT***	***RF***	***LR***
DB1	0.0	0.53	0.53	0.26	0.26	0.43
DB2	0.0	0.38	0.38	0.0	0.0	0.38
DB3	0.0	0.55	0.55	0.0	0.0	0.36

V. Results

This inquiry utilised Jupyter Notebook, Python as Machine Learning tool, to assess five Machine Learning Algorithms Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF)) in software default estimation.

The accuracy of Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF) classifiers for the three database are presented in Table VI. As exhibited in Table VI, the five Machine Learning algorithms attained a high accuracy value.

The typical estimate for the accuracy value in all database for the five classifiers is over 75% on average. Though, the lowermost estimate emerges for SVM and KNN algorithm in the DS3 database. This is for the reason that the database does not have greater than 20 defects and SVM and KNN algorithm requires a significant quantity of defects so as to attain a better accuracy rate. Thus, SVM and KNN got a greater accuracy value in DS2 database that are comparatively larger than the DS1 and DS3 database.

The precision measures for employing NB, SVM, KNN, DT and RFs classifiers on DS1, DS2 and DS3 database are exhibited in Table VII. Outcomes indicate that the five Machine Learning algorithms can be utilised for defect extrapolation successfully with a right precision value. The typical precision rates for every classifier in the three database are greater than 85%.

The next assessment quantity is the amount of recall. Table VIII exhibits the recall rates for the five classifiers on the three database. Correspondingly, the Machine Learning algorithms attained a suitable recall rate. The highest recall rate was attained by NB classifier that is 100% in all database. Whereas, the typical recall rates for SVM, KNN, DT and RM algorithms are 84%, 80%, 97% and 97%, correspondingly.

Further, to evaluate the five classifiers concerning recall and precision quantities, we employed the F-measure rate. Table exhibits the F-measure rates for the utilised Machine Learning algorithms in the three database. As presented in the table, NB has the maximum F-measure rate in all database trailed by DT and RF then SVM and KNN classifiers.

The outcomes represent that NB, DT and RF classifiers have improved rates than LR models. The typical RMSE amount for all Machine Learning classifiers in the three database is 0.28, whereas the typical RMSE estimates for LR model is 0.39.

VI. Conclusions and Future Work

Software fault estimation is a procedure in which an extrapolation prototype is generated so as to envisage the anticipated software defects created on past data. Numerous methodologies have been propositioned utilising distinct database, distinct metrics and distinct functioning quantities. This article assessed the application of Machine Learning Algorithms in software defect estimation. Five machine learning methods have been employed, Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF). The assessment procedure is applied utilising three database. Investigational outcomes are accumulated built on accuracy, precision, recall, F-measure, and RMSE quantities. Outcomes showed that the Machine Learning procedures are effective methods to envisage the impending software faults. The evaluation outcomes exhibited that the NB classifier has the greatest outcomes in comparison to others. Furthermore, investigational outcomes presented that employing Machine Learning method imparts an improved functioning for the estimation prototype in comparison to other methods, such as LR model. For future scope, new Machine Learning procedures can be adopted and an extensive assessment between them can be performed. Moreover, inserting additional software metrics in the study procedure is a feasible method to foster the correctness of the estimation model.

References

[1] D. Cubranic and G.C. Murphy, "Automatic bug triage using text classification," Proceedings of Software Engineering and Knowledge Engineering, pp. 92–97, 2004.

[2] G. Sharma, S. Sharma and S. Gujral, "A Novel Way of Assessing Software Bug Severity Using Dictionary of Critical Terms;" Procedia Computer Science, vol 70, pp. 632–639, 2015.

[3] A. Sheta and D. Rine, "Modeling Incremental Faults of Software Testing Process Using AR Models," Proceeding of 4th International Multi-Conferences on Computer Science and Information Technology (CSIT 2006), Amman, Jordan 3, 2006.

[4] Y. Tohman, K. Tokunaga, K., S. Nagase and M. Y, "Structural approach to the estimation of the number of residual software faults based on the hyper-geometric distribution model," IEEE Trans. on Software Engineering, pp. 345–355, 1989.

[5] P. He., B. Li, X. Liu, J. Chen and Y. Ma, "An empirical study on software defect prediction with a simplified metric set," Information and Software Technology, vol. 59, pp. 170-190, 2015.

[6] A. Chug and S. Dhall, "Software defect prediction using supervised learning algorithm and unsupervised learning algorithm," Confluence 2013: The Next Generation Information Technology Summit, pp. 5-10, 2013.

[7] A. Hammouri, M. Hammad, M. Alnabhan and F. Alsarayrah, "Software Bug Prediction using Machine Learning Approach," International Journal of Advanced Computer Science and Applications, vol. 9(2), pp. 78-83, 2018.

[8] L. Perreault, S. Berardinelli, C. Izurieta and J. Sheppard, "Using Classifiers for Software Defect Detection," 26th International Conference on Software Engineering and Data Engineering, SEDE, 2017.

[9] R. Malhotra, "A systematic review of machine learning techniques for software fault prediction," Applied Soft Computing, vol. 27, pp. 504-518, 2015.

[10] P. Singh and A. Chug, "Software defect prediction analysis using machine learning algorithms," 7th International Conference on Cloud Computing, Data Science & Engineering Confluence, IEEE, 2017.

[11] R. Malhotra and Y. Singh, "On the applicability of machine learning techniques for object oriented software fault prediction," Software Engineering: An International Journal, vol. 1(1), pp. 24-37, 2011.

[12] T. Minohara and Y. Tohma, "Parameter estimation of hyper-geometric distribution software reliability growth model by genetic algorithms", in Proceedings of the 6th International Symposium on Software Reliability Engineering, pp. 324–329, 1995.

[13] Olsen, L. David and Delen, "Advanced Data Mining Techniques," Springer, 1st edition, pp. 138, ISBN 3-540-76016-1, Feb 2008.

Multi-category Bangla News Classification using Machine Learning Classifiers and Multi-layer Dense Neural Network

Sharmin Yeasmin[1], Ratnadip Kuri[2], A R M Mahamudul Hasan Rana[3*]
Ashraf Uddin[4], A. Q. M. Sala Uddin Pathan[5], Hasnat Riaz[6]
Department of Computer Science and Telecommunication Engineering
Noakhali Science and Technology University, Bangladesh[1, 2, 3, 5, 6]
Department of Computer Science, American International University-Bangladesh, Bangladesh[4]

***Abstract*—Online and offline newspaper articles have become an integral phenomenon to our society. News articles have a significant impact on our personal and social activities but picking a piece of an appropriate news article is a challenging task for users from the ocean of sources. Recommending the appropriate news category helps find desired articles for the readers but categorizing news article manually is laborious, sluggish and expensive. Moreover, it gets more difficult when considering a resource-insufficient language like Bengali which is the fourth most spoken language of the world. However, very few approaches have been proposed for categorizing Bangla news articles where few machine learning algorithms were applied with limited resources. In this paper, we accentuate multiple machine learning approaches including a neural network to categorize Bangla news articles for two different datasets. News articles have been collected from the popular Bengali newspaper Prothom Alo to build Dataset I and dataset II has been gathered from the famous machine learning competition platform Kaggle. We develop a modified stop-word set and apply it in the preprocessing stage which leads to significant improvement in the performance. Our result shows that the Multi-layer Neural network, Naïve Bayes and support vector machine provide better performance. Accuracy of 94.99%, 94.60%, 95.50% has been achieved for SVM, Logistic regression and Multi-layer dense neural network, respectively.**

Keywords*—Bangla news classification; supervised learning; feature extraction; category prediction; machine learning; neural network

I. Introduction

A newspaper is known as a powerhouse of information. People get the latest information about their desired content through online or offline newspapers. Thousands of newspapers are published in different languages all over the world. Whatever happens around the world may be a thousand miles away but reaches us within a second through online news content. In the recent years, the importance of online articles has also increased rapidly due to the rapid rise and availability of smart devices. Bangla is the fourth most spoken language and vast amounts of Bangla news articles are produced every hour worldwide. Choosing the appropriate information from the sea of web is difficult as the news has no categorization based on its content. Online news websites provide subject categories and sub-categories [1] which significantly vary newspaper to newspaper. So, these might not be sufficient for fulfilling users' choice of interest. Readers like to explore news from various news sources rather than one source and recommending suitable news to the readers based on its contents can improve the readers' experience.

The paper's main motivation is to help in recommending relevant news to the Bengali online news readers using multi-category classification. Readers are only attracted to the news articles of their interest [2]. For this purpose, the readers have to explore all the news articles of different news sites to get the desired items. For example, a user interested in entertainment-related news has to go through all the news articles from various news sites and analyze information from multiple tiresome sources. A user would prefer such a system or framework that would gather news articles of interest from various news sites and access the system anywhere on any electronic device. Although frameworks are available to notify the readers about news' on their desire categories, manually categorizing thousands of online Bangla news articles is challenging. Moreover, appropriate categorization of Bangla news articles considering their content is essential for the readers and designing an automated system for this purpose is a crying need.

Several approaches have been proposed for news categorization for different languages, i.e. Indonesian [4], Hindi[5], Arabic[6][11], Spanish [7], and these approaches mainly based on traditional machine learning algorithms such as Naïve Bayes, decision tree, K-Nearest Neighbors etc. Since Bengali is morphologically rich and complex considering the large scale of alphabets, grapheme and dialects, it needs special consideration of its features in the training phase for classification on Bangla news based on its context. However, some approaches are available in Bangla language [13-16], but these researches were limited to some traditional methods and dealt with small datasets. Due to the scarcity of resources and the complex structure of Bangla text, it's been a challenging task to classify the Bangla news.

In this paper several popular machine learning models and a multi-layer dense neural network are implemented on two different datasets. Dataset I has been built of five categories called Economics, Entertainment, International, Science and Technology and, Sports containing 1425 documents from

*Corresponding Author

popular Bangla newspaper Prothom Alo available on [20] and collected a dataset named dataset II from the Kaggle website [17] which has a total of 532509 records with nineteen categories. But, 169791 records of five categories are used from that dataset in this paper. A list of Bangla stop words are built containing 875 words [21] to remove from the newspaper contents for preprocessing purpose. Similar preprocessing steps are applied for both datasets separately and achieved better accuracy for multiple machine learning models. The accuracy of 92.63% and 95.50% for dataset I and dataset II was achieved for the multi-layer dense neural network, respectively.

The remaining part of the paper is organized as follows - Section II reviews several related works on different types of news classification both for Bangla and other languages. Section III presents research methodology which describes datasets and proposed methods. Section IV depicts result analysis. Finally, this work is concluded and provides future direction in Section V.

II. Related Work

Text classification is the process of assigning labels to text according to its content. It is one of the most fundamental tasks in Natural Language Processing (NLP) with broad application such as sentiment analysis, topic labeling, spam detection, intent detection etc. Nowadays, many tasks have been conducted on this field. Especially it is done for English language as there are enough resources for English language [3]. On the other hand, there are not enough resources except English for the task because very few works have been carried out for the task. However, working on this field is also increasing day by day in recent times. Some works of text classification on non-English languages are overviewed in the following:

Naïve Bayes and Two-Phase Feature Selection Model were used to predict the test sample category for Indonesian news classification. Naive Bayes classifier is quicker and efficient than the other discriminative models. In text classification applications and experiments, Naive Bayes (Naïve Bayes) probabilistic classifier is often used because of its simplicity and effectiveness using the joint probabilities of words and categories given a document [4]. M. Ali Fauzi et al. [4] used Naïve Bayes for Indonesian news classification. Abu Nowshed Chy et al. [10] used Naïve Bayes for Bangla news classification.

Machine learning approach was used for the classification of indirect anaphora in Hindi corpus [5]. The direct anaphora has the ability to find the noun phrase antecedent within a sentence or across few sentences. But, indirect anaphora does not have explicit referent in the discourse. They suggested looking for certain patterns following the indirect anaphora and marking demonstrative pronoun as directly or indirectly anaphoric accordingly. Their focus of study was pronouns without noun phrase antecedent.

A method was designed for classification of Arabic news, the classification system that best fits data given a certain representation [6]. A new method was presented for Arabic news classification using field association words (FA words). The document preprocessing system generated the meaningful terms based on Arabic corpus and Arabic language dictionary. Then, the field association terms were classified according to FA word classification algorithm. It is customary for people to identify the field of document when they notice peculiar words. These peculiar words are referred to as Field Associating words (FA words); specifically, they are words that allow us to recognize intuitively a field of text or field-coherent passage. Therefore, to identify the field of a passage FA terms can be used, and to classify various fields among passages FA terms can be also used.

Cervino U et al. applied machine learning techniques to the automatic classification of news articles from the local newspaper La Capitaolf Rosario, Argentina [7]. The corpus (LCC) is an archive of approximately 75,000 manually categorized articles in Spanish published in 1991. They benchmarked on LCC using three widely used supervised learning methods: k-Nearest Neighbors, Naive Bayes and Artificial Neural Networks, illustrating the corpus properties.

This paper delineates the Bangla Document Categorization using Stochastic Gradient Descent (SGD) classifier [8]. Here, document categorization is the task in which text documents are classified into one or more of predefined classes based on their contents using Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. In text classification and natural language processing, SGD has been successfully applied to large-scale and sparse machine learning problems often encountered.

Fouzi Harrag, Eyas EI Qawasmah [11] used ANN for the classification of Arabic language document. In this paper Singular Value Decomposition (SVD) had been used to select the most relevant features for the classification.

Neural network was used for web page classification based on augmented PCA [12]. In this paper, each news web page was represented by term weighting schema. The principal component analysis (PCA) had been used to select the most relevant features for the classification. Then, the final output of the PCA is augmented with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks. According to this paper it's evident that, in case of Sports news, WPCM provides most acceptable classification accuracy based on their datasets. Their experiment evaluation also demonstrates the same.

A research group of Shahjalal University of Science & Technology used different machine learning based approaches of baseline and deep learning models for Bengali news categorization [13]. They used baseline models such as: Naïve Bayes, Logistic Regression, Random Forest and Linear SVM and deep learning models like BiLSTM, CNN. They found out that the highest result comes from the Support Vector Machine in the base model and CNN in deep learning where CNN gave the best performance for their Dataset.

In paper [14] authors used multi-layer dense neural network for Bangla document categorization. As feature selection technique they used TF-IDF method. They used three dense layers and 2 dropout layers. They got 85.208% accuracy.

Authors on [15] used four supervised learning methods namely Decision Tree, K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine for categorization of Bangla web documents. They also build their own dataset corpus but they didn't publish it. Their corpus included 1000 documents with a total number of words being 22,218. Their Dataset included five categories such as business, health, technology, sports and education. As feature selection they used TF-IDF method and they got 85.22% f-measure for Naïve Bayes, 74.24% for K-Nearest Neighbor, 80.65% for Decision Tree and 89.14% for Support Vector Machine.

An exploration group used Bidirectional Long Short Term Memory (BiLSTM) for classification of Bangla news articles [16]. They used Gensim and fastText model for vectorization of their text. Their Dataset contained around 1 million articles and 8 different categories. They got 85.14% accuracy for BiLSTM for their Dataset.

III. Methodology

The goal of this proposed model is to categorize Bangla news automatically based on the content of the document. In order to meet this up, some steps are performed such as 1) Data collection, 2) Data preprocessing, 3) Feature selection and extraction, 4) Dividing Dataset into training and testing set, 5) Building and fitting models, 6) Category prediction. Fig. 1 depicts an overview of the approach. The details of the steps are explained in following paragraphs.

A. *Data Collection*

Data is crucial in machine learning which required a lot of data to come up with somewhat generalizable models. The Bangla dataset corpus is built for this research task & the news articles have been collected from the popular news portal Prothom Alo online newspaper. News articles of five categories such as 'International', 'Economics', 'Entertainment', 'Sports', 'Science and Technology' has been used for the dataset. This dataset corpus consists of 1425 documents. Each category contains 285 documents, which can be found at [20]. Details of the dataset are represented in Table I.

Another dataset is also downloaded from the Kaggle website [17]. This Dataset contains newspaper articles from 2013 to 2019 from Prothom Alo. The newspaper articles have already been classified into different categories such as International, State, Economy, etc. Only five categories, namely, Entertainment, International, Economic, Sports, and Technology. In the Table II, the details and statistical analysis of the whole Dataset are given.

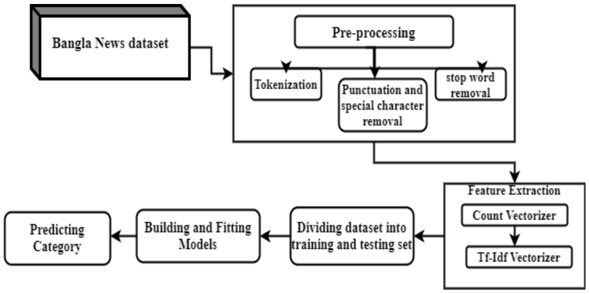


Fig. 1. Overview of Bangla News Classification System.

TABLE I. Details of Dataset I

Category	**No. of Docs**	**Words/Doc(Average)**
Economics	285	433
Entertainment	285	380
International	285	299
Science &Technology	285	381
Sports	285	349
Total	1425	367

TABLE II. Details of Dataset II

Category	**No. of Docs**	**Words/Doc**
Economics	20858	277
Entertainment	36791	237
International	37176	235
Science & Tech	15117	231
Sports	59849	261
Total	169791	250

B. *Data Pre-processing*

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Preprocessing the data is an important task and it is essential for getting better accuracy. In the experiment, the data was processed by several techniques such as removing empty data from document, tokenization, punctuation removal and stop word removal, white space removal, number removal.

1) Tokenization: Splitting a text into sentences, then words, and then characters. Based on spaces, texts are broken down into words and using the list function; words are broken down into characters.

2) Punctuation, special character and number removal: Punctuation like ; : । ' " ' , ? !, etc. and special character like @, #, $, %, ^, &, (, *,), etc. and number that is not important for classification are removed from the whole Dataset.

3) Stop word removal: High-frequency words common in every document and have not much influence in the text are called stop words. Stop words are collected from two different sources [18]&[19], and combined unique stop words and increased the number of stopwords. The stop words list that was build contains 875 stop words, and it can be found at [21]. The list of 361 bangali stop words like "অবশ্য, অনেকে, এ, এবং, ইত্যাদি, করেছিলেন, নিতে, হয়, etc." All the stop words are removed from the Dataset for getting better accuracy.

4) Categorical encoding: There are two types of categorical encoding entitled label encoding and one-hot encoding. In label encoding, each label is assigned a unique integer based on alphabetical ordering. On the other hand, each category is represented as a one-hot vector in one hot encoding. That means only one bit is hot or true at a time. An example of a one-hot encoding of a dataset with two categories is given in Table III. Label encoding technique has been used for encoding category in machine learning algorithms and one-hot encoding for multi-layer dense neural network.

TABLE III. EXAMPLE OF ONE HOT ENCODING

	Label 1	Label 2	Label 3
Doc 1	0	0	1
Doc 2	1	0	0

After the data preprocessing step, statistical analysis step is performed on both Dataset to see if data preprocessing step is successfully performed and how words are related to each category. Fig. 2 illustrates the flow chart of the data preprocessing system. Fig. 3 and Fig. 4 illustrate the 14 most frequent words of each category of Dataset I and Dataset II. It is seen that these words are strongly related to corresponding categories that help the model successfully predict a document category. After pre-processing step, structure and number of word is changed on datasets. The detailing after pre-processing step of the two dataset is given in Table IV and Table V.

C. *Feature Selection and Extraction*

In this step, string features are converted into numerical features. Bag of words and TF-IDF model are used for converting string features into numerical features for performing the mathematical operation. Dataset I consists of 43404 unique words, and Dataset II that is downloaded from the Kaggle website [17] consists of 915428 unique words after data preprocessing. All the words do not have impact on the classification. So, the most frequent words have been used as features that have importance to classification. For selecting features, a Count vectorizer was utilized, which works based on the frequencies of words. Both datasets' model accuracy are observed in the Count vectorizer approach by considering different minimum document frequencies and maximum document frequencies. And for Dataset I, the best result is found by considering minimum document frequency 10, which means the words are excluded that are only on 10 or less than 10 documents and maximum document frequency 0.6, which means the words that are on the 60% document or more than that. For Dataset II, the highest accuracy is got by considering minimum frequency 10 and maximum document frequency 0.6 because those words have no significance in determining the class. In this paper 1320 most frequent words are used as a feature vector for Dataset I, and the rest of the words are excluded. For Dataset II, 10,000 most frequent words are used as a feature vector.

After selecting features, the TF-IDF vectorizer has been used for feature extraction because count vectorizer doesn't return the proper value. As it is known, count vectorizer only returns 0 or 1 as the value of a different word which does not states the transparent frequency of different words from a document.

1) Bag of words: It is a basic model used in natural language processing. A bag-of-words is a representation of text that describes the occurrence of words within a document.

2) TF-IDF: TF-IDF stands for Term Frequency-Inverse Document Frequency which says the word's importance in the corpus or Dataset. TF-IDF contain two concept Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency is defined as how frequently the word appears in the document or corpus. Term frequency can be defined as:

TF = No. of time word appear in the doc. / Total no. of word in the doc.

Inverse document frequency is another concept that is used for finding out the importance of the word. It is based on the fact that less frequent words are more informative and essential. IDF is represented by the formula:

IDF = No of Docs / No of Docs in which the word appears

TF-IDF is a multiplication between TF and IDF value. It reduces the importance of the common word that is used in a different document. And only take important words that are used in classification. TF-IDF matrix of first 10 docs and first six words of dataset I is given in Table VI.

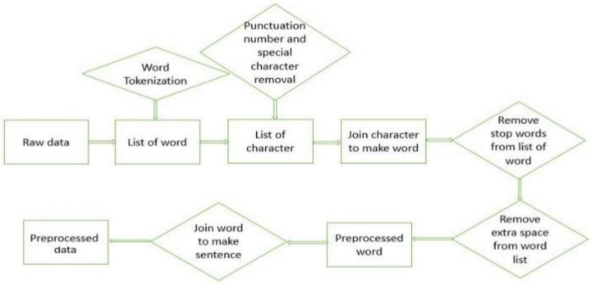

Word Tokenization
Punctuation number and special character removal
Raw data
List of word
List of character
Join character to make word
Remove stop words from list of word
Remove extra space from word list
Preprocessed word
Join word to make sentence
Preprocessed data

Fig. 2. Flow Chart of Data Preprocessing System.

TABLE IV. DATASET I DETAILS AFTER PREPROCESSING

Category	Total Words	Words/Doc	Unique Words
Economics	75379	282	12462
Entertainment	64895	243	15722
International	53174	199	11883
Science and Technology	65830	244	14381
Sports	62449	230	13300
Total	341702	240	55748

TABLE V. DATASET II DETAILS AFTER PREPROCESSING

Category	**Total Words**	**Words/Doc**	**Unique Words**
Economics	5071921	243	271773
Entertainment	4972334	135	290961
International	3319266	89	158302
Science and Technology	1971826	130	147524
Sports	8879569	148	379027
Total	24214916	142	1247587

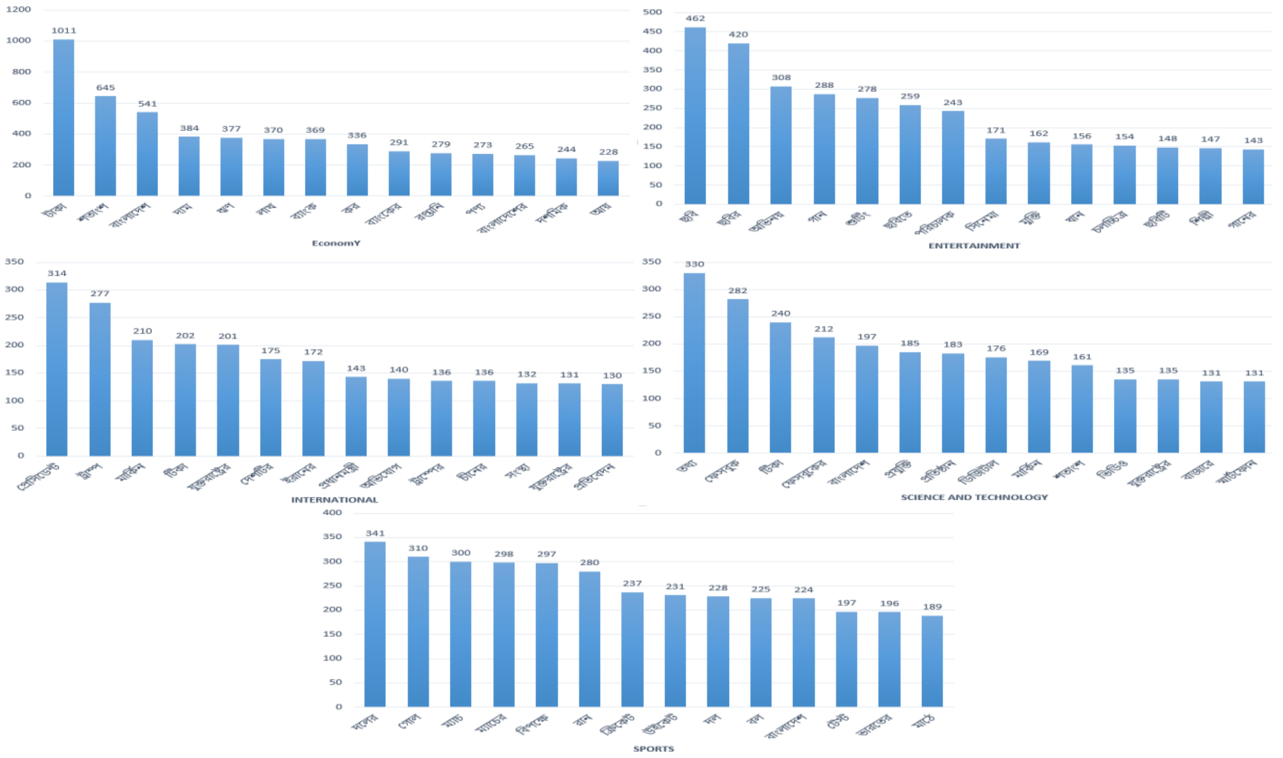


Fig. 3. Fourteen most frequent words of each category after data cleaning of Dataset I

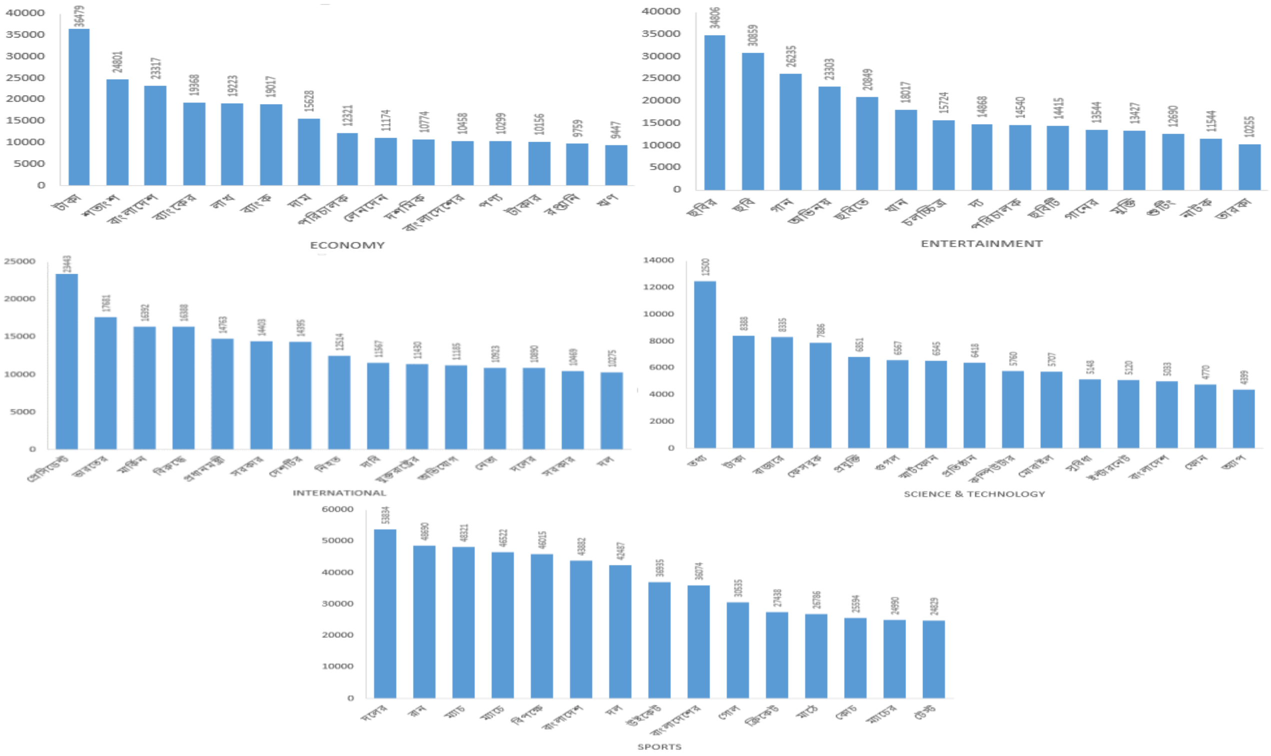


Fig. 4. Fourteen most Frequent Words of each Category after Data Cleaning of Dataset II.

TABLE VI. TF-IDF MATRIX OF FIRST TEN DOCS AND FIRST SIX WORDS OF DATASET I

	WORD 1	WORD 2	WORD 3	WORD 4	WORD 5	WORD 6
DOC 1	0	0	0	0	0	0
DOC 2	0	0	0	0	0	0
DOC 3	0	0	0.038	0	0	0
DOC 4	0	0	0	0	0	0
DOC 5	0	0.038	0	0	0	0
DOC 6	0	0	0	0	0	0
DOC 7	0.032	0.033	0	0	0	0
DOC 8	0	0	0	0	0	0
DOC 9	0	0	0	0	0	0
DOC10	0	0	0	0	0	0

E. Splitting Dataset into Training and Testing Set

After Successful feature extraction, Datasets are split into train and test datasets. Both datasets are divided into 4:1. Four portions of dataset are used for the training set, and the rest portion is for testing. That means 80% of data from the datasets are used for training, and the rest 20% is considered as the testing. This step is done using the sklearn library, which is very simple.

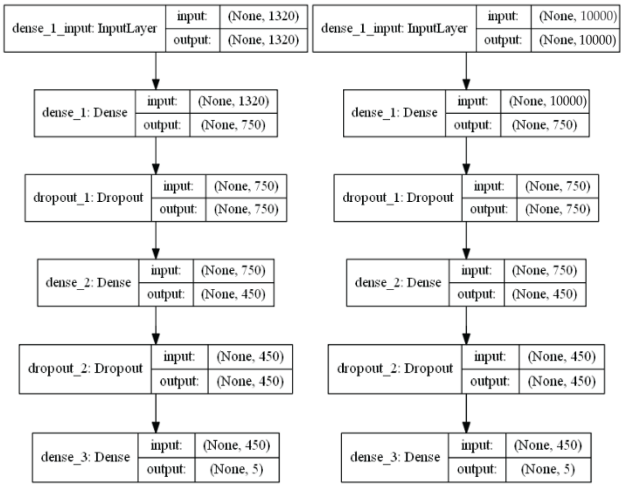
F. Building and Fitting Models

In this stage, the datasets are fitted into different machine learning classifier algorithms and neural network.

1) Using machine learning classifier algorithm: Here several machine learning classifiers used such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, Decision Tree for the classification of Bangla news & import sklearn built-in classifier for this.

2) Using multi-layer dense neural network: Here the data preprocessing technique and feature extraction technique is the same as machine learning algorithms. However, one hot encoder is used for encoding the encoding category. For categorizing task, feed forward neural network is used as classification algorithm. It is organized in the form of multiple layers. In the proposed model, dense layer has been used. Feed forward neural network consists of the input layer, the hidden layers and the output layer. Dataset generated Input patterns are transmitted from input layer to next layer which is also called by first hidden layer. Later output from the first hidden layer is being used as the input of the second hidden layer. The same process continueing untill reach the last hidden layer. Finally, the output of the last hidden layer is being used as the input of output layer or last layer. For building such model, Sequential model has been used. This model uses a linear stack of layers. The most common layer is a dense layer which is a regular densely connected neural network layer with all the weights and biases. In the first layer input shape is determined since the following layers can make automatic shape inference. To build the Sequential model, layers one by one are added in order. Total three dense layers used. After each dense layer, one dropout layer added with a 20% dropout rate.

For the layers between the input and output layer, relu activation function is used, and in the output layer, the softmax function is used as an activation function. Before training the model, the learning process is configured using optimizer and loss function. Here adam optimizer and categorical_crossentropy is used as optimizer and loss function respectively to train the model. Different numbers of dense layers, different numbers of nodes between layers, and different epochs are used for checking the accuracy of the model. Finally, the model is built by adding three dense layers. On the first layer, 750 nodes are added, on the second layer 450 nodes, and on the third layer, 5 nodes are added as the dataset is of five categories. The multi-layer dense neural network model for the two datasets is depicted in Fig. 5. The model is trained in 200 epochs. By doing so, the highest accuracy is got for both Datasets. Step by step procedure that is done for building a multilayer dense neural network model for getting the highest accuracy Bangla news classification is represented in Fig. 6. How the training loss and accuracy changes in both Dataset is represented in Fig. 7 and Fig. 8, respectively.



(a) For dataset I (b) For dataset II

Fig. 5. Multi-layer Dense Neural Network Model.

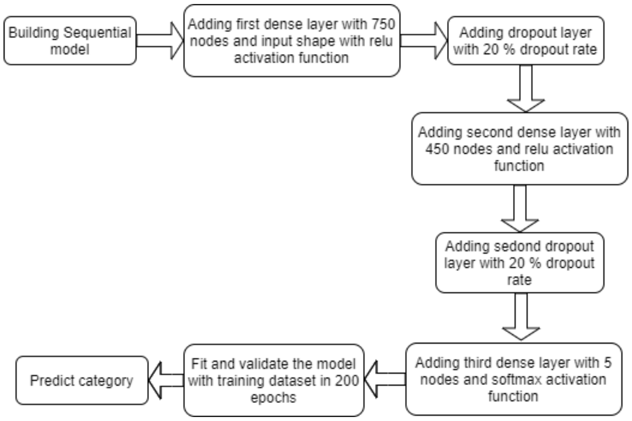

Building Sequential model
Adding first dense layer with 750 nodes and input shape with relu activation function
Adding dropout layer with 20 % dropout rate
Adding second dense layer with 450 nodes and relu activation function
Adding sedond dropout layer with 20 % dropout rate
Adding third dense layer with 5 nodes and softmax activation function
Fit and validate the model with training dataset in 200 epochs
Predict category

Fig. 6. Step by Step Procedure of Feed-forward Multi-layer Dense Neural Network for Bangla News Classification.

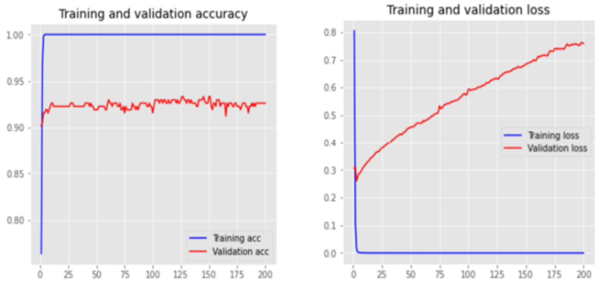


Fig. 7. Training and Validation Accuracy and Loss for Dataset I.

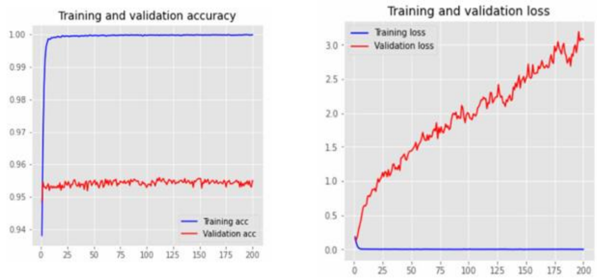
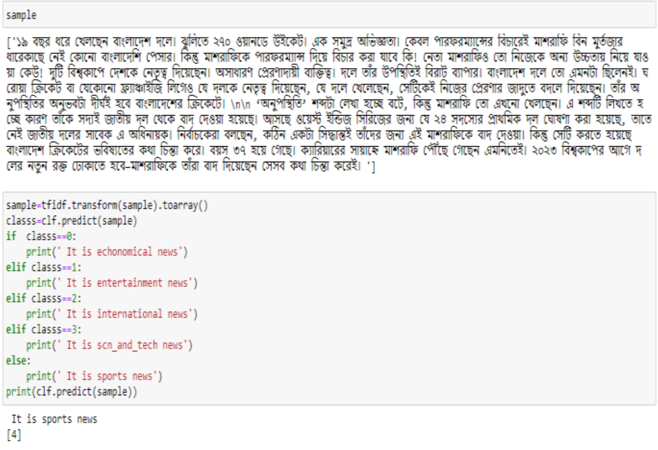


Fig. 8. Training and Validation Accuracy and Loss for Dataset II.

G. Category Prediction

After fitting Dataset to classifier, the main job is to predict category. In this stage, the model is trained to predict the test data that are unseen to the machine. If any vectorized sample of Bangla news is given to the model, it can predict the category of the sample data. Here for training, as different classifiers are used such as Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and also the model which is made by neural network can predict the category of vectorized sample data.

The model and vectorizer are saved as a pickle file and then it is used to classify. A random sample is given to the model and it successfully classifies the random sample. A screenshot of the random testing is given in Fig. 9.



```
sample
```

```
['১৯ বছর ধরে খেলছেন বাংলাদেশ দলে। ঝুলিতে ২৭০ ওয়ানডে উইকেট। এক সমুদ্র অভিজ্ঞতা। কেবল পারফরম্যান্সের বিচারেই মাশরাফি বিন মুর্তজার
ধারেকাছে নেই কোনো বাংলাদেশি পেসার। কিন্তু মাশরাফিকে পারফরম্যান্স দিয়ে বিচার করা যাবে কি! নেতা মাশরাফিও তো নিজেকে অন্য উচ্চতায় নিয়ে যাও
য়া কেউ! দুটি বিশ্বকাপে দেশকে নেতৃত্ব দিয়েছেন। অসাধারণ প্রেরণাদায়ী ব্যক্তিত্ব। দলে তাঁর উপস্থিতিই বিরাট ব্যাপার। বাংলাদেশ দলে তো এমনটা ছিলেনই। ঘ
রোয়া ক্রিকেট বা যেকোনো ফ্র্যাঞ্চাইজি লিগেও যে দলকে নেতৃত্ব দিয়েছেন, যে দলে খেলেছেন, সেটিকেই নিজের প্রেরণার জাদুতে বদলে দিয়েছেন। তাঁর অ
নুপস্থিতির অনুভবটা দীর্ঘই হবে বাংলাদেশের ক্রিকেটে। \n\n ‘অনুপস্থিতি’ শব্দটা লেখা হচ্ছে বটে, কিন্তু মাশরাফি তো এখনো খেলছেন। এ শব্দটি লিখতে হ
চ্ছে কারণ তাঁকে সদ্যই জাতীয় দল থেকে বাদ দেওয়া হয়েছে। আসছে ওয়েস্ট ইন্ডিজ সিরিজের জন্য যে ২৪ সদস্যের প্রাথমিক দল ঘোষণা করা হয়েছে, তাতে
নেই জাতীয় দলের সাবেক এ অধিনায়ক। নির্বাচকেরা বলছেন, কঠিন একটা সিদ্ধান্তই তাঁদের জন্য এই মাশরাফিকে বাদ দেওয়া। কিন্তু সেটি করতে হয়েছে
বাংলাদেশ ক্রিকেটের ভবিষ্যতের কথা চিন্তা করে। বয়স ৩৭ হয়ে গেছে। ক্যারিয়ারের সায়াহ্নে মাশরাফি পৌঁছে গেছেন এমনিতেই। ২০২৩ বিশ্বকাপের আগে দ
লের নতুন রক্ত ঢোকাতে হবে—মাশরাফিকে তাঁরা বাদ দিয়েছেন সেসব কথা চিন্তা করেই। ']
```

```
sample=tfidf.transform(sample).toarray()
classs=clf.predict(sample)
if  classs==0:
    print(' It is echonomical news')
elif classs==1:
    print(' It is entertainment news')
elif classs==2:
    print(' It is international news')
elif classs==3:
    print(' It is scn_and_tech news')
else:
    print(' It is sports news')
print(clf.predict(sample))
```

```
 It is sports news
[4]
```

Fig. 9. Screenshot of Output of Successfully Classifying given Sample.

IV. Result and Discussion

In this section, the performances of the model are analyzed on different machine learning algorithms and neural network for both Datasets.

A. Accuracy of Model

Onfusion matrix is a presentation for summarizing the performance of a classification algorithm. The confusion matrix for all classifier algorithms is given in Fig. 10 and Fig. 11 for Dataset I and Dataset II, respectively. By judging the confusion matrix the best model can be decided. From this matrix, the accuracy, precision, recall, and f1-score of the built model can be calculated. For different classifiers, different confusion matrices are built and from those confusion matrices, the accuracy, precision, recall, and f1-score of different classifiers is calculated. The performances of the different models are represented in Tables VII and VIII for dataset I and dataset II respectively and the highest precision, recall and accuracy category wise for all the classifiers are shown as bold font. Overall performance of different classifier model is shown in Tables IX and X for both dataset respectively. Here, highest accuracy, precision and recall, and f1-score of algorithms are shown also as bold font. Fig. 12 is a plot of the accuracy of different classifiers for both datasets. F1-score of classifiers according to news type is shown on Fig. 13 and Fig. 14 for Dataset I and Dataset II respectively.

1) Naïve bayes: Both dataset have five categories. From Tables VII and VIII it is clear that there are variations in different performance rate for different types of news. For dataset I Entertainment has the lowest f1-score and for dataset II sports has the lowest f1-score when Naïve Bayes classifier is used. If all the categories are combined the accuracy for Naïve Bayes model is 91.23% and 92.76% for dataset I and dataset II, respectively.

2) K-Nearest neighbor: The accuracy of this model is 84.81 % for dataset I and 70.4% for dataset II. The lowest performance rate is found for entertainment category in dataset I and the same is international category in dataset II.

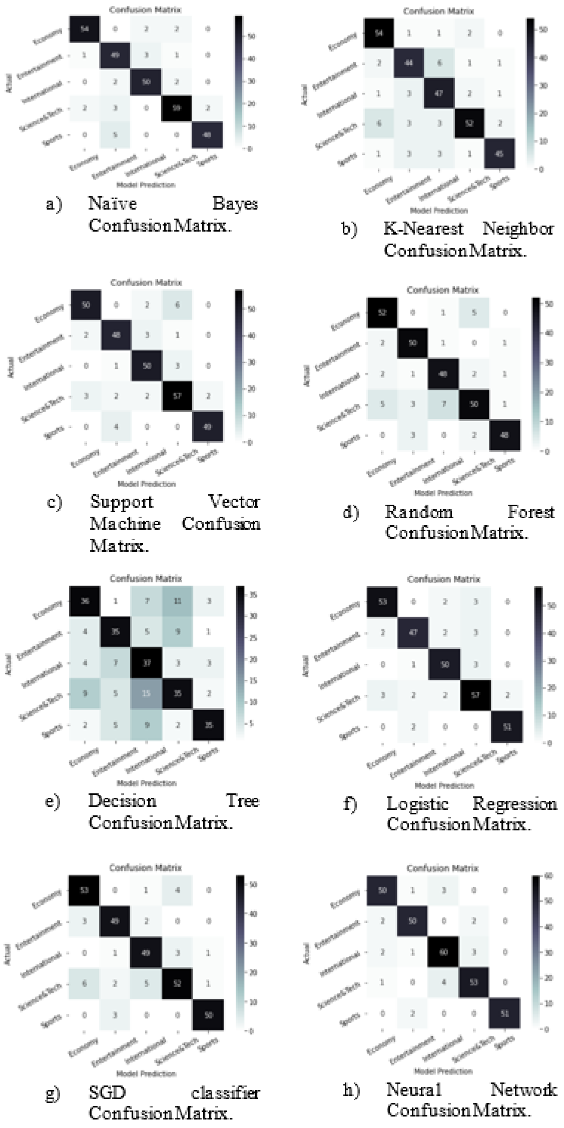
3) Support vector machine: Support Vector Machine can be defined over a vector space where the problem is for finding a decision surface that “best” separates the data points in two classes [9]. Support Vector Machine has different types of kernels. The linear kernel is used for the purpose. Overall accuracy of this model is 89.12% for dataset I and 94.99% for dataset II. Here science and technology category shows the lowest performance rate for dataset I and for dataset II sports category shows the lowest performance.

4) Random forest: For building Random Forest classifier model entropy criterion and 50 n_estimators are used here and got accuracy 87.01% for dataset I and 91.4% for dataset II. In the prediction model the lowest rate of performance is found in science and technology category for dataset I and sports for dataset II.

5) Decision tree: The overall accuracy of this model is 62.45% for dataset I and for dataset II accuracy is 79.87%. Again for dataset I science and technology has the lowest rate of performance and for dataset II sports has the lowest rate of performance in this prediction model. On the other hand the highest rate of performance is found in sports category for dataset I and science and technology is for dataset II.

6) Logistic regression: Again, the lowest accuracy is found for science and technology category in dataset I and sports category in dataset II in this model.. The highest rate of performance in the prediction model is sports for dataset I and science and technology for dataset II. The overall accuracy of this model is 90.52 % for dataset I and 94.6% for dataset II.

7) SGD classifier: If all the categories are combined to get the accuracy of the SGD classifier, the accuracy is 88.77% and 93.78% for dataset I and dataset II, respectively. The SGD confusion matrix is shown in Fig. 10(g) and 11(g) for both dataset which shows which category has high performance. For dataset I sports has the highest f1-score and science and technology category has the lowest f1-score and for dataset II science and technology has the highest f1-score and sports has the lowest f1-score.



a) Naïve Bayes Confusion Matrix.

b) K-Nearest Neighbor Confusion Matrix.

c) Support Vector Machine Confusion Matrix.

d) Random Forest Confusion Matrix.

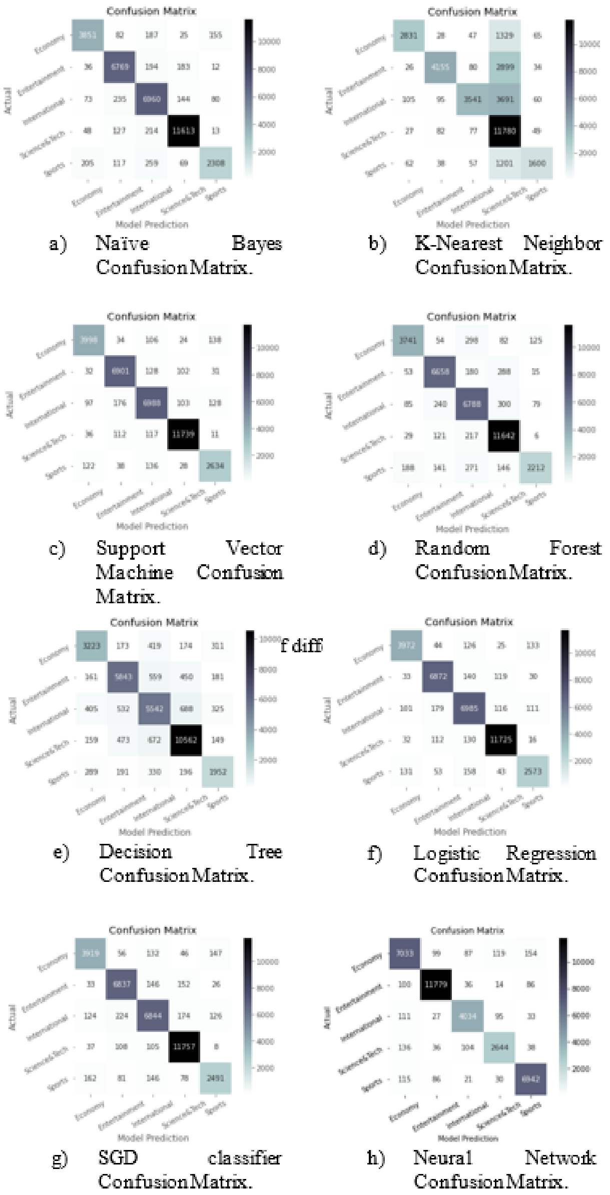
e) Decision Tree Confusion Matrix.

f) Logistic Regression Confusion Matrix.

g) SGD classifier Confusion Matrix.

h) Neural Network Confusion Matrix.

Fig. 10. Confusion Matrix of different Classifiers for Dataset I.



a) Naïve Bayes Confusion Matrix.

b) K-Nearest Neighbor Confusion Matrix.

c) Support Vector Machine Confusion Matrix.

d) Random Forest Confusion Matrix.

e) Decision Tree Confusion Matrix.

f) Logistic Regression Confusion Matrix.

g) SGD classifier Confusion Matrix.

h) Neural Network Confusion Matrix.

Fig. 11. Confusion Matrix of different Classifiers for Dataset II.

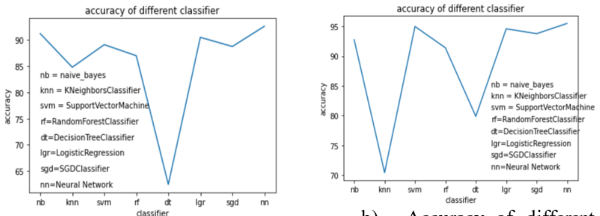
TABLE VII. RESULT COMPARISON BETWEEN MACHINE ALGORITHMS OF DATASET I

Classifiers	Category	Precision	Recall	F1-score
Naïve Bayes	Economics	0.94	**0.93**	**0.93**
	Entertainment	0.83	0.90	0.86
	International	0.91	0.92	0.91
	Science&Tech	0.92	0.89	0.90
	Sports	**0.96**	0.90	**0.93**
K-Nearest Neighbor	Economics	0.84	**0.93**	**0.88**
	Entertainment	0.81	0.81	0.81
	International	0.78	0.87	0.82
	Science&Tech	0.89	0.78	0.83
	Sports	**0.91**	0.85	0.87

Support Vector Machine	Economics	0.90	0.86	0.88
	Entertainment	0.87	0.88	0.87
	International	0.87	**0.92**	0.89
	Science&Tech	0.85	0.86	0.85
	Sports	**0.96**	**0.92**	**0.94**
Random Forest	Economics	0.85	0.89	0.87
	Entertainment	0.87	**0.92**	0.89
	International	0.84	0.89	0.86
	Science&Tech	0.84	0.75	0.79
	Sports	**0.94**	0.90	**0.92**
Decision Tree	Economics	0.65	0.62	0.63
	Entertainment	0.66	0.65	0.65
	International	0.51	**0.68**	0.58
	Science&Tech	0.58	0.53	0.55
	Sports	**0.79**	0.66	**0.72**
Logistic Regression	Economics	0.91	0.91	0.91
	Entertainment	0.90	0.87	0.88
	International	0.89	0.92	0.90
	Science&Tech	0.86	0.86	0.86
	Sports	**0.96**	**0.96**	**0.96**
SGD Classifier	Economics	0.85	0.91	0.88
	Entertainment	0.89	0.90	0.89
	International	0.85	0.90	0.87
	Science&Tech	0.88	0.78	0.82
	Sports	**0.96**	**0.94**	**0.95**
Multi-layer Dense Neural Network	Economics	0.91	0.93	0.92
	Entertainment	0.93	0.93	0.93
	International	0.90	0.91	0.90
	Science&Tech	0.91	0.91	0.91
	Sports	**1.00**	**0.96**	**0.98**

TABLE VIII. RESULT COMPARISON BETWEEN MACHINE LEARNING ALGORITHMS OF DATASET II

Classifiers	Category	Precision	Recall	F1-score
Naïve Bayes	Economics	0.91	0.89	0.90
	Entertainment	0.92	0.94	0.93
	International	0.89	0.93	0.91
	Science&Tech	**0.96**	**0.96**	**0.96**
	Sports	0.90	0.78	0.83
K-Nearest Neighbor	Economics	0.92	0.66	**0.77**
	Entertainment	**0.94**	0.57	0.71
	International	0.93	0.47	0.62
	Science&Tech	0.56	**0.98**	0.71
	Sports	0.88	0.54	0.67
Support Vector Machine	Economics	0.93	0.93	0.93
	Entertainment	0.95	0.96	0.95
	International	0.93	0.93	0.93
	Science&Tech	**0.97**	**0.97**	**0.97**
	Sports	0.89	0.89	0.89
Random Forest	Economics	0.91	0.87	0.89
	Entertainment	0.92	0.92	0.92
	International	0.87	0.90	0.88
	Science&Tech	**0.93**	**0.97**	**0.95**
	Sports	0.90	0.74	0.81
Decision Tree	Economics	0.76	0.75	0.75
	Entertainment	0.81	0.81	0.81
	International	0.73	0.74	0.73
	Science&Tech	**0.87**	**0.88**	**0.87**
	Sports	0.67	0.66	0.66
Logistic Regression	Economics	0.93	0.92	0.92
	Entertainment	0.94	0.95	0.94
	International	0.92	0.93	0.92
	Science&Tech	**0.97**	**0.97**	**0.97**
	Sports	0.90	0.87	0.88
SGD Classifier	Economics	0.91	0.91	0.91
	Entertainment	0.93	0.95	0.94
	International	0.93	0.91	0.92
	Science&Tech	**0.96**	**0.98**	**0.97**
	Sports	0.89	0.84	0.86
Multi-layer Dense Neural Network	Economics	0.94	0.94	0.94
	Entertainment	**0.98**	**0.98**	**0.98**
	International	0.94	0.94	0.94
	Science&Tech	0.91	0.89	0.90
	Sports	0.96	0.96	0.96



a) Accuracy of different classifier for Dataset I

b) Accuracy of different classifier for Dataset II

Fig. 12. Accuracy of different Classifier for both Dataset.

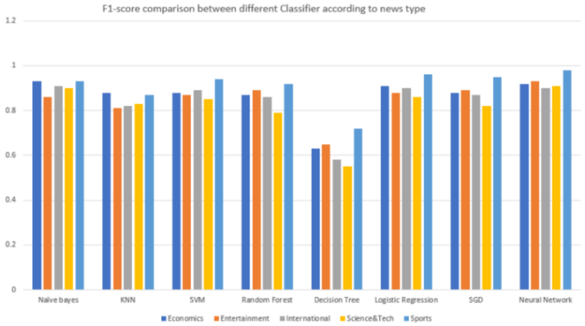


Fig. 13. Comparison of f1-score of Dataset I of different Classifier according to the News Types.

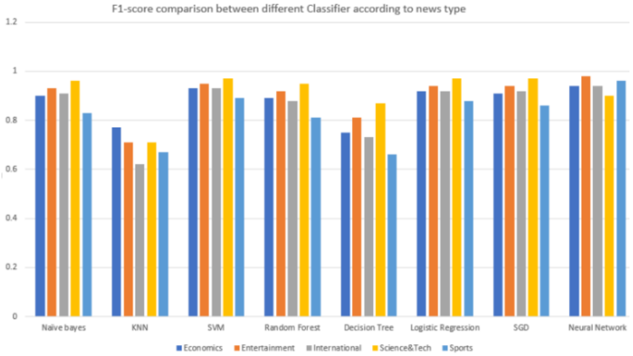


Fig. 14. Comparison of f1-score of Dataset II of different Classifier according to the News Types.

8) Multi-layer dense neural network: In this case, international news for dataset I and science and technology news for dataset II have the lowest rate of performance. In Machine learning model it can be seen that, In dataset I, all classifiers returned lowest performance in science & Technology category, except Naive Bayes and K-Nearest Neighbor. Where in dataset II, all classifiers gives lowest performance in sports category where only K-nearest neighbour gives lowest performance on other category. But ,it doesn't happen in multi-layer dense neural network model. The accuracy of this model is also quite impressive. For dataset I sports and for dataset II, entertainment has the highest rate of performance. This model has the highest accuracy comparing the other traditional machine learning models. The overall accuracy of this model is 92.63% for dataset I and 95.50% for dataset II.

B. *Comparison of Algorithms*

Table IX shows that in the traditional machine learning algorithms for dataset I the highest result comes from the Naïve Bayes classifier model and Table X shows for dataset II Support Vector Machine has the highest result. But for both dataset the highest accuracy comes from multi-layer dense neural network. That means multi-layer dense neural network gives the best performance for both dataset. In Table IX it is also shown that decision tree classifier gives the worst result for dataset I. On the other hand from Table X, it is shown that k-nearest neighbor classifier gives the worst result. If the confusion matrix of Fig. 11(b) is observed it is seen that the highest false classification is found for k-nearest neighbor. From Tables IX and X of overall performance, it is shown that most of the classifiers have low variance and low bias which indicates the proposed model doesn't have underfitting and overfitting.

TABLE IX. OVERALL PERFORMANCE OF DIFFERENT CLASSIFIER MODEL ON DATASET I

Classifiers		**Accuracy**	**Precision**	**Recall**	**F1-score**
Type	**Name**				
Machine Learning Algorithm	Naïve Bayes	**91.23%**	**91.37%**	**91.28%**	**91.32%**
	K-Nearest Neighbor	84.81%	85.13%	85.06%	85.09%
	Support Vector Machine	89.12%	89.41%	89.30%	89.35%
	Random Forest	87.01%	87.20%	87.49%	87.34%
	Decision Tree	62.45%	64.01%	62.29%	63.44%
	Logistic Regression	90.52%	90.72%	90.71%	90.72%
	SGD Classifier	88.77%	88.96%	89.19%	89.08%
Neural Network	Multi-layer Dense Neural Network	**92.63%**	**93%**	**92.8%**	**92.8%**

TABLE X. OVERALL PERFORMANCE OF DIFFERENT CLASSIFIER MODEL ON DATASET II

Classifiers		**Accuracy**	**Precision**	**Recall**	**F1-score**
Type	**Name**				
Machine Learning Algorithm	Naïve Bayes	92.76%	91.84%	90.24%	91.03%
	K-Nearest Neighbor	70.4%	85.05%	64.6%	73.42%
	Support Vector Machine	**94.99%**	**93.84%**	**93.78%**	**93.81%**
	Random Forest	91.4%	91.07%	88.36%	89.7%
	Decision Tree	79.87%	77.03%	76.81%	76.92%
	Logistic Regression	94.6%	93.54%	93.14%	93.34%
	SGD Classifier	93.78%	92.68%	91.92%	92.30%
Neural Network	Multi-layer Dense Neural Network	**95.50%**	**94.6%**	**94.2%**	**94.4%**

V. Conclusion and Future Work

The main focus of this research is to build an automatic classification system for Bangla News documents. This system provides users an efficient and reliable access to classified news from different sources. Different as well as most widely used machine learning classifiers and multi-layer dense neural network are used for categorization and a comparison has been conducted between them. Among the classifier algorithms, Support Vector machine Classifier provides the best result. In the model, TF-IDF technique is used for vectorization to fit data to the classifier.

In future, word2vec model will be used for better result and for preventing the limitation of TF-IDF model. In TF-IDF model, more importance is put on the uncommon words. But, semantic information of the words is not stored in TF-IDF model.

In this research, multi-layer dense neural network and some built in classifier like Naïve Bayes classifier, k-nearest neighbor classifier, random forest classifier, support vector machine classifier and decision tree classifier were used. In future, CNN, RNN and other neural network model will be examined to build the model for better performance.

References

[1] Tenenboim, L., Shapira, B. and Shoval, P. 2008. “Ontology-based classification of news in an electronic newspaper”.

[2] Pendharkar, B., Ambekar, P., Godbole, P., Joshi, S. and Abhyankar, S. 2007. “Topic categorization of rss news feeds, Group”.

[3] Carreira, R., Crato, J. M., Gonçalves, D. and Jorge, J. A. 2004. “Evaluating adaptive user profiles for news classification”. 9th international conference on Intelligent user interfaces, pp. 206–212.

[4] Fauzi, M. A., Arifin, A. Z., Gosaria, S. C., & Prabowo, I. S. (2016). “Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model”. Indonesian Journal of Electrical Engineering and Computer Science, 2(3), 401-408.

[5] Dutta, K., Kaushik, S., & Prakash, N. (2011). “Machine learning approach for the classification of demonstrative pronouns for Indirect Anaphora in Hindi News Items”. The Prague Bulletin of Mathematical Linguistics, 95(1), 33-50.

[6] El-Barbary, O. G. (2016). “Arabic news classification using field association words”. Advances in research, 1-9.

[7] Beresi, U. C., Adeva, J. G., Calvo, R. A., & Ceccatto, A. H. (2004, August). “Automatic classification of news articles in Spanish”. In Actas del Congreso Argentino de Ciencias de Computación (CACIC) (pp. 1588-1600).

[8] Kabir, F., Siddique, S., Kotwal, M. R. A., & Huda, M. N. (2015, March). “Bangla text document categorization using stochastic gradient descent (sgd) classifier”. In 2015 International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 14). IEEE.

[9] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’99. New York, NY, USA: ACM, 1999, pp. 42–49. [Online]. Available: http://doi.acm.org/10.1145/312624.312647.

[10] Chy, Abu Nowshed & Seddiqui, Hanif & Das, Sowmitra. (2014). “Bangla news classification using naive Bayes classifier”. 16th Int'l Conf. Computer and Information Technology, ICCIT 2013. 10.1109/ICCITechn.2014.6997369.

[11] FouziHarrag,Farhat ABBAS University,Eyas EI Qawasmah,JUST University,“Neural Network for Arabic Text Classification”, in 2009 Second International Conference on the Applications of Digital Information and Web Technologies, doi:10.1109/ICADIWT.2009.5273841.

[12] Selamat, A., & Omatu, S. (2003, July). Neural networks for web page classification based on augmented PCA. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 3, pp. 1792-1797). IEEE.

[13] Hossain, M. R., Sarkar, S., & Rahman, M. “Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization”. International Journal of Computer Applications, 975, 8887.

[14] Manisha Chakraborty, and Mohammad Nurul Huda, “Bangla Document Categorization using Multilayer Dense Neural Network with TF-IDF”, International Conference on Advances in Science, Engineering Robotics Technology (ICASERT 2019), May 3-5, 2019, Dhaka, Bangladesh, pp. 1-4.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[15] Mandal, A. K., & Sen, R. (2014). “Supervised learning methods for bangla web document categorization”. arXiv preprint arXiv:1410.2045.

[16] M. M. H. Shahin, T. Ahmmed, S. H. Piyal and M. Shopon, "Classification of Bangla News Articles Using Bidirectional Long Short Term Memory," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1547-1551, doi: 10.1109/TENSYMP50017.2020.9230737.

[17] twintyOne, “Prothom alo [2013 - 2019]”, Aug 2019. [Online]. Available: https://www.kaggle.com/twintyone/prothomal .

[18] stopwords-iso stopwords-bn. [Online]. Available: https://github.com/ stopwords-iso/stopwords-bn.

[19] Bengali Stopwords, [Online]. Available: https://www.ranks.nl/ stopwords/bengali?fbclid=IwAR3gFJxN8Yo3BT6S9bMGY87NQzudq Mf7z9kxX4veH0aYMLZBBrDaCrZH1jo.

[20] sharminyeasmin198, “Bengali News Dataset (Prothom alo)”, March 2021. [Online]. Available: https://www.kaggle.com/sharminyeasmin198 /bengali-news-dataset-prothom-alo.

[21] sharminyeasmin198, “Bengali Stopwords 2021”, March 2021. [Online]. Available: https://www.kaggle.com/sharminyeasmin198/bengali-stopwords-2021.

The Effect of using Light Stemming for Arabic Text Classification

Jaffar Atwan[1]
Department of Computer Information Systems
Al-Balqa Applied University
Al-Salt, Jordan

Mohammad Wedyan[2]
Faculty of Artificial Intelligence
Al-Balqa Applied University
Al-Salt, Jordan

Qusay Bsoul[3]
Faculty of Science and Technology
Universiti Sains Islam Malaysia
Bandar Baru Nilai, Malaysia

Ahmad Hamadeen[4]
Department of Computer Science
Al-Balqa Applied University
Al-Salt, Jordan

Ryan Alturki[5]
Department of Information Science
College of Computer and Information Systems
Umm Al-Qura University, Makkah, Saudi Arabia

Mohammed Ikram[6]
Computer Science Department
University College in Al-Jamoum
Umm Al-Qura University

***Abstract*—Arabic is one of the Semitic languages in antiquity and one of the six official languages of the UN. Also, Arabic classification plays a significant and essential role in modern applications. There is a big difference between handling English text and Arabic text classification; preprocessing is also challenging for Arabic text. This paper presents the implementation of a Naïve Bayes classifier for Arabic text with and without stemmer. A set of four categories and 800 documents were used from the Text Retrieval Conference (TREC) 2001 dataset. The results showed that Naïve Bayes with light stemmer achieves better results than Naïve Bayes without stemmer. The findings of the classifier accuracy by employing stemmer and without stemmer are as preprocessing. It reveals that the accuracy resulted from the light stemmer was better than the classifier without stemmer detection, which Naïve Bayes Classification with light stemmer got 35.0745 higher than the Naïve Bayes Classification 33.831% without stemmer. After contrasting them, the stemmer got better accuracy than the classifier.**

Keywords—Arabic language; light stemming; information retrieval; Naïve Bayes classification

I. Introduction

Machine-readable information is available in large and increasing quantities that complicate comprehension and use. Machine learning (ML) provides tools that help organize vast numbers of texts and detect them automatically [1]. Feature selection is the most significant application in ML. Feature selection, by identifying the most salient features for learning, the interest of the learning algorithm sheds the lights on the most valuable data for analysis and future prediction [1-4]. The ideas were adopted from test theory to develop a technique for correlation-based feature selection and estimate a group of ML algorithms that taught a different group of natural and artificial problems. The feature selection is easy and fast to implement; it removes irrelevant and repetitive data and its ability to enhance the learning algorithms' performance in many cases. The results of this technology can be contrasted with a modern feature selector elicited from the literature but require much less computation. In the current research, the domain of Arabic documents will be used to decrease features number and promote the performance detection of the Arabic document.

There is no comprehensive definition of the Arabic language. Some Arabic language documents or emails may have some unwanted words or samples; this email detects spam due to unwanted email messages; not all Arabic emails are spam. And the commercial email that is spam [5], spam as the name suggests, is unwanted emails, but there is a general question, what are junk emails? Although Arabic is known to many users, it is difficult to see the definition of both the Arabic language and unwanted messages.

This paper is organized as the following. Section 2 provides a review of previous related studies. The background of the study on Arabic classifiers is tackled by Section 3. The steps followed during the experiment present in Section 4. Section 5 shows the experimental setup and results. Finally, Section 6 discusses the conclusion and future work.

II. Literature Review

Text Classification (TC) is considered as a form of supervised learning task which involves assigning documents with predefined category labels depending on the suggested likelihood by a set of labeled training documents. Before the ML approach emerged in TC, knowledge engineering was the most common way in which the expertise of those working in the field was utilized. A set of rules was manually created into document classification within predefined categories [6] explained that ML recruitment in TC presented several advantages as it lower cost and time in terms of the expert workforce only without any effect on precision. TC problems often present as a group of D documents and a group of S predefined categories with the significant purpose of assigning each (d_i, c_j) pair with a Boolean value (d_i = ∈ D and c_j = ∈ S). An indeed given (d_i, c_j) value represents the decision of allocating document d_i to class c_j; however, a falsely assigned (d_i, c_j) value represents otherwise. Formally speaking, the primary task is the approximation of the obscure objective function f: D × S → {True, False} [4], which clarifies an actual method of classifying the documents through the classification function f': D × S → {True, False} in a manner there will be a minimal number of decisions off and f' that do not correspond.

Several learning algorithms have been developed and applied to TC. Among them includes k-Nearest Neighbor (KNN), Decision Tree (DT), Neural Networks, and Naive Bayesian (NB) [7-10].

The author in [11] compared these learning techniques and concluded that they all perform equally when there are more than 300 documents in each category. However, DT, KNN, and LLSF performed better than neural networks and Naïve Bayes Classification (NB) when there are fewer than ten positive training documents in each category.

Although several supervised ML approaches have been applied to TC, the task still demands the extra effort to predefine the categories and assign category labels to the training set documents. In large and dynamic text databases, this can be a complicated task. According to [6], inter-indexer inconsistency is another phenomenon that makes TC complicated. Accordingly, two experts may disagree on the category to designate a document. For example, a story about Bill Clinton and Monika Lewinsky could be classified under politics, gossip, both types, or neither category based on the subjective decision of human indexer [6]. The first synthesis that motivates us to discuss this is as follows:

- A large number of classification text.
- A large feature space.

TC's applications include document organization, hierarchical web page categorization, and text filtering. Document organization indicates the task of structuring documents into folders (maybe flat or hierarchical). For example, the incoming adverts to an editorial office may be grouped into categorized like Cars, Real Estate, Computers, etc., before publication. Conversely, text filtering indicates to the classification of a dynamic group of documents to relevant and irrelevant groups. This is illustrated in a news system in which articles in a newspaper are filtered from a news agency [6]. For instance, the delivery of news unrelated to sports in a sports newspaper will be blocked. Likewise, incoming messages may be classified as Arabic or not Arabic by a document filter in a bid to block Arabic message delivery [12]. The hierarchical classification process is one of the most flexible processes in the web browsing process due to the ease of navigating the hierarchical form of the categories and directing the search to a specific type instead of putting a general query on a search engine for general purposes. The hierarchical classification of documents generally requires the subdivision of the classification problem into smaller classification tasks. Some of the previous studies that addressed hierarchical document classification are [13-16].

TC using ML techniques entails a preprocessing of the texts, which require the transformation of the document into suitable forms for applying the learning algorithms. [17] introduced the commonly used vector space model called document representation. This model represents each document as a vector with each separate dimension corresponding to the word distinct occurring for all the words in the document. Text classifiers are applied for crime detection and weather prediction. In addition, they are also used to detect and track Arabic over documents. Table I showed the Arabic classification process is both detailed and benchmarked with previous works.

TABLE I. THE PROCESS OF ARABIC CLASSIFICATION

Training Phase	**Testing Phase**
Language preprocessing	***Language preprocessing***
Stemming and stop word removal, Tokenization, Normalization.	Stemming and stop word removal, Tokenization, Normalization.
Removal	***Removal***
Collection of Word	Collection of Word
Representation	***Representation***
TFIDF	TFIDF
Classifier	Classifier
Naïve Bayesian	Naïve Bayesian
Evaluation Accuracy	Evaluation Accuracy

III. BACKGROUND OF STUDY

A. Document Preprocessing

As indicated earlier, the scheme of document grouping in which intra-group similarities are high and low [18]. The essential process is preprocessing due to its significant role in improving and developing these schemes for any classifiers. The text analyst should take into account the relationship between preprocessing and similar measures on Document Classification. Performing preprocessing of documents is an essential process for classifying documents that are implemented by applying ML techniques.

Reference [18] pointed out that the significance of such stage in Document Classification is attributed to the presence of large numbers of unnecessary words present in the documents, many of which negatively influence classification rather than help in that. Using documents as a whole with unnecessary words is complicated. There are many of these words. The researchers themselves presented some non-essential utterances that may be found in the documents. The terms are classified into conjunctions, other grammatically based categories, particles, and that are typically employed without providing any assistance to the researcher in the classification of the document. In addition to some of the words presented by researchers like these words in the English language "ate, eaten, eat" and "أكلت , يؤكل , الاكل "in Arabic language, it is possible to decrease the number of the unique document words. Hence, we conclude that documents free of unnecessary words are applied to them with appropriate prior treatment, and this would improve and increase the performance of the Document Classification approach.

After applying this step, the documents must be converted into a form appropriate for the representation process. Thus, the learning algorithms application is conducted. Then work will be done to remove unnecessary words like special markers and punctuation marks. To perform this process, many commonly used tasks, namely normalization, tokenization, stop-word removal, and mainly stemming, needed to be done. These tasks will be illustrated below, according to the review of previous studies.

B. *Classification Algorithm*

In the supervised algorithms, it is assumed that the categorical structure of a given database is already known. The supervised algorithms require a set of labeled documents to map documents into the predefined labels. As mentioned previously, it is challenging to determine the category and correct label of the training sets, particularly in large databases. Hence, this section will focus on the commonest supervised algorithm, NB.

NB is one of the common ML techniques. It depends on the Bayes' theorem, which claims to have strong (naive) and cumulative independence assumptions. A thorough description of the fundamental probability model theorem serves as the independent feature for the NB. A proper NB classifier could simply presume that there is no relationship between the existence or absence of a given class feature with that of any other feature. Also, and considers a simple probabilistic-based classifier. This assumption is expressed as follows:

$$\mathrm{P}(C_i|d) = \frac{P(C_i)P(d|C_i)}{P(d)} \qquad (1)$$

Where $P(\mathrm{C_i}\,|d)$ it indicates a previous possibility of category C_i in the presence of another instanced, $P(\mathrm{C_i})$ represents the possibility of category C_i which might be calculated through:

$$P(C_i) = \frac{N_i}{N} \qquad (2)$$

where N_i = the number of documents belonging to the category C_i, and N represents categories number, $P(d|\mathrm{C_i})$ represent the possibility of d document belongs to the given categoryC_i, as well as P(d) is the possibility of instance d. clarifies the complete NB pseudocode.

- BEGIN: about all the available values.
- Follow the rules for every individual value as:
  - Calculate and count the values of the classes appearing.
  - Obtain the class, which is frequently occurring.
  - Make the rule, which connects this particular class with instance values.
  - Find out the rate at which the error occurred for the rule.
  - Choose the rules with the minor error rate END.

IV. Experiment Process

A. *DataSet*

The dataset includes a set specifically designed to assess the extraction of the Arabic text for Arabic classifiers created as part of TREC 2001. The group has 383,872 Arabic documents, mostly newswire dispatches issued by Agence France Press (AFP) between 1994 and 2000 [19]. Ground truth and standard TREC queries have been created for such collection: 25 queries were considered part of TREC 2001 (Technology, 2001). The collection of queries has matching relevance judgments produced utilizing the pooling technique. Based on that, part of TREC 2001 is defined for classifiers as in Table II, which include four classes along with a group of documents (a gross of 800 documents).

B. *Normalization*

The steps of normalization ensure a specific characters' order that allows multiple variants. The reason behind the importance of data normalization for Arabic experiments is attributed to the fact that different encoding guidelines might either be used or not used at all by newspaper article and sometimes occurs in the same language. The following steps pinpoint the normalization of corpus and quires employed by [20]:

- Punctuation Removal.
- The removal of diacritics. Some entries consisted of weak vowels. Such elimination enabled text to become compatible.
- The removal special characters and numbers.
- Substituting آ, أ , and إ with ا.
- Substituting the end of ى with ي.
- Substituting the end of ة with ه.

C. *The Removal of Stop-Words*

Terms that usually frequently appeared in every document are so-called stop-words. These terms give no hint of their core document contents. Stop-words are determined so that a stop-words list could be established [21] . For that, an important thing in the Arabic system to omit stopping words from documents during preprocessing. Consequently, it is omitted from the group of indexed terms. Since there is no unified stop-words list of Arabic systems that can be used in classification systems. In this study, Khoja stop-words list tested with light10 stemmer [22].

TABLE II. Summary Description of Arabic Data Set

Categories	# of document
1	200
2	200
3	200
4	200

D. Tokenization

Converting a word into a distinctive word in text processing is a highly significant step. The tokenization process bears the responsibility of splitting the text into tokens, defining boundaries, words, numbers, and abbreviations. Arabic text tokenization is an important initial step as part of pre-processing phase. To define the complete word, in such a paper, the word was considered bound by white space marks to tokenize the Arabic text. Most importantly, the stemming process is regarded as the follow-up step after tokenization and the removal of stop-words.

E. Stemming

The term stemming indicates an approach of conflation that seeks to locate a common stem for a group of words in a text [23, 24]. For the conflation process, we used light10 stemmer by following the same process used by Larkey's [25]:

- They are deleting "و" ("and") if the rest of the word has greater than or equal to three letters. Regardless of the importance of deleting"و ," it considers also problematic due to the fact that many popular Arabic words begin with this letter. Therefore, the length standard is more stringent here than the specific definite articles.
- Omitting every definite article, as this omission leaves word length greater than or equal to two letters.
- Dwelling into the list of suffixes once in the order (right to left) as shown in Table III, omitting every one of them that were at the end of the word if these omitting leaves word length greater than or equal to two letters.
- Table III shows the eliminated strings. Both conjunctions and definite articles are considered 'prefixes'. Light10 stemmer does not omit any character that can be considered an Arabic prefix.

TABLE III. PREFIXES AND SUFFIXES THAT ARE ELIMINATED THROUGHOUT LIGHT 10

Prefixes	Suffixes
ال، وال، بال، كال، فال، لل، و	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي

F. Estimation

To complete estimation, Term Frequency (TF) x Inverse Document Frequency (IDF)) (TF x IDF) weighting was used for the weighting calculated as.

$$w_i = tf_i . log\left(\frac{N}{df}\right) \quad (3)$$

In classification problems, the estimation measures are generally determined from a matrix by employing a group of incorrectly and correctly categorized for each class (called the confusion matrix). Table IV revealed the confusion matrix for a binary categorization problem with classes that are merely positive and negative.

Here, FP, FN, TP, and TN are described thus:

- False Positives (FP): The negative examples are wrongly projected as positive.
- False Negatives (FN): positive instances which are wrongly predicted as negative.
- True Positives (TP): The positive examples that are correctly projected as positive.
- True Negatives (TN): The negative instances which are correctly predicted as negative.

The accuracy rate (ACC) is the commonly used estimation measure on the ground that estimates the classifier efficiency depends on its proportion of correct projections. The ACC of a classifier is calculated as follows:

$$ACC = ((TP + TN)/(TP + TN + FP + FN)) * 100 \quad (4)$$

TABLE IV. CONFUSION MATRIX

Projected Class		
True Class	***Positive***	***Negative***
Positive	TP	FN
Negative	FP	TN

V. EXPERIMENTAL SETUP AND RESULTS

As soon as the documents of the text are processed, they go through the classification tasks that occur by both stemming and converting them to an appropriate format. The collection of the document is classified twice for evaluation objectives. The first categorization is without stemmer, in which document collection occurs before stemming and applies to documents, whereas the second categorization is referred to as light stemmer that is used for the document collection. In respect of the available classifier approach, the overall dataset is classified for cross-validation. In conclusion, the cross-validations are classified as well as those in [26] for estimation. The remaining folds are employed for testing objectives, whereas K-fold cross-validation, K - 5 folds are used for validation and training.

Table V reveals the findings of the classifier accuracy by employing stemmer and without stemmer as preprocessing. It reveals that the accuracy resulted from the light stemmer was better than the classifier without stemmer detection, which the NB with light stemmer got 35.0745 higher than the NB 33.831% without stemmer. After contrasting them, the stemmer got better accuracy than the classifier. The last evaluation by employing the number of features revealed that without stemmer worse in decreasing the number of features, whereas stemmer was better, as shown in Table V.

TABLE V. EFFECT OF THE LIGHT STEMMER AND WITHOUT STEMMER USING NB AS A CLASSIFIER

Stemmer	Without Stemmer	Light Stemmer
Accuracy	33.830846	35.074627
#features	91756	46167

VI. Conclusion and Future Work

The experiments revealed that applied stemmer as preprocessing on our data set considers significant. Therefore, it has a substantial impact on the NB classifier on side accuracy and features number. Without stemmer, it does not significantly impact our dataset, as illustrated in Fig. 1 and 2. Major weakness concerning the without stemmer in preprocessing pertains to the dimensionality of terms requiring the second contribution to fill the huge term called feature selection. The critical concepts to be taken from the web to improve Arabic preprocessing are indicated. Therefore, the real datasets are employed in the experiments of this study. The primary finding regarding stemmer is essential for categorization. As stated earlier, it is better than with stemmer. In general, the finding is inconclusive due to the effect of classification or the chosen features revealed.

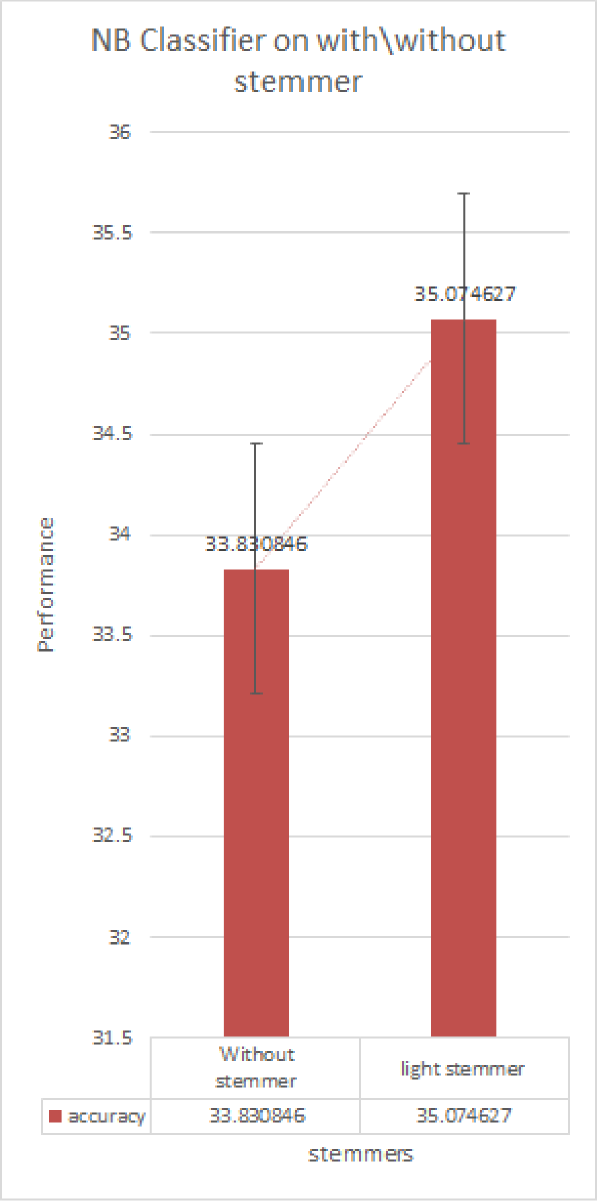


Fig. 1. Performance of Arabic Classifier using with/without Stemmer.

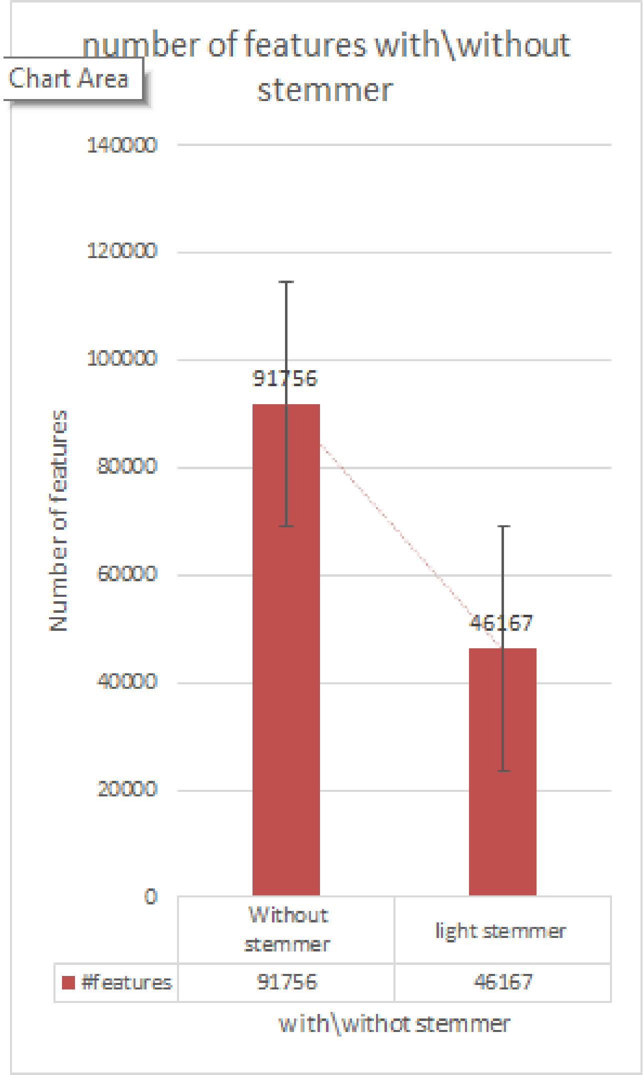


Fig. 2. Effect of Number of Features using with/without Stemmer.

Nevertheless, the central related weakness without stemmer in preprocessing is the dimensionality of terms revealed, which requires future work to bridge the gap concerning huge terms number called feature selection. This study answers the question regarding the effects of classifiers algorithms on Arabic classifiers with/without employing stemming of words and the purpose of investigating the performance of the used NB as Arabic classifiers on the classification performance with/without the employment stemming of terms.

A lot of open questions of the study are unanswered. Content classification considers a significant research area that provides various directions for additional studies. As a result, this section presents some suggestions for future research. Future studies conducted in this field are recommended to read this paper to try different classifiers and contrast their performance by employing various stemmer's algorithms. Also, using feature selection strategies for reducing the dimensionality of terms in the datasets as well as choosing the best features depending on their relevancy and significance to the subject class.

ACKNOWLEDGMENT

The researchers wish to thank (LDC) the Linguistic Data Consortium for their supplement of LDC2001T55 Arabic Newswire Part 1 at no cost, as one of the students of the Fall 2012 LDC Data Scholarship program.

REFERENCES

[1] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, p. e01802, 2019.

[2] G. Jain, M. Sharma, and B. Agarwal, "Spam detection on social media using semantic convolutional neural network," International Journal of Knowledge Discovery in Bioinformatics (IJKDB), vol. 8, no. 1, pp. 12-26, 2018.

[3] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," in 2018 10th International Conference on Communication Systems & Networks (COMSNETS), 2018: IEEE, pp. 380-383.

[4] M. Revanasiddappa and B. Harish, "A new feature selection method based on intuitionistic fuzzy entropy to categorize text documents," IJIMAI, vol. 5, no. 3, pp. 106-117, 2018.

[5] J. A. Zdziarski, Ending spam: Bayesian content filtering and the art of statistical language classification. No starch press, 2005.

[6] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1-47, 2002.

[7] T. Joachims, "Making large-scale SVM learning practical," Technical Report, 1998.

[8] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," Stanford InfoLab, 1997.

[9] B. Masand, G. Linoff, and D. Waltz, "Classifying news stories using memory based reasoning," in Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992, pp. 59-65.

[10] Y. Yang, "An evaluation of statistical approaches to text categorization," Information retrieval, vol. 1, no. 1-2, pp. 69-90, 1999.

[11] Y. Yang and X. Liu, "A re-examination of text categorization methods," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 42-49.

[12] L. Özgür, "Adaptive Anti-Spam Filtering Based on Turkish Morphological Analysis, Artificial Neural Networks and Bayes Filtering," Bogazici University. Institute for Graduate Studies in Science and Engineering, 2003.

[13] M. Sahami, "Using machine learning to improve information access," Stanford University, Department of Computer Science, 1998.

[14] S. Dumais and H. Chen, "Hierarchical classification of web content," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 256-263.

[15] M. Wedyan, B. Alhadidi, and A. Alrabea, "The effect of using a thesaurus in Arabic information retrieval system," Int. J. Comput. Sci, vol. 9, pp. 431-435, 2012.

[16] G. Kanaan and M. Wedyan, "Constructing an automatic thesaurus to enhance Arabic information retrieval system," in The 2nd Jordanian International Conference on Computer Science and Engineering, JICCSE, 2006, pp. 89-97.

[17] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.

[18] W. B. Croft, D. Metzler, and T. Strohman, Search engines: Information retrieval in practice. Addison-Wesley Reading, 2010.

[19] A. Cole, D. Graff, and K. Walker, "Arabic Newswire Part 1 Corpus (1-58563-190-6)," Linguistic Data Consortium (LDC), 2001.

[20] J. Atwan, M. Mohd, H. Rashaideh, and G. Kanaan, "Semantically enhanced pseudo relevance feedback for arabic information retrieval," Journal of Information Science, vol. 42, no. 2, pp. 246-260, 2016.

[21] B. Alhadidi and M. Alwedyan, "Hybrid Stop-Word Removal Technique for Arabic Language," Egyptian Computer Science Journal, vol. 30, no. 1, pp. 35-38, 2008.

[22] J. Atwan, M. Mohd, and G. Kanaan, "Enhanced arabic information retrieval: Light stemming and stop words," in International Multi-Conference on Artificial Intelligence Technology, 2013: Springer, pp. 219-228.

[23] J. Atwan and M. Mohd, "Arabic Query Expansion: A Review," Asian Journal of Information Technology, vol. 16, no. 10, pp. 754-770, 2017.

[24] J. Atwan, M. Wedyan, and H. Al-Zoubi, "Arabic text light stemmer," Int. J. Comput. Acad. Res, vol. 8, no. 2, pp. 17-23, 2019.

[25] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light stemming for Arabic information retrieval," in Arabic computational morphology: Springer, 2007, pp. 221-243.

[26] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning (no. 10). Springer series in statistics New York, 2001.

Travel Behavior Modeling: Taxonomy, Challenges, and Opportunities

Aman Sharma[1], Abdullah Gani[2], David Asirvatham[3], Riyath Ahmed[4],
Muzaffar Hamzah[5], Mohammad Fadhli Asli[6]
Faculty of Innovation and Technology, Taylor's University Malaysia[1,3,4]
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia[2,5,6]

***Abstract*—Personal daily movement patterns have a longitudinal impact on the individual's decision-making in traveling. Recent observation on human travel raises concerns on the impact of travel behavior changes on many aspects. Many travel-related aspects like traffic congestion management and effective land-use were significantly affected by travel behavior changes. Existing travel behavior modeling (TBM) were focusing on assessing traffic trends and generate improvement insights for urban planning, infrastructure investment, and policymaking. However, literature indicates limited discussions on recent TBM adaptation towards future technological advances like the integration of autonomous vehicles and intelligent traveling. This survey paper aims to provide overview insights on recent advances of TBM including notable classifications, emerging challenges, and rising opportunities. In this survey, we reviewed and analyzed recently published works on TBM from high-quality publication sources. A taxonomy was devised based on notable characteristics of TBM to guide the classification and analysis of these works. The taxonomy classifies recent advances in TBM based on type of algorithms, applications, data sources, technologies, behavior analysis, and datasets. Furthermore, emerging research challenges and limitations encountered by recent TBM studies were characterized and discussed. Subsequently, this survey identified and highlights open issues and research opportunities arise from recent TBM advances for the future undertaking.**

***Keywords*—Travel behavior; travel behavior modeling; prediction modeling; intelligent traveling**

I. Introduction

Personal daily movement habit cumulatively influences the decision-making approach of the individual when traveling. Recent observation on human travels raises concerns on the impact of recent travel behavior changes towards many aspects including traffic management, transport planning, and decision-making. Researchers carried out travel behavior studies with the goal of understanding human travel behavior via analytical approach. The main goal of travel behavior studies is to generate translational travel behavior model for facilitating prediction analysis for travel planning insights. Therefore, Travel Behavior Modeling (TBM) is introduced as cost-efficient and timesaving approach for supporting that goal. Traffic management are severely affected by increasing traffic congestion occurrence due to human travel behavior. Recent observation pointed out the increases of inferior travel habit like slow driving style that further amplified by faulty road condition and crowded area. These behaviors and external factors eventually cause indirect losses on many aspects. Massive transport congestion in the US costed the economy with average wasted 5.5B work hours, more than 2.9B gallons of fuel, and more $100 billion per year. China also faced similar problem with consequences of increasing annual travel cost that currently priced at $1,126 per person [1]. Victoria Transport Policy Institute highlighted that traffic congestion is the leading cause of critical delay, high fuel consumption and economic wastage in the US. This issue has considerably contributed towards decline in trade and growth. This expense is expected to rise by 2020 to the extent of $121B [2].

Future travel behavior is expected to change drastically soon due to mainstream popularity of autonomous vehicles and smart travel technologies. Increasing adaptation of autonomous vehicles eventually change how human travel, consequently impacting travel behavior norms and environment [3]. The use of autonomous vehicles increases travel mileage with reduced share of slowing traffics with other slow-mode public transports. These changes are already reflected in growth pattern in many dispersed urban cities. Eventually, autonomous vehicles reduce the need for private vehicles and provide extraneous conveniences for user despite its minor contribution towards parking congestion. A recent study reported the implications of substituting all private motor vehicles with shared autonomous vehicles (SAV) in Berlin, Germany [4]. Their findings indicates that the substitution is likely to decrease the number of cars that accommodates private travel demand by 91%. In addition, the number of vehicles that accommodates travel demand in Lisbon are expected to decrease by 95% with the introduction of SAV [5]. Moreover, travel demand involving family-sharing private vehicles are also expected to significantly decrease due to redundant purpose [6]. However, such substitution in Atlantis, US only observe minor decrease of only 9.5% due to lower priority on accommodating individual travel demand. In addition, the association of travel behavior with socio-economic (income, household composition) indicates land-use patterns that is leveraged to assess the mobility of individual traveler [7]. TBM is expected to play significant roles in realizing these future endeavors due to its direct role towards predicting traveler choices.

Many studies were conducted over the years in developing methods, algorithms, and models for TBM. Such examples of modeling are Short-term Travel Behavior Prediction [8], Intra-Household TBM [9], Sense Place TBM [10], Markov Travel Model [11], and Inferring Structural Models [12]. These modeling were built for facilitating prediction on travel demand and travel decision. Many TBM studies analyzes and diagnoses travel behavior pattern at individual level [13], [14], [15], [16]. Eventually, specific modeling for this scale were developed like Individual TBM [14], Regret-based Choice

Model [13], Rethinking TBM [16] and Heterogeneous TBM [15]. With the upcoming changes from autonomous vehicles adaptation, land-use transposition, we can anticipate that TBM holds many potentials with lasting impact towards improving this aspect.

This survey aims to provide overview insights on recent advances of travel behavior modeling including emerging trends, challenges, and opportunities. Gaining latest insights on travel behavior modeling eventually contribute towards driving forward many sectors like transportation, tourism, and urban development. In this survey, recent published works on TBM from quality journals and established conferences indexed by Web of Science (WoS) were reviewed. A taxonomy was devised based on the survey to classify recent TBM advances based on many notable aspects. Furthermore, the problems and issues addressed by recent work were reviewed to derive the existing research challenges in TBM. Moreover, this survey analyzed the reported limitations and future works to highlight any persistent issues or opportunities in TBM for future studies. The main contributions of this survey paper are:

1) A taxonomy of travel behavior modeling that classifies TBM based on algorithms, applications, data collection techniques, technologies, travel behaviors, models, and datasets.
2) A summarization of emerging trends on research challenges and limitations in implementing TBM in present applications.
3) A summarization of open issues and opportunities from recent advances in TBM for future undertaking.

The content of this paper is organized as follows. Section 2 explains the fundamentals of travel behavior modeling including definitions, process, and terminologies. Section 3 introduces the devised taxonomy of travel behavior modeling based on the conducted survey including characterizations and instances of each classification. Section 4 then presents the summarization of identified research challenges and limitations from existing TBM. Section 5 then highlights the open issues and opportunities arises from recent TBM advances for future studies. Finally, Section 6 presents the conclusions for this paper.

II. Travel Behavior Modeling

This section introduces the fundamentals of travel behavior modeling including the underlying concepts, definitions, and working terminologies. The presented fundamentals offer general information on travel behavior modeling for understanding the remaining contents in this paper.

The origin of travel behavior modeling started by the introduction of behavioral modeling to systems theory and control theory in 1970s [17]. Behavior modeling addresses the conventional approaches' inconsistencies on state-space, transfers, and convolution representations. Eventually, behavior modeling is adapted in many event or activity analysis over the years including travel behavior analysis. Travel behavior analysis is important in supporting sustainable development in parallel with society growth and technological advances. Travel behavior studies are carried out to explore human travel behavior's impacts towards fuel consumption, environment, urban development, and citizen upbringing.

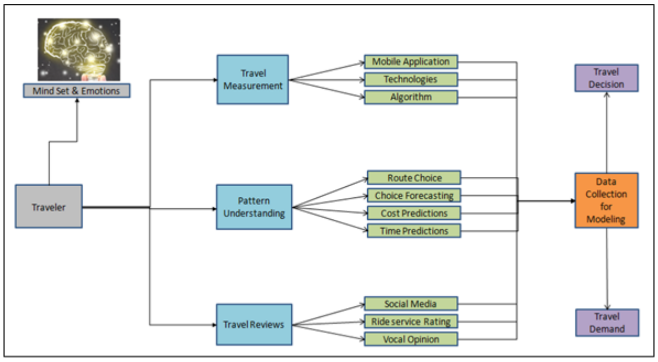


Fig. 1. Functional Model of Travel Behavior Modeling.

Travel behavior analysis starts by anticipating the mindset of traveler associated with travel measurement, pattern understating, and travel reviews. In general, TBM is utilized to process these variables to generate prediction on travel decision and travel demand. TBM learns user activities based on recorded travel experiences information like time, place, and mode of transportation. TBM then diagnose, understand, and response to the change of activities and subsequently predict future travel decision. Eventually, the generated prediction should accommodate travel habits and transportation in future travel based on the preferences and lifestyle of the user. Fig. 1 summarizes the framework of overall processes and related attributes in travel behavior modeling.

To help readers in understanding the remaining contents in this paper, we specify the working terminologies that are commonly used as follows:

- Algorithms: Processing operations or rule sets of the analysis models created by travel behavior modeling.
- Applications: Computer programs or software that implemented travel behavior modeling as core function with the purpose of facilitating travel behavior analysis.
- Models: Machine learning models created based on statistical or mathematical approach to extract and learn behavior patterns from travel datasets.
- Industries: Industrial domains, sectors, or fields of the applied travel behavior modeling.

III. Taxonomy of Travel Behavior Modeling

This section introduces the taxonomy on travel behavior modeling based on the conducted survey as illustrated in Fig. 2. This taxonomy classifies travel behavior modeling based on algorithms, applications, data sources, technologies, analysis, models, and testing datasets. Characterizations of each travel behavior modeling classification are presented as follows.

A. Travel Behavior Modeling Algorithms

Our taxonomy classifies the algorithms of recent TBM advances based on the algorithm objectives and approaches. The algorithms of recent TBM advances are classified and characterized as follows.

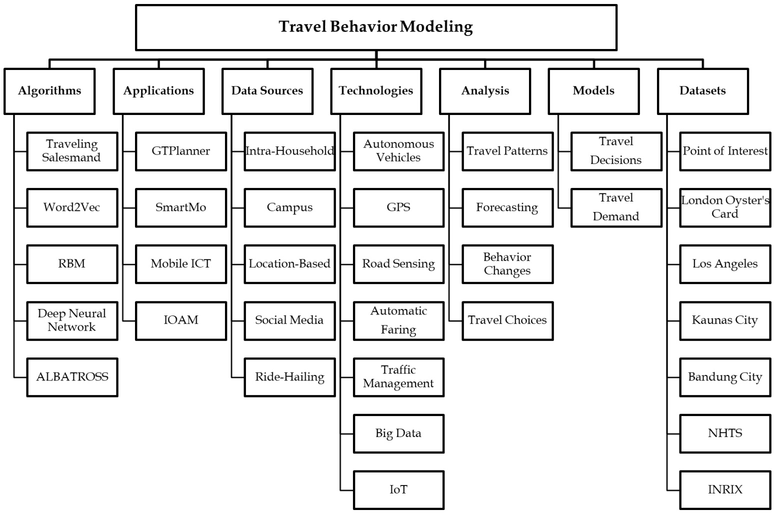


Fig. 2. Taxonomy of Travel Behavior Modeling.

1) Traveling Salesman: The crucial aspect in travel behavior modeling is the efficiency of behavior modeling optimization. Optimizing algorithm like Traveling Salesman are used in several instances such as benchmarking, performance comparison, algorithm selection, algorithm analysis, and algorithm comparison. Optimizing algorithm model is achieved using two ways: (1) fitting nonlinear curves with fixed semantics and (2) training artificial neural networks on benchmark results. A study demonstrated effective travel behavior modeling based on Traveling Salesman algorithm by using model optimization approach [18].

2) Word2Vec: Albeit the various methods available for analyzing travel demand prediction, automatic translation and word prediction are crucial in facilitating the analysis. The Word2vec algorithm allows the recalibration of PyTre packages on supporting automated translation and prediction via natural language processing (NLP). By leveraging natural language processing, performance measurement on testing or training dataset is carried out swiftly. Despite the swift advantages, the algorithm is limited by its stochastic nature that eventually requires implementation via deliberate experimental design. In addition, integrating model like Rethinking Travel Behavior (RTB) is viable in the algorithm development to support performance measurement. RTB model predicts the travel demand based on mode, trip purpose, education level, family type, and occupation. Furthermore, the model utilizes the same datasets throughout processing for consistency [16].

3) Restricted Boltzmann Machine: This algorithm enables consideration of information heterogeneity and variable correlations in processing that commonly occurs on large scale datasets. A recent study investigated on generative machine learning approaches by analyzing multiple discrete-continuous (MDC) travel behavior event from 293,330 travel data [19]. RBM-based algorithm performs significantly better in modeling behavior with larger datasets due to increased modeling accuracy and future forecasting. The study proposed a framework that recommends RGM-based algorithm implementation based on different instances of route choice, dynamic road pricing, and traffic simulation application.

4) Deep Neural Network: Taking advantages of the prowess of deep learning, researchers start to implement deep neural network model to predict travel choices pattern. A recent study proposed a deep neural network structure model that allows many additional structural settings like hidden layer structure, activation functions, number of training epochs, dropout ratio, and learning set ratio [20]. The proposed approach leverages deep learning capabilities to models traveler behavior using deep neural network structure.

5) ALBATROSS: Machine learning identifies that versatile application which is related to the ALBATROSS model using decision trees. Through this model, researchers can focus more on freight transport research, feature engineering, driver behavior, data extraction from digital media and the GPS trajectories that are focusing on transport research [21].

B. Travel Behavior Modeling Applications

In general, travel behavior modeling applications aim to support event pattern analysis on travel demand and planning. However, the applications comprise of different features depending on the travel sectors or the scale of users it covers. Instances of application and its distinct features are presented as follows.

1) GTPlanner: GTPlanner uses TBM for facilitating transportation system planning by generating travel insights based on user preferences and duration [22]. GTPlanner is an exemplary educational travel advisor by promoting the impact of individual traveling modes which consequently fostering quality travel behaviors. GTPlanner also features rudimentary trip planner that allows en route behavior changes tracking and generates latest insight prospects.

2) SmartMo: SmartMo is a mobile-based application for gathering traveler real-time information and generate real-time travel insights. SmartMo gathers and assess event information like pre-trip (mode of transportation and trip purpose), on-trip (start trip, route tracking and end trip), and post-trip(upload). Taking advantages of traveler event information, SmartMo enables real-time route checking, insights update, data autonomy, data security, and time efficiency for monitoring traveler behavior [23].

3) Mobile ICT: Nowadays, mobile devices like smartphone and tablets plays significant role in facilitating travel activities with travel planning and social networking features. Travel planning is simpler than ever before, when smart devices and applications have become mainstream in human culture. Planning tasks like browsing travel date, mode of travel, destination, and coordinating trip decisions are easily carried out via mobile devices. With mobile devices, travelers can fully utilize any online advantages to maximize the travel outcomes throughout the trip duration. Such instances of outcomes including increased count of places visited, social gatherings attended, and trips planned in groups [24]. Empowered by mobile computing, travel behavior and tourist business also change drastically over the years [25]. Travel behavior comprises 5 core phases: dreaming (reviews, blogs, travelers' experiences), planning (information and travel guidance), booking (flight, accommodation, bus, car, and tours), experiencing (on demand booking and maps) and sharing (friends, relatives, and social network) [26].

4) Interactive Online Accessbility Mapping (IOAM): Interactive web usability mapping facilitates many traveling purposes like commuting, school, healthcare, lodging, social activities, and entertainment. IOAM features four transportation modes: walking, biking, vehicle transit, and selected destination requirements. A recent study investigated the role of IOAM in facilitating users on understanding travel behavior. The study reported that participants with access to IOAM valued the tool in understanding residential location and travel-related decision-making process [27].

C. Travel Behavior Modeling Data Sources

Working data for TBM are originated from various sources like social media, crowdsourcing, or service providers. Many TBM studies work with data generated using survey-based approach that utilizes user interviews and online questionnaires. Such instances of data sources that were found in the survey are related with intra-household, campus, location-based tool, social media, and ride-hailing services. The characterizations of each related data sources are discussed as follows.

1) Intra-Household: Modeling and forecasting traveler's time preference takes account on several individual aspects into consideration like demographics, situational, psychological, and sociological. For instance, predicting bus travel time preferences largely dependent on parameters of different modes and actors like bus drivers, other bus users, and other metro users [11]. Another instance is the modeling of individual travel behavior based on intra-household interactions via datasets of 67K trips made by 3.5K individuals [9]. In general, suitable data collection techniques for this type of TBM relies heavily on the individual familiar sense of the location. Another instance is the modeling of travel decision among users between two shopping centers in California via 719 survey data [10]. The utilized TBM assess movement patterns between two traveler vehicles based on two factors: energy conservation and environmental protection. Another instance of intra-household consideration in TBM is travel activity assessment for addressing fuel consumption, car uses, and gas emission in South India [28].

2) Campus: The goal of using TBM for modeling individual travel behavior in campus is to generate mapping on daily transport travel demand. For instance, Dalhousie University, Canada carried out a travel survey among its campus residents pertaining travel activities and environmental awareness. The survey gathered information from campus residents like resident's work, entertainment activities, social activities, formal education activities, and shopping activities [29]. In addition, many travel projects especially from European countries often generated their travel behavior datasets using GPS with volunteer participants from urban cities.

3) Location-Based Tool & Services: The rapid advances of location-based services have turned location-based tool into a powerful data gathering and communication methods. For instance, a study observed travel check-in behavior from social media data by analyzing its location-based social network (LBSN) features [30]. Their findings reveal contrasting pattern of check-in behavior between weekday and weekend activities, with emergences of unique patterns exhibited by different gender. Their study demonstrated on how spatial pattern of check-in mapping surface density and smoothness are ascertained using kernel density estimation (KDE) associated ArcGIS.

4) Social Media: Social media has become the mainstream online interaction platform for discussing or sharing opinions and experiences on traveling. Social media usage on traveling topics holds myriads of information like patterns and reviews of past traveling events recorded by many travelers. Social media network like Facebook, Twitter, and Yelp can generate enriched travel movement behavior data via location-based sharing features [31]. In their study, Rashidi et al. reviewed many possible methods for analyzing travel-related social media data effectively [32]. They highlighted that the geo-tagging features are leveraged by city planners and business premises to understand the urban dynamic of their areas. Gaining insights on recent urban dynamic can help these entities to adapt instead of relying on outdated information from old records. Featuring services like check-in logs and hotspot, social media data helps researchers to understand the travel demand of visited locations. Location-based information from these services allows TBM to predict individual travel behavior with better accuracy and efficiency. However, limited access towards those social media data remains to be the main issue for data mining and natural language processing. Restricted data fetching indirectly reduces the quality of insights on travel behavior that are obtained from social media [33], [34]. Another instance is how a study implemented logistic regression model for analyzing Facebook check-in logs to predict future check-in patterns [35]. Their findings indicate how majority of Singaporeans adapted and modified their leisure travel behavior via social media influences. Another study highlighted the significant impacts of social media in travel itinerary planning and social sharing norm [36].

The advances of social media platform also amplify social media data usage generation, resulting massive collection of user interaction. There are several technical challenges in analyzing social media big data for understanding travel behavior like incompatible text parsing, data format inconsistencies, and hidden sentiments. For instance, Twitter social media data contains check-in logs enabled by geo-tagging features along with text posting information [37]. Travel studies can leverage the geo-tagged information to identify the relationship between conventional model of travel demands [38], [39]. The prospect of traffic incidents identification and verification via Twitter opens the possibilities for efficient and real-time data management. Moreover, Instagram allows location sharing and posting of travel experiences via pictorial, posts, and videos [40]. Other social media platform like Foursquare also records individual travel behavior information based on check-in logs, number of visits, and activity patterns [41]. A movement behavior study on bike-sharing services in Washington used probabilistic models and analysis to understand bike user behaviors [42]. The study explores the relationship between bike-sharing user activities in different event throughout using the services. The employed probabilistic model allows for information fetching on weekly activity patterns from check-in logs via Foursquare.

5) Ride-Hailing Services: Ride-hailing services has turned into mainstream methods for public to access transportation thanks to the advances of mobile computing and smart devices. The services facilitate many advantageous features including quick availability, on-demand, timesaving, flexible payment methods, privacy, and security. Previous travel behavior studies often work with datasets associated with Uber services due to its reputation in ride-hailing services. A transportation analysis study explored the rising popularity of Uber as mainstream method of transportation in the Greater Toronto and Hamilton Areas (GTHA) [43]. The study investigated the behavioral process of combined probabilistic decisions with consideration of process of formation and conditional semi-compensatory decision. However, the developed model does not indicate intricate comparison between Uber and private cars, public transport, and motorless vehicles. They also pointed that most young adults prefer Uber services due to its compatibility and flexibility for accommodating the trip. Nowadays, most ride-hailing services embedded their facilities on smart devices and online platforms, providing users with quick access to transportation. TBM is used for supporting these facilities by facilitating timely demand availability, ride sharing organization, and travel punctuality. The current increasing preferences towards ride-hailing services indicates the increases of demand for automated vehicles in the future. Consequently, public adaptation towards mainstream ride-sourcing norm also triggers changes in transportation aspect like carpooling, biking, ride-hailing (Lyft, Uber), and micro-transit (Bridj, Chariot). Such instance of changes was recorded in a dataset comprising attributes like times, distance, and earning from rides, and travel behavior via social demographic interviews with 311 passengers [44].

D. Travel Behavior Modeling Data Embedded Technologies

Based on the survey, TBM are commonly embedded in many transportation technologies like autonomous vehicles, road sensing, GPS analytics, automatic fare, real-time traffic monitoring, big data, and smart vehicles. The role of TBM on empowering these technologies are discussed as follows.

1) Autonomous Vehicles: The premise of innovating autonomous vehicles (AV) is to allow drivers to perform other en-route tasks or activities apart from piloting the vehicle. Powered by the integration of diffusion model and spatial travel demand model ingenuity, AV poses significant impact on future travel behavior. With mainstream adaptation especially in Germany and USA, AV are gaining attraction due to long-term travel duration and cost efficiency with least mobility impairment. TBM is essential in anticipating the travel behavior changes prompted by this adaptation and predicting its impact on travel demand. Consequently, policy recommendations to support the adaptation of AV and mitigating its negative consequences are crucial [45].

2) Global Positioning System (GPS): The travel pattern generated from GPS logs effectively reflects the daily movement of an individual and their travel behavior. A recent study observed and compared individual movement activities when inside and outside of their home using TBM, specifically C-Means (FCM) clustering algorithm [46]. The use of GPS also contributes towards generating many travel behavior data, that eventually leveraged for strategic urban planning. For instance, Dar es Salaam travel behavior data is generated using citizen GPS to help Tanzanian government in improving public transport policies and infrastructure [47]. TBM help processing these GPS data to generate insights on effective planning for street infrastructures, landmarks, and land use [48].

3) Road Sensing: In their study, Su et al. explores the utilization of road events to enhance the quality of road sensing system [49]. Road events like traffic jams, road accidents, potholes, bumps, and road signs further enrich the compatibility comparison of the suggested route. Subsequently, they proposed a low-cost crowd-sourcing data collection and integration system for detecting en-route risks and generating alternative routes. Safety risks and complications like slippery roads, pavement roughness, and travel regulation are considered based extracted information from crowd-sourced big data.

4) Automated Faring: In his work, Kang implemented scalable inference methods for analyzing the relationship between public transit system route and travel behavior [50]. The study focuses on the association between traveler's origin-destination choices with multi-modal travel environment. Two preference vectors and true OP-pairs were investigated using Automatic Fare Collection (AFC) system-type data (stop-level ODs). The investigation was possible after gathering and analyzing data for route choice with stochastic travel environment specifically for paying travel fare.

5) Real-Time Traffic Management: Such instance of TBM usage in real-time traffic management technologies is the transport model algorithm development by Wen and Chunming. The presented algorithm tracks and predicts real-time traffic dynamics based on agent-based transport model and independent observations of the trajectory of several hundred vehicles [51]. The study involves arrays of TBM technologies including big data, machine learning, and sensor networks for generating efficient travel time and fuel consumption.

6) Big Data Infrastructure: When working with big data, TBM is crucial to facilitate large scale spatial analysis on human mobility via movement monitoring. The analysis itself comprised of processing on many variables including preliminary analysis, path, trip purpose, origin-destination matrices, mode, path choices, and unresolved travel errors. Nevertheless, utilizing TBM with big data requires adaptation depending on the technique and methodologies used to process the data. Future TBM studies should work with larger data sample sizes to explore its scalability on facilitating accurate movement analysis [52].

7) Internet of Things (IoT): With mainstream use of travel route system, travel behavior data is generated in real-time using smart devices and IoT technologies. The data generated via this method often utilized to map travel routes of many travelers based on their point of interest. Using automated event detection, the system gathers and process information like visit duration and activities to generate the next best travel routes. Future TBM studies can develop algorithms that maps frequent tangible travel routes and personalize suggested route ranking.

E. Travel Behavior Modeling Analysis

In general, travel behavior studies examine and analyze human daily travel behavior and its impact to society development. This taxonomy classifies the spectrum of related behavior analysis into 4 classes: travel patterns, forecasting, behavior changes, and travel choices. Summarization of work examples on each type of behavior analysis are presented as follows.

1) Travel Patterns: Behavior analysis on travel patterns focuses on analyzing instances and patterns of travel activities over a period or duration. For example, Krueger et al. proposed an integrated model that facilitates travel pattern assessment with consideration of moral values, norms, modalities, and travel behavior [53]. The model enables rapid identification on any travel pattern modification with concurrent shifts in normative beliefs. However, further refinement of the model is crucial such as including latent parameter constructs like values, aims, and personal identity. In addition, the study plans refining the model to generate travel decisions in a comprehensive representation of social context. Another study highlighted the implications of recent travel behavior of working adults by observing their weekly travel movement. The observation involves working adults without young children in California districts, with the goal of increasing auto use and reducing active transit use. A recent study investigated on travel patterns and its relationship with household compositions in major cities for sustainable urban development [54]. Their findings reveal that insights from urban travel patterns helps city planners and policymakers in addressing societal barriers and sustainable traveling.

2) Travel Forecasting: Travel forecasting predicts predetermined outcomes based on existing event records to helps encounters future travel difficulties. A user can plan their travel accordingly based on the insights generated by travel forecasting system. In this context, the forecast of travel demand via user travel planning helps transportation scheduling and policy making. To generate simulated user travel planning, random forecast is utilized to generate ensembles of random decisions. A robust random forecast can predict and classify user travel preferences to further measure the model capacities and interpretability. Prediction of user travel preferences eventually contributes towards facilitating behavioral interpretation and pattern discovery in complex datasets.

3) Behavior Changes: The changes of dynamic between travel behavior and residential point revolves around three major changing conditions: relocation of travel purpose, residential location choice, and modification of residential environment [55]. Behavior changes in travel activities often characterized by factors or traits like travel satisfaction, travel attitudes, and travel behavior. Behavior changes also influenced by point of travel origin that is commonly home location, reflected towards the entirety of the travel plan. Maneuvering alternatives but unfamiliar routes during travel also triggers behavior changes depending on the travel locations.

4) Travel Choices: Predicting traveler's choice of destination based on emerging trends from previous travel behavior record is challenging. Researchers started to develop specific behavior modeling for assessing destination travel choices to allow predictions on multi-instances of destination choices. Inclusion of real market subdivision with regards to attitudinal parameters is crucial for model plausibility and statistical testing. The destination choice behavior modeling eventually leads towards enabling cross-function classification of other supporting parameters like estimation, goodness, and fitness. The entire process consists of two-stages: starting with common choice of functional destination, then adjustment based on rules allocation [56].

F. Travel Behavior Modeling Models

TBM observes travel demand conditions that are based on a variety of factors including location, place, and time. Travel demand and travel decision are interlinked due to travel decision dependency on variables like cost and transport mode. However, the true common factors between travel demand and travel decisions are typically the location, place, and destination. TBM models enables early prediction on travel demands but heavily reliant on given input like transport mode and precise location or destination. Such instances of TBM models are Individual TBM, Regret-based TBM, and Rethinking TBM, Short-term TBM. Derived from the survey, our taxonomy classifies TBM models into two major categories: facilitating travel decisions and facilitating travel demand.

1) Travel Decision: Travel decision is influenced by traveler's choices with cost, mode of transportation, time, weather, location, and route. The need for understanding traveler's choice bring forth the development of many variation of travel decision-based model. Generating ideal destination choices based on previous travel patterns are challenging especially with consideration of multitude of instances [56].

With increasing traffic congestion occurrence in many metropolitan cities, real-time traffic prediction and routing generation now becomes essential. Using provided information like ongoing routes, traffic density, and incident vicinity, TBM helps by generating optimal routes to destination. Such instance of TBM uses is in GPS routing that monitor trips via rapid route identification using probabilistic modeling. However, TBM implementation in GPS routing can still be improved in terms of accuracy with latest hardware fitting and en-route approach [8].

The goal of intra-household TBM is to facilitate accurate and effective travel decision-making based on daily intra-household interactions. Individual travel decisions often influ-

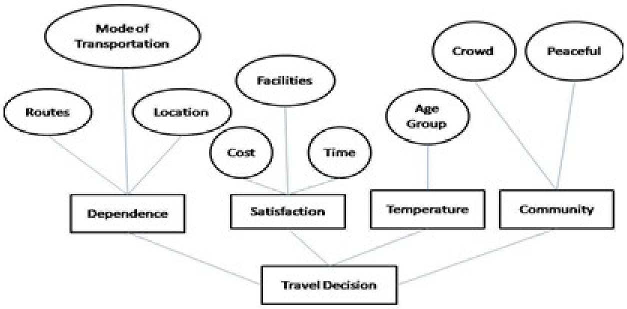


Fig. 3. Relational Variables in Travel Decisions Analysis.

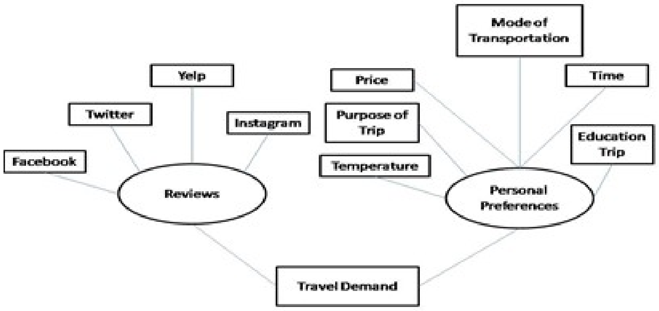


Fig. 4. Functional Model of Travel Behavior Modeling.

enced by opinions, prior experience, and needs of household members or acquaintances. A study developed a method using this model for eliminating decision bias that restrained viable outcomes in accommodating intra-household travel necessity effectively. Bayesian-based multivariate spatial specification is implemented to allow individuals to take household readiness into account during generating travel decision.

Sense Place TBM was built to help identifying relations among the derives of six different variables representing a site. SPTBM is facing many problems with regards to data collection, planning practice, and modeling. A study used SPTBM with the aim to inspect the sense of a location's structure and measurement using survey data. There are three factors in sense of places that are fetched using the three-tiered structural equation model, which parallels estimated sense analysis factors. Sense analysis factors are mostly related to environment [10]. The integrated model of the system involves moral values, norms, modalities, and traveling behavior. This model can identify travel pattern modification during the time of travel in response to concurrent shifts in normative beliefs and behavior. The drawbacks of this model were addressed by bringing improvement in the framework in terms of more latent construct parameters like values, aims, and personal identity. Such approach allows for more comprehensive representation for travel decisions in a social context [53].

Markov Travel Model (MTM) focuses on travel decision concerning travel time, future foresting and individual preferences and behavior. Such TBMs that utilizes MTM concept are Class-Specific Mode Choice Model, Initialization Model, and Transition Model. These models would identify the evolution of preferences or modality styles. Possible enhancements to this study are based on preference dependencies for the individuals based on habits and framework improvements [11].

Another instance of travel decision model is Inferring Structural TBM that obtains socially optimal feedbacks on the utilization of a public-funded transportation infrastructure. This model leads to understanding the dynamics for governing the adoption of sustainable alternatives to driving with primary commute mode. Thus, a realistic decision-making approach is generated to improve the accuracy of travel decisions [12].

Fig. 3 summarizes the relational factors considered in travel decision analysis. Prime factors like dependence, satisfaction, temperature, and community possess additional sub-components. For instance, dependence has routes, mode of transportation and location. Routes may have different choices including clear route, rush route, less polluted route, or proper space for the vehicles. Mode of transportation has many options such as public transport, private vehicles, or bicycle walk, and location may have different choices like natural attraction, vacation planning, and business trip. Satisfaction factor depends on three parameters namely cost, facilities and time. Temperature has different effects on individuals depending on their age, young adults usually endure better compared to elders during winter. The community factor is dependent on whether the surrounding community emanating crowd or peaceful environment.

2) Travel Demand: Travel demand relies on the travel forecasting which depends upon varying parameters such as the most visited place, convenient hours, and seasons. There are several notable models that were developed to facilitate travel demand prediction analysis as discussed as follows.

Liang et al. developed Individual Travel Behavior (ITB) model focusing on identifying travel demand from public transport passengers travel activities. This model was built to analyze travel behavior individually for supporting feasible public transport implementation. Also known as the travel behavior graph-based model, ITB predicts individual graphs for passengers based on different public transport choices and demand, and provides an analysis of travel features with the help of customized public transport. There are several drawbacks of travel behavior graph-based models like fetching and scheduling the demand of research is higher with the optimization of the transport operating network justifiable on a large scale [14].

Regret-based choice models estimate travel demand on many levels and relate to various travel choice contexts. The models show outcomes for different traditional utility-based models in comparison with behavioral interpretations. The objective of this type of model is to present regret based decision-making process to provide more quality and behavioral insights. Regret-based choice models are enhanced in terms of the framework by better disentanglement in various sources of heterogeneity [13].

Heterogeneous Travel Behavior (HTB) aims to facilitate trip generation and simulation. The HTB model skips the entire path to an activity-based model and can predict the

TABLE I. CHARACTERISTICS COMPARISON BETWEEN VARIOUS TBM MODELS.

Reference	Model	Data Type	Data Size	Future Work
[8]	Travel Decision	GPS	260 individuals.	En-route prediction before and beginning of trip.
[9]	Travel Decision	Survey	67,000 trips, 3,500 individuals.	More accurate and effective travel decision.
[10]	Travel Decision	Survey	719 individuals.	Enhancement using socio-demographic indicators.
[9]	Travel Decision	Survey, mobile data	Not reported.	Transportation planning applications.
[53]	Travel Decision	Survey	170 individuals.	Enhancement of framework.
[11]	Travel Decision	Interviews	1st wave (303), 2nd wave (286), 3rd wave (258) individuals.	Enhancement of framework.
[12]	Travel Decision	Survey, Open Street map data	Not reported.	Realistic decision-making approach.
[14]	Travel Demand	Smart card data	6 passengers.	Automatic graph generation with larger scale computing.
[13]	Travel Demand	Survey	242 individuals.	Significantly improves the performance of the random regret minimization model to enhance the framework.
[16]	Travel Demand	Glove embedding vectors	6 billion input word vectors from 37 million texts.	Stochastic nature of the algorithm that requires careful experimental design.
[15]	Travel Demand	Survey	Trips range from 0 to 5, cross-classification model with 1.24 trips.	Inclusion of additional variables for better accuracy rate.

development of travel demand with least computer processing time. HTB possess many advantages such as prediction of true travel speeds, mode-defined constants, and improved assignment results. Another model is Microscopic Trip Generation Module that maintain an operating model while improving the research and its design. However, there are some limitations in the model such as lack of accuracy, auto-ownership, and microscopic tour generation [15].

Fig. 4 summarizes the relational factors considered in travel demand analysis like reviews and personal references. Reviews are collected through social media. Personal preferences are more dependent to each traveler. Comparison of models in the term of data type, size, and future improvement suggestion is summarized in Table I.

The survey findings indicates that travel decision-based TBM future direction eventually leads towards precise en-route prediction. Precising the prediction requires investigation on crucial factors like pre-planning, transportation accommodation planning, and effective travel decision framework. The prospect of precise en-route prediction is viable and have lasting impact towards adapting TBM with realistic decision-making capabilities. Furthermore, existing travel demand based TBMs were largely constrained or limited by technical components and scalability aspects. These TBMs often hindered by computing limitation or handling large-scale data for automatic graph generation. Improving these components can increase the performance of applied random regret minimization model and framework enhancement. The inclusion of additional variables from datasets with greater size are recommended to enrich or enhance the model prediction accuracy.

G. Travel Behavior Modeling Common Datasets

Existing travel behavior studies have worked with many travel datasets from various sources ranging from open data, crowdsourcing, and conventional surveys. In the following discussions, we present the summarization of several notable travel behavior datasets based on our conducted survey.

1) Point of Interest Dataset: This dataset was shared to help researchers in finalizing trip purpose parameter for developing destination prediction algorithm with higher accuracy [8]. This dataset is generated via crowdsourcing Open Street Map with 260 participants, allowing their vehicle's location to be recorded for every 70 minutes. The dataset comprises attributes including trip time, weekdays, origin, purpose, user, trip number and most common destination visited. The original XML data was drawn via Geographic Information Systems (GIS) in the form of layers of files linked with GPS data. This dataset features relational travel information of roughly estimated 26,000,000 points of interest in American states of Maryland, Washington D.C., and Virginia.

2) London's Oyster Smart Card & Survey Dataset: Since 2012, data generated by smart cards usage in London's transportation system has been used to create the dataset. There are two different smart cards and two different modes of transportation: tube and bus. 2.18 million journeys of 9708 passengers were recorded in the dataset. Furthermore, the smart cards dataset often paired with the London Travel Demand Survey dataset. The survey contains collected household information including household demographics, socio-economic information, and travel-related information. The survey inquires common information like ID, age, manager, total number of owned vehicles, household income, working status, occupation, weekly work frequency, daily commuting distance, frequency of becoming driver or passengers [57].

3) Los Angeles Dataset: Using survey with 352 participants from 219 households over 2395 days, this dataset is generated to map individual daily travel activities. The dataset contains variables including information on total number of daily trips distributed by types of transport, demographic, employment status, household, owned vehicle, and annual income. Daily individual travel activities were recorded in weekly basis with exclusion of repetitive activities in daily basis [58].

4) Neighborhoods for Active Kids Dataset: This dataset is generated to help management to build and implement active school travel planning. The dataset includes communities

TABLE II. Recommendations for Optimizing uses on Common TBM Datasets.

Datasets	Descriptions	Suggestion
Point of Interests [8]	Destination prediction	Develop more efficient algorithm and use age factor parameter in the datasets for understating about age group destination prediction.
London's Oyster Card and Survey [57]	Passenger's movement and travel demand forecasting	Develop an advisory system to be used during travel time.
Los Angeles [58]	Transportation professionals and policy makers consider shifting from the conventional one-day approach toward a multi-day approach.	Collaboration with IT industry and transportation industry for making bridge between them.
Neighborhoods for active kids [59]	Independent mobility by providing child friendly social and safe environment, traffic safety, and policies that promote local schools and safe vehicle-free zones around schools.	Development and enhancement of the ideas in terms of algorithm and software.
Kaunas City [60]	Travel time data collection	Develop a method for data security.
Bandung City [61]	Measurement of transport demand	Develop a strong policy for sustainable transportation in developing countries.
National Household Travel Survey (NHTS) - City of Long Beach [62]	Trip destination prediction	Develop an algorithm for trip designation prediction.
INRIX [63]	Traffic management	Development of traffic management algorithm with increased accuracy and efficiency.

to support independent mobility by providing traffic safety, social environment, local promotion school policy and safe vehicle areas around the children schools. Dataset variables consists of school travel mode and child characteristics such as age, gender, ethnicity, physical activity, child beliefs regarding traffic safety, neighborhood, independent mobility, and household characteristics. The household characteristics are education, employment, number of adults, number of children, car ownership, household beliefs, distance to the school, traffic safety, stranger danger, convenience, and social interaction. Social environment covers neighborhood, safety, neighborhood cohesion, neighborhood connection and other related factors such as residential density, street connectivity, high traffic exposure and low traffic exposure [59].

5) Kaunas City Dataset: This dataset original use is for facilitating state government to analyze car travel time pattern around the city. Developed using Python, a Geographic Information System (GIS) was created with the purpose of extracting, analyzing, and visualizing car travel data. The extracted data constitutes the dataset with information like coordinates of origin, coordinates of destination, mode of travel, date, and departure time [60].

6) Bandung Metropolitan City Dataset: Bandung Metropolitan City dataset recorded traveler's check-in behavior, comprising enriched insights on understand traveler's preferences. Personal verification is required including user ID, gender, date, time, and geo-location (longitude and latitude), and address. The attributes contained in this dataset is closely similar to Shanghai dataset including total number of check-ins, total number of processed check-ins, and total number of users distributed by gender [61].

7) National Household Travel Survey (NTHS) - City of Long Beach Dataset: The NHTS dataset is generated via conventional survey conducted in the city of Long Beach, California. The survey was conducted to gather consensus on environmental factors at trip destinations that affect non-motorized travel behavior in the area. The national household travel survey was conducted between 2008 to 2009 comprising information on motorized users, area density, diversity, and design at destinations that significantly affect mode choice decisions [62].

8) INRIX Dataset: The Floating Car Data (FCD) of Netherlands focuses on traffic information, traffic management and automated vehicles. The INRIX dataset recorded travel time of persons and goods from place of origin to destination (O-D) and hours of delay [63].

Based on the condition of presented datasets, several recommendations are suggested to help future TBM studies in optimizing uses of these datasets. The suggested recommendations for each dataset are summarized in Table II.

IV. Research Challenges in Travel Behavior Modeling Applications

TBM acts as simulator and generate decision-making insights and trends for travel decision and travel demand analysis. The output insights are used to strategize effective land use, vehicle sales, infrastructure investment and planning, and service improvement. Despite the many advantages, leveraging TBM optimally are challenging in many aspects including limitation of enabler technologies, data sources, and collection methods. The classifications based on our taxonomy reveals several emerging trends and challenges in utilizing TBM. We derived these challenges by analyzing available common parameters in recent TBM works and classified them based on elements in our TBM framework in Fig. 1. The research challenges are deliberately discussed as follows.

A. Technological Limitation

Many studies on TBM methods and technologies have been carried out over the years. However, these studies also encounter limitations like model inaccuracy or inefficiency, partial framework, transport-dependent planning, unrealistic decision manner, and inconstant socio-demographic. The survey results indicates that future work on TBM eventually heading towards the application of generalized probabilistic models. Future TBM possibly requires intricate programming development of behavior models for processing massive mobile apps data with increased efficiency and accuracy.

TABLE III. SUMMARIZATION ON CHALLENGES, RECOMMENDATIONS, AND FUTURE DIRECTIONS FOR TBM.

Challenges	Objectives	Recommendations	Future Directions
Behavior Modeling	- Predicting individual travel behavior - Individual TBM Graph	- Develop methods for suspicious behavior prediction - Advanced case detection on sudden behavior change	- Travel demand prediction - Automatic behavior graph generation
Destination Prediction	- Trip prediction purpose - Predicting destination location - Advanced pre-input route generation	- Long-term behavior prediction - Integrating emergency alert service - Remote prediction without internet dependencies	- En-route prediction of destination - Increasing prediction accuracy rate
Policy	- Feedback investigation - Travel planning environment - Schedule composition and activity duration	- False feedback prediction - Smart cities development support - Policy proposals	- Decision-making to match real-world behavior
Expensive Autonomous Vehicles	- Reducing public transportation - Ride sharing network promotion	- Public adaptation - Low-cost maintenance	- Low-cost development of autonomous vehicles

TBM is used in many instances including travel management, traffic management, medical access, healthcare planning, and sustainable urban planning. TBM can generate timely decision-making insights in case of additional input like dietary preferences, emergency events, or destination changes. Despite the many usage instances, TBM is in fact heavily used in many processes of traffic management and planning. Several methods employ TBM to facilitate traffic management like spatiotemporal monitoring [64], closed-circuit TV camera [65], smart vehicle technology [66], traffic control [67], and movement behavior [68].

B. *Quality of Generated Data for Modeling*

TBM output relies on the dataset quality it works with, depending on the reliability of its data source and collection method. Albeit the dataset quality often reduced due to data privacy measures, these measures are necessary to protect data sources and users from external threat. The prospect of combining TBM with effective user mobile data access control and monitoring are considered viable for future TBM. Accurate direct access can further enrich generated insights for user travel planning, preferences, routes, and safety using real-time information. Moreover, data quality from conventional surveys like questionnaire and interview are diminished by high inaccuracy rate due to invalid feedback and lack of verification.

Nowadays, social media usage serves as the prime data sources for analyzing travel behavior and experiences. Social media data holds rich information like user posts, opinions, or discussions on their travel experiences. However, real-time social media datasets are not available for the public due to provider's policies [31]. For instance, users can retrieve data generated by Google Maps via application programming, however, certain personal information like travel speed, gender, vehicles, and cost may be redacted. Therefore, mutual collaborations between researchers and social media providers in research should be intensified productively to fully utilize the potential of TBM.

Apart from social media, the mainstream use of ride-hailing services also generates many useful datasets for travel behavior studies. Travel datasets generated by ride-hailing services features unique perspectives on travel behavior like on-demand services, quick booking, and time-critical traveling. However, similar to policies restriction of social media, ride services like Uber and Grab does not share their data to the public [44]. Due to the nature of the business, corporate reputation is of utmost important that any allegation of misconduct or misuse are devastatingly damaging. Albeit the unique characteristics of the ride services datasets, its data quality is also affected by many external factors and dependencies. External factors like internet connectivity, unavailability of drivers during holidays, and apps usability eventually determines the quality of the generated datasets.

V. OPEN ISSUES AND OPPORTUNITIES FOR TRAVEL BEHAVIOR MODELING

This section discusses and summarize the identified open issues emerging from the research challenges found in the survey. Despite the recent advances that attempted to address the challenges, TBM itself revolves around human user which is a complex system. Considering the complexities and difficulties in understanding human behavior, future TBM research should focuses on personalization aspect to address the remaining issues. The remaining open issues are summarized in Table III and discussed as follows.

1) *Quality Behavior Modeling:* Modeling travel behavior duration is crucial in facilitating individual travel behavior prediction and model graph generation for supporting analysis on travel decisions and demands. The modeling procedure generates automatic travel behavior graph using computing [14] and models the behavior prediction [13]. The constant need for accurate rapid event prediction in person travelling activity, particularly if the behavior has become routine, is a persistent problem in behavior modelling. Therefore, methods for understanding sudden behavior changes in the middle of traveling trip is necessary to anticipate risks and maneuver. Future studies also can concentrate on enhancing travel activity data collection approaches and channels for generating insights on encouraging quality personal travel behavior.

2) *Real-Time Destination Prediction:* Destination prediction remains as one of the major issues in TBM that is influenced by many goal variables like en-route trip, destination location, and advance travel route information. Modeling destination prediction often constrained by dependencies on high accuracy rate of en-route prediction of destination [8]. In addition, the accuracy rate is often affected by internet connectivity especially for facilitating long-term travels or remote locations. To address this issue, travel application should provide comprehensive planning mode that generate insights by taking account trip preparation. This mode also should actively

identify mapping point with available communication access and store the coordinates for offline viewing. Furthermore, future studies should focus on the scope of facilitating real-time route changes with consideration of cost, duration, and safety.

3) Intelligent Travel Policies: Existing policies focuses on intelligent travel decisions including feedback investigation, environmental impact, schedule composition planning, activity duration, and emergency readiness. However, there are limited policy guidelines for encouraging quality travel decision-making that matches real-world travel behavior [12]. The increased possibility of faulty travel decision-making increases the likelihood of ensuing problem and its negative impact. Therefore, methods to differentiate between valid feedback and invalid feedback is critical to eliminate any change outlier during behavior modeling.

4) Cost-Efficient Autonomous Vehicles: Adaptation and usability of autonomous vehicles remains the main issue for anticipating future travel behavior and building ride-sharing practices. This issue stemmed from the persisting high cost of adopting or implementing autonomous vehicles in transportation ecosystem. Therefore, the development of low-cost autonomous vehicles is paramount for large-scale public adaptation into mainstream transportation [4], [5], [6].

VI. Conclusions and Future Work

In this paper, we presented overview insights on recent advances in TBM encompassing modeling classifications, emerging challenges, and arising opportunities. Leveraging insights from travel demand and travel decision analysis, TBM enables the prediction of future possibilities of travel scenarios. A taxonomy of travel behavior modeling along with the characterization and summarization of works on each classification were presented. Based on the survey, recent advances in TBM are distinctly classified based on algorithms, data sources, technologies, travel behavior, models, and datasets. Despite its heavy uses in facilitating transportation sectors, TBM also indirectly supports other sectors like tourism, marketing, and urban planning. Based on the reported challenges in recent TBM advances, we learn that model inaccuracy and inefficiency issues persist and largely constrained by current technological limitation and dataset qualities. The prospect of improving TBM via large scale integration and crowdsourcing are considered viable due to increased likelihood for better accuracy and efficiency. Based on the survey, we identified several notable issues and opportunities arising from recent TBM advances revolving aspects like behavior modeling, destination prediction, policy, and expensive autonomous vehicles. Our findings envision that future travel behavior are expected to change drastically due to many substantial aspect evolutions. Therefore, TBM is expected to play significant roles in realizing these visions. Researchers and industry analysts can investigate the TBM open issues and research opportunities highlighted in this survey for future undertaking. Future studies can also leverage the proposed taxonomy and identified challenges to enrich understanding on TBM for improvement purposes.

References

[1] Z. Kan, L. Tang, M.-P. Kwan, C. Ren, D. Liu, and Q. Li, "Traffic congestion analysis at the turn level using Taxis' GPS trajectory data," *Computers, Environment and Urban Systems*, vol. 74, pp. 229–243, 2019.

[2] H. M. Amer, H. Al-Kashoash, M. Hawes, M. Chaqfeh, A. Kemp, and L. Mihaylova, "Centralized simulated annealing for alleviating vehicular congestion in smart cities," *Technological Forecasting and Social Change*, vol. 142, pp. 235–248, 2019.

[3] A. Soteropoulos, M. Berger, and F. Ciari, "Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies," *Transport Reviews*, vol. 39, no. 1, pp. 29–49, 2019.

[4] J. Bischoff and M. Maciejewski, "Simulation of city-wide replacement of private cars with autonomous taxis in Berlin," *Procedia Computer Science*, vol. 83, pp. 237–244, 2016.

[5] L. M. Martinez and J. M. Viegas, "Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal," *International Journal of Transportation Science and Technology*, vol. 6, no. 1, pp. 13–27, 2017.

[6] W. Zhang, S. Guhathakurta, and E. B. Khalil, "The impact of private autonomous vehicles on vehicle ownership and unoccupied VMT generation," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 156–165, 2018.

[7] J. M. Dargay and M. Hanly, "The impact of land use patterns on travel behaviour," in *European Transport Conference, France*, 2003.

[8] C. M. Krause and L. Zhang, "Short-term travel behavior prediction with GPS, land use, and point of interest data," *Transportation Research Part B: Methodological*, vol. 123, pp. 349–361, 2019.

[9] C. Kim and O. Parent, "Modeling individual travel behaviors based on intra-household interactions," *Regional Science and Urban Economics*, vol. 57, pp. 1–11, 2016.

[10] K. Deutsch, S. Y. Yoon, and K. Goulias, "Modeling travel behavior and sense of place using a structural equation model," *Journal of Transport Geography*, vol. 28, pp. 155–163, 2013.

[11] F. E. Zarwi, A. Vij, and J. Walker, "Modeling and forecasting the evolution of preferences over time: A hidden Markov model of travel behavior," *arXiv preprint arXiv:1707.09133*, 2017.

[12] S. Feygin, "Inferring Structural Models of Travel Behavior: An Inverse Reinforcement Learning Approach," Ph.D. dissertation, 2018.

[13] S. Jang, "Regret-based travel behavior modeling: an extended framework," Ph.D. dissertation, 2018.

[14] Q. Liang, J. Weng, W. Zhou, S. B. Santamaria, J. Ma, and J. Rong, "Individual travel behavior modeling of public transport passenger based on graph construction," *Journal of Advanced Transportation*, vol. 2018, 2018.

[15] R. Moeckel, L. Huntsinger, and R. Donnelly, "From macro to microscopic trip generation: representing heterogeneous travel behavior," *The Open Transportation Journal*, vol. 11, no. 1, 2017.

[16] F. C. Pereira, "Rethinking travel behavior modeling representations through embeddings," *arXiv preprint arXiv:1909.00154*, 2019.

[17] J. C. Willems, "Models for dynamics," in *Dynamics reported.* Springer, 1989, pp. 171–269.

[18] Q. Qi, T. Weise, and B. Li, "Optimization algorithm behavior modeling: A study on the traveling salesman problem," in *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 2018, pp. 861–866.

[19] M. Wong and B. Farooq, "A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data," *Transportation Research Part C: Emerging Technologies*, vol. 110, pp. 247–268, 2020.

[20] D. Nam, H. Kim, J. Cho, and R. Jayakrishnan, "A model based on deep learning for predicting travel mode choice," Tech. Rep., 2017.

[21] A. N. P. Koushik, M. Manoj, and N. Nezamuddin, "Machine learning applications in activity-travel behaviour research: a review," *Transport Reviews*, vol. 40, no. 3, pp. 288–311, 2020.

[22] G. Sierpiński, M. Staniek, and I. Celiński, "Travel behavior profiling using a trip planner," *Transportation Research Procedia*, vol. 14, pp. 1743–1752, 2016.

[23] M. Berger and M. Platzer, "Field evaluation of the smartphone-based travel behaviour data collection app "SmartMo"," *Transportation Research Procedia*, vol. 11, pp. 263–279, 2015.

[24] S. Jamal and M. A. Habib, “Smartphone and daily travel: How the use of smartphone applications affect travel decisions,” *Sustainable Cities and Society*, vol. 53, p. 101939, 2020.

[25] S. F. Dias and V. A. Afonso, “Innovative Business Models in Tourism and Hospitality: Going Mobile?” in *Strategic Business Models to Support Demand, Supply, and Destination Management in the Tourism and Hospitality Industry*. IGI Global, 2020, pp. 164–184.

[26] M. Christian, “Mobile Application Development in the Tourism Industry and its Impact on On-Site Travel Behavior,” *Modul Vienna University*, 2015.

[27] Y. Guo and S. Peeta, “Impacts of personalized accessibility information on residential location choice and travel behavior,” *Travel Behaviour and Society*, vol. 19, pp. 99–111, 2020.

[28] B. G. Menon and B. Mahanty, “Modeling Indian four-wheeler commuters’ travel behavior concerning fuel efficiency improvement policy,” *Travel Behaviour and Society*, vol. 4, pp. 11–21, 2016.

[29] N. S. Daisy, M. H. Hafezi, L. Liu, and H. Millward, “Understanding and modeling the activity-travel behavior of university commuters at a large Canadian university,” *Journal of Urban Planning and Development*, vol. 144, no. 2, p. 4018006, 2018.

[30] M. Rizwan and W. Wan, “Big data analysis to observe check-in behavior using location-based social media data,” *Information*, vol. 9, no. 10, p. 257, 2018.

[31] S. A. Golder and M. W. Macy, “Digital footprints: Opportunities and challenges for online social research,” *Annual Review of Sociology*, vol. 40, pp. 129–152, 2014.

[32] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, “Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges,” *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.

[33] H. Cramer, M. Rost, and L. E. Holmquist, “Performing a check-in: emerging practices, norms and’conflicts’ in location-sharing using foursquare,” in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011, pp. 57–66.

[34] A. P. Timalsena, A. Marsani, and H. Tiwari, “Impact of Traffic Bottleneck on Urban Road: A Case Study of Maitighar–Tinkune Road Section,” in *Proceedings of IOE Graduate Conference*, vol. 5, 2017.

[35] J. Chang and E. Sun, “Location3: How users share and respond to location-based data on social,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, 2011.

[36] V. Jadhav, S. Raman, N. Patwa, K. Moorthy, and J. Pathrose, “Impact of Facebook on leisure travel behavior of Singapore residents,” *International Journal of Tourism Cities*, 2018.

[37] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[38] J. H. Lee, S. Gao, and K. G. Goulias, “Can Twitter data be used to validate travel demand models?” in *14th International Conference on Travel Behaviour Research*, 2015.

[39] J. H. Lee, A. W. Davis, S. Y. Yoon, and K. G. Goulias, “Activity space estimation with longitudinal observations of social media data,” *Transportation*, vol. 43, no. 6, pp. 955–977, 2016.

[40] E. Parra-López, J. Bulchand-Gidumal, D. Gutiérrez-Taño, and R. Díaz-Armas, “Intentions to use social media in organizing and taking vacation trips,” *Computers in Human Behavior*, vol. 27, no. 2, pp. 640–654, 2011.

[41] S. Hasan and S. V. Ukkusuri, “Social contagion process in informal warning networks to understand evacuation timing behavior,” *Journal of Public Health Management and Practice*, vol. 19, pp. S68–S69, 2013.

[42] C. Coffey and A. Pozdnoukhov, “Temporal decomposition and semantic enrichment of mobility flows,” in *Proceedings of the 6th ACM SIGSPATIAL international workshop on location-based social networks*, 2013, pp. 34–43.

[43] K. N. Habib, “Mode choice modelling for hailable rides: An investigation of the competition of Uber with other modes by using an integrated non-compensatory choice model with probabilistic choice set formation,” *Transportation Research Part A: Policy and Practice*, vol. 129, pp. 205–216, 2019.

[44] A. Henao, *Impacts of Ridesourcing-Lyft and Uber-on Transportation Including VMT, Mode Replacement, Parking, and Travel Behavior*. University of Colorado at Denver, 2017.

[45] L. Kröger, T. Kuhnimhof, and S. Trommer, “Does context matter? A comparative study modelling autonomous vehicle impact on travel behaviour for Germany and the USA,” *Transportation Research Part A: Policy and Practice*, vol. 122, pp. 146–161, 2019.

[46] H. Millward, M. H. Hafezi, and N. S. Daisy, “Activity travel of population segments grouped by daily time-use: GPS tracking in Halifax, Canada,” *Travel Behaviour and Society*, vol. 16, pp. 161–170, 2019.

[47] L. Joseph, A. Neven, K. Martens, O. Kweka, G. Wets, and D. Janssens, “Measuring individuals’ travel behaviour by use of a GPS-based smartphone application in Dar es Salaam, Tanzania,” *Journal of Transport Geography*, vol. 88, p. 102477, 2020.

[48] G. Lue and E. J. Miller, “Estimating a Toronto pedestrian route choice model using smartphone GPS data,” *Travel Behaviour and Society*, vol. 14, pp. 34–42, 2019.

[49] L. Su, J. Gao, and Q. He, “Towards Quality-Aware Big Data Integration for Crowdsourced Road Sensing System,” 2017.

[50] J. E. J. Kang, “Inferring Origin-Destination Demand and Utility-Based Travel Preferences in Multi-Modal Travel Environment Using Automatic Fare Collection Data,” Ph.D. dissertation, 2016.

[51] D. Wen and Q. Chunming, “Variational Inference for Agent-Based Models with Applications to Achieve Fuel Economy,” 2017.

[52] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, “The promises of big data and small data for travel behavior (aka human mobility) analysis,” *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.

[53] R. Krueger, A. Vij, and T. H. Rashidi, “Normative beliefs and modality styles: a latent class and latent variable model of travel behaviour,” *Transportation*, vol. 45, no. 3, pp. 789–825, 2018.

[54] S. Chakrabarti and K. Joh, “The effect of parenthood on travel behavior: Evidence from the California Household Travel Survey,” *Transportation Research Part A: Policy and Practice*, vol. 120, pp. 101–115, 2019.

[55] B. Devos and C. Werner, *Culture and cultural politics under Reza Shah: The Pahlavi state, new bourgeoisie and the creation of a modern society in Iran*. Routledge, 2013.

[56] J. A. Ansah, “Destination choice set definition in travel behaviour modelling,” *Transportation Research*, vol. 11, no. 2, pp. 127–140, 1977.

[57] Y. Zhang, T. Cheng, and N. S. Aslam, “Exploring the Relationship Between Travel Pattern and Social-Demographics using Smart Card Data and Household Survey,” *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[58] C. Li, L. Hou, B. Y. Sharma, H. Li, C. Chen, Y. Li, X. Zhao, H. Huang, Z. Cai, and H. Chen, “Developing a new intelligent system for the diagnosis of tuberculous pleural effusion,” *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 211–225, 2018.

[59] E. Ikeda, E. Hinckson, K. Witten, and M. Smith, “Assessment of direct and indirect associations between children active school travel and environmental, household and child factors using structural equation modelling,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 16, no. 1, p. 32, 2019.

[60] V. Dumbliauskas, V. Grigonis, and A. Barauskas, “Application of Google-based data for travel time analysis: Kaunas city case study,” *Promet-Traffic&Transportation*, vol. 29, no. 6, pp. 613–621, 2017.

[61] T. B. Joewono, A. K. M. Tarigan, and M. Rizki, “Segmentation, classification, and determinants of in-store shopping activity and travel behaviour in the digitalisation era: The context of a developing country,” *Sustainability*, vol. 11, no. 6, p. 1591, 2019.

[62] D. Kim, J. Park, and A. Hong, “The role of destination’s built environment on nonmotorized travel behavior: A case of Long Beach, california,” *Journal of Planning Education and Research*, vol. 38, no. 2, pp. 152–166, 2018.

[63] H. van der Loopa, M. Kouwenhovenb, P. van Bekkumc, and H. Meursc, “Validation and usability of floating car data for transportation policy research,” 2018.

[64] Y. Xu, Q.-J. Kong, and Y. Liu, “Short-term traffic volume prediction using classification and regression trees,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 493–498.

[65] V. Jain, A. Sharma, and L. Subramanian, “Road traffic congestion in the developing world,” in *Proceedings of the 2nd ACM Symposium on Computing for Development*, 2012, pp. 1–10.

[66] X. Chang, H. Li, J. Rong, Z. Huang, X. Chen, and Y. Zhang, "Effects of on-board unit on driving behavior in connected vehicle traffic flow," *Journal of Advanced Transportation*, vol. 2019, 2019.

[67] S. B. Raheem, W. A. Olawoore, D. P. Olagunju, and E. M. Adeokun, "The cause, effect and possible solution to traffic congestion on Nigeria road (a case study of Basorun-Akobo road, Oyo state)," *International Journal of Engineering Science Invention*, vol. 4, no. 9, pp. 10–14, 2015.

[68] Z. Li and J. Huang, "How to mitigate traffic congestion based on improved ant colony algorithm: A case study of a congested old area of a metropolis," *Sustainability*, vol. 11, no. 4, p. 1140, 2019.

A Novel Pornographic Visual Content Classifier based on Sensitive Object Detection

Dinh-Duy Phan[1], Thanh-Thien Nguyen[2], Quang-Huy Nguyen[3], Hoang-Loc Tran[4],
Khac-Ngoc-Khoi Nguyen[5], Duc-Lung Vu*[6]
Faculty of Computer Engineering, University of Information Technology
Vietnam National University Ho Chi Minh City, Viet Nam

***Abstract*—With the increasing amount of pornography being uploaded on the Internet today, arises the need to detect and block such pornographic websites, especially in Eastern cultural countries. Studying pornographic images and videos, show that explicit sensitive objects are typically one of the main characteristics portraying the unique aspect of pornography content. This paper proposed a classification method on pornographic visual content, which involved detecting sensitive objects using object detection algorithms. Initially, an object detection model is used to identify sensitive objects on visual content. The detection results are then used as high-level features combined with two other high-level features including skin body and human presence information. These high-level features finally are fed into a fusion Support Vector Machine (SVM) model, thus draw the eventual decision. Based on 800 videos from the NDPI-800 dataset and the 50.000 manually collected images, the evaluation results show that our proposed approach achieved 94.06% and 94.88% in Accuracy respectively, which can be compared with the cutting-edge pornographic classification methods. In addition, a pornographic alerting and blocking extension is developed for Google Chrome to prove the proposed architecture's effectiveness and capability. Working with 200 websites, the extension achieved an outstanding result, which is 99.50% Accuracy in classification.**

Keywords*—Computer vision; image proccessing; object detection; pornographic recognition and classification; blocking extension; machine learning; deep learning; CNN

I. Introduction

In the digital era, information has become a powerful weapon to manipulate the development of a society. People nowadays are easy to find and upload any information they want on the Internet. On the one hand, these pieces of information sometimes are good and bring positive value to the human race. On the other hand, many harmful kinds of negative information can also be found with only some keywords by anybody, even children. Pornography content is one of them. Many women worldwide are victims of sexual cybercrimes because their private videos are spread on the social network. Furthermore, pornographic content is even restricted in many countries. From the above problems, the need for an effective pornographic visual content detector is necessary.

Many efforts have been made recently to classify pornographic images among normal ones. In the early stages, the skin-based method has been applied. These approaches check whether images have nude people or not based on the ratio of the exposed skin. Another approach is handcrafted features-based, which uses various descriptors to extract key point low-level features in an image. A visual codebook may be learned by applying the k-means algorithm on a training set. After that, the trained codebook could represent any images, and a classifier may detect pornographic ones. Applying low-level features to identify obscene images, it achieved significantly higher performance than skin-based methods. However, representing images by visual words still suffers a severe problem since it ignores the spatial relationship, which is very important to represent the image's content. The state of the art approaches for this classifying problem are based on deep learning methods. These approaches build models with neural networks that let it learns features from the image's contents itself.

Previous studies [1], [2] [3], have implemented the above approaches and achieved some particular success, especially the deep-learning-based approach has proposed a potential development for this problem. However, the context in images is very complicated, and there are many similarities between a pornographic image and a normal one. The normal image that contains a large region of exposed skin (e.g., swimming, wrestling, people wearing bikinis) or contains people with sexy poses may be misclassified as pornography. Misclassification may seriously affect the user experience while using the internet. To avoid this problem, we have to clarify what is pornography. According to Oxford Advanced Learner Dictionary, "Pornography is magazines, DVDs, websites, etc., that describe or show naked people and sexual acts to make people feel sexually excited, especially in a way that many other people find offensive."[1] From the definition, an image can be determined as pornographic if it contains naked people. In other words, pornographic images are images that consist of human's sensitive objects and organs such as breasts, anus and genitals. We called this method is the sensitive object-based approach.

Based on that insight, this paper presents a novel approach for pornographic content detection and classification, which not only leverages the advantages of previous approaches but also compensates for these methods' weaknesses. Our main approaching strategy is using the effectiveness of object detection to identify pornographic elements in visual content with steady prediction. Additionally, skin and human recognition are also integrated into our method to distinguish between images with humans from images without humans, but with human like skin colors such as sand or wood. These two modules not only capable to augment the classification's decision but also can be served as the counterweight to prevent the potential bias that comes from object detection. Ultimately, a linear classification

[1] https://www.oxfordlearnersdictionaries.com/definition/english/pornography

model SVM is adapted to make the eventual prediction using features from these three modules, eliminate the limitations during the process while keeping their advantages.

In summary, our main contributions are as follows:

- A method is proposed to detect sensitive objects on the human body. Then, these objects were combined with skin and human features feeding to SVM model to conclude if an image is pornographic or not.
- Based on the pornographic textual and visual detection, a pornographic alerting and blocking extension is developed as an initial gate warning user before accessing inappropriate websites.

The rest of the paper is organized as follows. Section II describes details of some relevant approaches included skin-based approaches, handcrafted features-based approaches, and deep learning-based approaches. Section III presents theoretical background with four main models that we used to detect sensitive objects: Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN. The detail of our proposed method will be presented in Section IV. In order to demonstrate the applicability of the proposed model, a web-based extension will be introduced in section V. After that, Section VI provides the details of our dataset, experiments, and the achieved results. Finally, we give conclusions and suggest some future works in Section VII.

II. Related Work

Pornographic recognition and classification have been a long-lasting problem with many existing studies. In this paper, previous studies are grouped into four primary categories according to their main approach, which are: skin-based approach, handcrafted features-based approach, deep learning-based approach and object-based approach.

The skin-based approach is considered the earliest and most basic method of recognizing visual pornography, as it focuses on estimating the exposed skin area within an image. Several low-level or high-level features such as shape, color, facial, or belly can be utilized in order to achieved higher estimation. The final decision of this method mostly depended on mathematic or statistic thresholds based on the predicted skin ratio and the image's ground-truth. Previous studies with skin-based approach can be determinating seperate skin ratios on different body areas [4], combining facial recognition and skin recognition [1], or utilizing a pre-train discrimination neural network following a skin extractor [5]. Although the effectiveness of recognizing most cases of pornographic visual content, as [6] pointed out, the skin-based approach comes with vital weakness as its performance can be affected highly by the quality and resolution of the input image. Additionally, the strong resemblance between athletic or sporting images with a vast amount of exposed skin and pornographic images under the skin-ratio threshold can be a serious problem of recognizing the right one, which certainly affects the performance of skin-based methods.

The handcrafted feature-based approach, mostly applying the Bag-of-Visual-Word technique, extracts key-point features inside visual data using feature descriptors and maps it into vectors. To obtain features, various feature descriptors can be adapted such as: SIFT and Hue-SIFT [2]; BossaNova [7], which is an image representation based on histogram of distance; or Temporal Robust Features [8], a spatial-temporal interest point descriptor that adapts Fisher Vector intermediate representation. When these representative vectors are formed, a visual codebook can be constructed and concatenated with discrimination algorithm likes SVM to determine the pornographic of input visual content. However, the diversity and discrepancy of pornographic visual content along with the omission of spatial relationship make it difficult to determine the appropriate features to describe visual pornography comprehensively.

The robustness of deep learning-based methods in recent years has brought a significant result in visual pornography identification. Rather than adapting descriptors to obtain representative features manually, the advantage of neural network architecture helps model extract features and refine learning parameters themselves, thus improve the performance of predicting pornographic visual content. Previous studies with pornography detection [9], [10] utilize pre-trained neural network models and fine-tune on a custom pornography dataset instead of training the architecture from scratch. However, training strategies and network customization must be made to optimize architectural's parameters and prevent the model suffer from over-fitting.

Lastest studies about visual pornography [11], [12], [13] identify that sensitive objects, organs or body parts, likes vulva, dildo, female breast or anus, have been known to carry rich information of describing a major part of pornographic content. By recognizing these organs or objects with object detection models, it is possible to draw the conclusion about the pornographic of input visual content with high accuracy. Normally, studies under the object-based method often adapted exist object detector and fine-tuned it on a custom annotated dataset. However, the selection of appropriate pornographic organs or objects as well as the method of annotating depended heavily on the study scale and perspective. Noticeably, Tabone et al. [14] proposed seven sexual organs and objects for pornography classification included buttocks, female breast, female genital (which are divided into two sub-classes: female genital posing and female genital active), male genital, sex toy and benign object. Eventually, they annotated those classes with a five-set labeled point: one center point and four perpendicularly offset for each. Additionally, some common sexual poses or intercourses can be learned in pattern to improve the performance of prediction. Although methods under object-based approach can ensure the high performance of prediction on most pornographic visual data, in some special cases, the strong resemblance of some everyday object with sexual organs under certain conditional (light, viewpoint, or shape) such as sausage and male genital make it difficult to make the right decision.

In our previous studies with sensitive objects and organs on object-based approach in [15] [16], not only we developed a pornographic object detector with Mask R-CNN to identify the most four common sensitive objects includes women breast, male/female genitals and anus, but also we utilized the training strategy of object detection model with two-step learning to overcome the false positive prediction of pornographic objects, thus improve the overall performance of prediction.

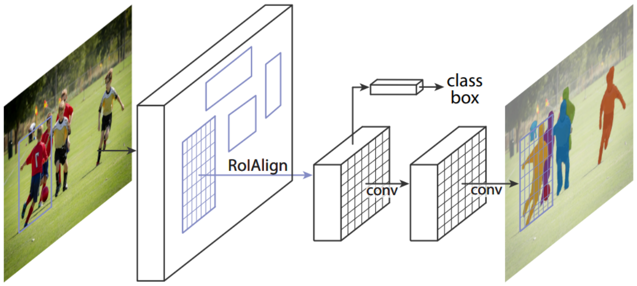


Fig. 1. Mask R-CNN Main Architecture [17].

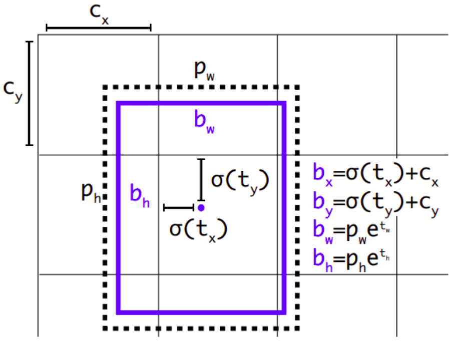


Fig. 2. YOLO Bounding Box with Location Prediction [21].

III. Theoretical Background

A. Mask R-CNN

Mask R-CNN [17] is a robust object detection and instance segmentation framework that draws bounding boxes for object detection and generates a high detail segmentation mask for instance segmentation. In this paper, the Matterport's version of Mask R-CNN [18] is adapted, which uses ResNet101 [19] and Feature Pyramid Network (FPN) [20] as main backbone.

The Mask R-CNN model included multiple stages and branches from the original Faster R-CNN framework, whith the main architecture can be seen in Fig. 1. Tts network hierarchy included the Region Proposal Network branch for object bounding boxes identification and Box Regression Network branch for bounding box regression. Furthermore, a simple FCN is added to predict the segmentation masks on each Region of Interest (RoI) in a pixel-to-pixel manner. Furthermore, for the core operation of feature extraction, Mask R-CNN applies pooling algorithm RoIAlign to extract small feature maps called Region of Interest (RoI) features to be well aligned and preserved with the corresponded object in the pixel level. Hence, RoIAlign helps Mask R-CNN achieves pixel-to-pixel alignment and high accuracy in prediction and mask segmentation. On each RoI, the new multi-task loss function of Mask R-CNN has been introduced, which combines the loss of classification, localization, and segmentation mask by Eq. 1.

$$\mathcal{L}_{\text{Mask R-CNN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} \tag{1}$$

The loss function $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{box}}$, which is adapted in Mask R-CNN, are defined by Eq. 2 and Eq. 3 respectively, with p_i, p_i^*, t_i, t_i^* are the predict class, ground truth label, predicted bounding box coordinates and ground truth bounding box coordiniates.

$$\mathcal{L}_{\text{class}} = \frac{1}{N_{\text{cls}}} \sum_i -p_i^* \log p_i - (1 - p_i^*) \log(1 - p_i) \tag{2}$$

$$\mathcal{L}_{\text{box}} = \frac{\lambda}{N_{\text{box}}} \sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*) \tag{3}$$

On the other hand, the loss function of instance segmentation mask $\mathcal{L}_{\text{mask}}$ is defined as the average binary cross-entropy loss between the ground-truth mask y_{ij} and the predicted mask y_{ij}^k, only in terms of class k-th in Eq. 4:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [y_{ij} \log y_{ij}^k + (1 - y_{ij}) log(1 - y_{ij}^k)] \tag{4}$$

B. Yolo

You Only Look Once (YOLO) is the state-of-the-art, real-time object detection system proposed by Joseph Redmon. YOLO [22] uses a single neural network to predict bounding boxes and class probabilities directly from full images in one evaluation. In YOLOv2 [21], they remove the fully connected layers from YOLO and use anchor boxes to predict bounding boxes. This research applied YOLOv3 [23], the latest version of YOLO using Darknet53 network architecture that helps the model achieves the outstanding benchmark comparing to others. Fig. 2 illustrates how YOLO model predicts the bounding box location on the image. Network model identifies 4 coordinates of the bounding boxes t_x, t_y, t_w, t_h, then the bounding box predictions is calculated by Eq. 5:

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \end{cases} \tag{5}$$

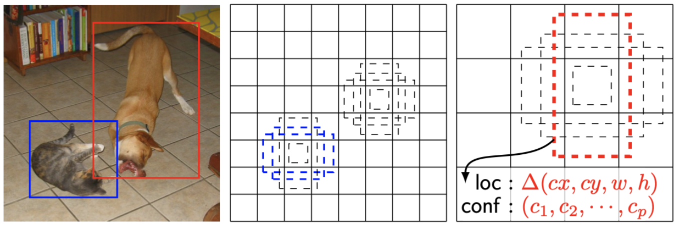
with the the top left corner point of the predicted box(c_x, c_y) as well as the width and height p_w, p_w, respectively. To optimized training parameters, YOLO adapted the loss function as shown in Eq. 6

$$\mathcal{L}_{\text{YOLO}} = \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{loc}} \tag{6}$$

in terms of

$$\begin{aligned} \mathcal{L}_{\text{loc}} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} \Big[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \\ + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \Big] \end{aligned} \tag{7}$$

and


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

Fig. 3. SSD Framework. [24] (a) SSD only needs an Input Image and Ground Truth Boxes for Each Object During Training. In a Convolutional Fashion, SSD Evaluates a Small Set (e.g. 4) of Default Boxes of Different Aspect Ratios at Each Location in Several Feature Maps with Different Scales (e.g. 8 x 8 and 4 x 4 in (b) and (c)).

$$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[1_{ij}^{\text{obj}} + \lambda_{\text{noobj}}(1 - 1_{ij}^{\text{obj}})\right](C_{ij} - \hat{C}_{ij})^2 + \sum_{i=0}^{S^2}\sum_{c\in\mathcal{C}} 1_i^{\text{obj}}\left[p_i(c) - \hat{p}_i(c)\right]^2 \quad (8)$$

C. SSD

Single Shot Multibox Detector (SSD) is proposed by Wei Liu et al. [24] which is known as the object detection model with high accuracy and speed. Unlike the other models, which need to hypothesize bounding boxes, followed by some complicated steps to handle them, SSD uses a fixed set of default boxes, then predicts scores and box offsets by using small convolutional filters to feature maps as illustrate in Fig. 3. One of the advantages of SSD is that it can detect objects of mixtures of scales with high accuracy. To achieve that, SSD produces predictions of different scales from feature maps, which are also in various scales and explicitly separates predictions by aspect ratio. For the balance between swiftness and precision, SSD runs a convolution network on the input image only once and then computes the feature map. SSD framework can be summarized as follows:

- The training phase inputs images with ground truth boxes for each object.
- With different scales of images, SSD evaluates a small set of default boxes in different aspect ratios at each location in several feature maps.
- For each default box, the framework predicts both offsets and confidences of shape for all object categories. At the training phase, SSD tries to match these default boxes to the ground truth boxes.

During the training process, SSD adapts the multi-task loss function described in Eq. 9:

$$\mathcal{L}_{\text{SSD}} = \frac{1}{N}(\mathcal{L}_{\text{conf}} + \alpha\mathcal{L}_{\text{loc}}) \quad (9)$$

in terms of:

$$\mathcal{L}_{\text{loc}} = \sum_{i,j}\sum_{m\in\{x,y,w,h\}} 1_{ij}^{\text{match}} L_1^{\text{smooth}}(d_m^i - t_m^j)^2 \quad (10)$$

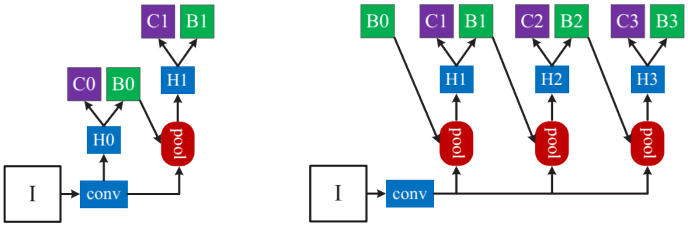


Fig. 4. Architecture of Faster-RCNN (left) and Cascade-RCNN (right) [25].

and

$$\mathcal{L}_{\text{cls}} = -\sum_{i\in\text{pos}} 1_{ij}^k \log(\hat{c}_i^k) - \sum_{i\in\text{neg}} \log(\hat{c}_i^0), \text{ where } \hat{c}_i^k = \text{softmax}(c_i^k) \quad (11)$$

D. Cascade Mask R-CNN

Cascade R-CNN [25] is a multi-stage object detection based on Faster R-CNN. It adds two extra stages to the standard two-stage Faster R-CNN architecture, as can be seen in Fig. 4. Besides, the data of each stage is trained with increasing IoU thresholds, which aims to reduce the overfitting problem between the training and testing processes.

The Cascade R-CNN architecture can be integrated with other models to improve the performance. For experiments, we use Cascade Mask R-CNN [26], a combination of Cascade R-CNN and Mask R-CNN, on Detectron2 [27]. As being the extended version of Mask R-CNN, Cascade Mask R-CNN shares a similar loss function with Mask R-CNN, as shown in Eq. 12.

$$\mathcal{L}_{\text{Cascade Mask R-CNN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} \quad (12)$$

IV. Methodology

In this article, instead of using low-level image features, a novel model for extracting high-level features, which can be used for classification afterward is proposed. The proposed method consists of two stages as shown in Fig. 5. The first stage is the high-level feature extraction which combines 10 high-level features from 3 blocks: sensitive objects detection (8 features), Human presence (1 feature) and Skin color extraction (1 feature). The second stage of the model is an SVM classifier for making the final decision of discriminating pornographic or non-pornographic content. The details of each block are described below.

A. First Stage - High-level Feature Extraction

1) Sensitive object detection module: When examining pornographic visual contents, including images and videos, we realize that the sensitive objects in the human body appear frequently, so that these objects can be used for detecting pornographic content effectively. In other words, we can apply object detection models to detect these objects and use it as high-level features. Throughout the image, four sensitive

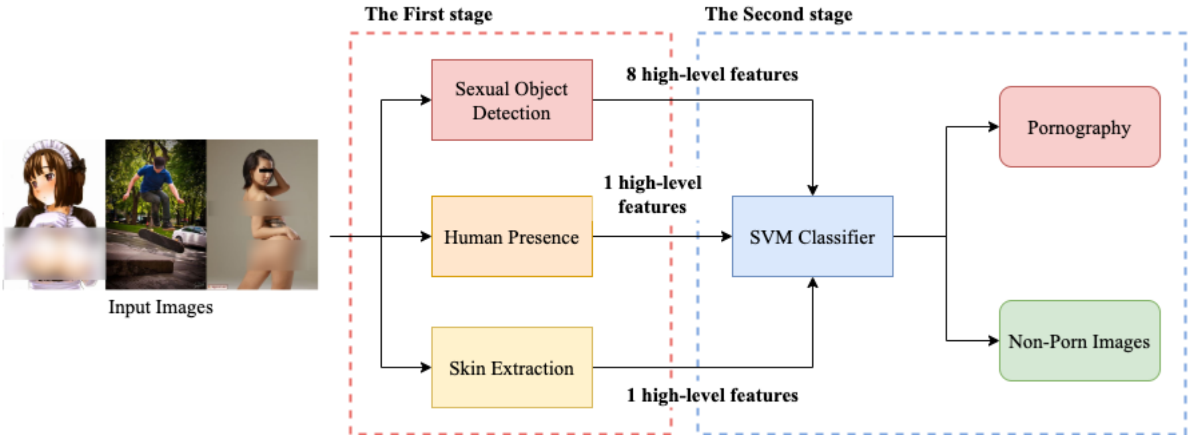


Fig. 5. Block Diagram of the Proposed Pornography Classifier.

objects and organs included female breasts, male/female genitals, and anus, are identified along with bounding boxes and corresponding confidence scores. Two high-level features are extracted for each type of sensitive object, including the number of detected objects and each type's highest confidence score, thus making eight high-level features for feeding to the main fusion model.

2) *Human presence module:* Images or videos are considered sexual if and only if it contains people or parts of people, therefore the second information can be used to solve the task in this paper is human presence. In other words, human detection is a sub-task of the object detection problem, thus the object detection model can be adapted to identify humans' presence on visual content. The existing human in images detected is another high-level feature we proposed to improve our model.

In this module, a human detector on Detectron 2, which used the pre-trained model on COCO dataset, is adapted to identify humans' existence from the input image. The segment mask of detected people was then used to recognize these people's presence on the image and adapted as the input for the skin extraction module. Thus the feature of human presence, in terms of a binary value, is adapted to the main fusion model.

3) *Skin extraction module:* For the skin extraction, we adapted two color spaces, HSV and YCbCr, to recognize skin areas from the image. While HSV is being used for its advantage in describing a high-quality color as well as reducing the problem of illuminating color identification, YCbCr is adapted for its advantage to describe skin on various races with significant results. Moreover, YCbCr is one of the most popular color spaces applied to skin extraction methods.

The color spaces thresholds that we applied for skin extraction in our proposed approach is described in Eq. 13. Based on these boundaries, we decided whether the pixel is describing skin or not by combining two extracted results from HSV and YCbCr range.The ratio of the total body skin areas on the image, which value varies between 0 and 1, as another high-level feature along nine upper describing features.

$$\begin{cases} 0 \leq H \leq 17 \\ 15 \leq S \leq 170 \\ 0 \leq V \leq 255 \end{cases} \text{and} \begin{cases} 0 \leq Y \leq 255 \\ 85 \leq Cb \leq 135 \\ 13 \leq Cr \leq 180 \end{cases} \tag{13}$$

The skin extraction task is performed on the human segmentation areas, which we get from the Human Presence module described above. We calculated the skin ratio only on human segments, not on the entire original image. This gives a better view of the human body's skin ratio, where the higher exposed skin, the higher pornographic probabilities.

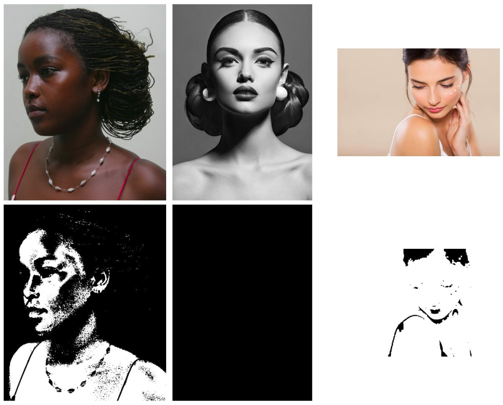


Fig. 6. Skin Extraction Algorithms Results.

However, we must admit that the skin extraction algorithm has a certain weakness when it deals with gray-scale or black/white images as well as images with skin-liked color, such as sand, wood, etc. Under these cases, the skin extraction neither working nor achieving great results, as described in Fig. 6. Thus, that will be one of the problems we have to overcome in the future.

B. *The Second Stage - SVM Classifier*

Ten high-level features from the three modules above then feed into a discriminative SVM model, which works as a fusion mechanism, to become the proposed approaching method. Generally, the overview of the fusion model is illustrated in

Fig. 5 where the sensitive object detection module can be replaced by Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN model, respectively. For the SVM model, kernel Radial Basis Function (RBF) is adapted as it comes with our proposed method's highest performance.

In the proposed model, sensitive object information plays the most crucial part in determining the sexuality of input visual content as it accounts for 80% of input features feeding to the fusion module. Moreover, the presence of humans from input images and the ratio of explicit body skin on these human segments play an additional role in ensuring the effectiveness of the method's performance. Ten high-level features feeding to the SVM model can leverage their effectiveness in prediction as well as eliminate the limitations from their own modules.

V. Pornographic Blocking Extension for Website

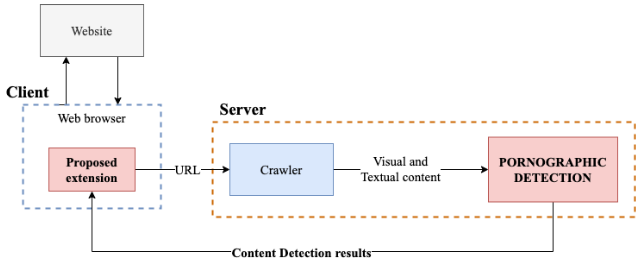


Fig. 7. Procedure of Pornographic Blocking Extension.

From our previous work in [28], [29] and [15], we developed an extension on Google Chrome for pornographic website recognition works as an initial gate alerting and blocking people before accessing the unappropriated website. The extension involves pornographic textual content identification using the NaiveBayes classifier and visual content identification based on YOLO sexual object detection to determine the pornography of the input website.

A. Procedure Flow

The main procedure of this extension can be observed in Fig. 7. Initially, visual and textual contents are crawled from the input website on the crawler module. Then, pornographic visual and textual identification modules are adapted to identify how pornography the website is.

In the visual identification module, a sensitive object detection was adapted to determined if website images are pornography or not. Due to the fast execution and great performance, YOLOv3 was chosen for the extension. On the other hand, the Vietnamese/English pornographic textual classifier was adapted from [29] for the textual identification module. This classifier uses the NaiveBayes algorithm to discriminate whether the sensitiveness of textual contents from the website.

Based on the predicted results of these contents, the extension alerting three safety levels to the user, which are: (1) Safety: textual and visual contents are both recognized as safety; (2) Porn-Type-1: one of the textual or visual content is recognized as pornography; (3) Porn-Type-2: textual and visual contents are both recognized as pornography.

TABLE I. The Website Dataset

	Safety	Porn-Type-1	Porn-Type-2	**Total**
Vienamese	65	5	11	81
English	35	15	69	119
Total	100	20	80	200

TABLE II. Result on the Manual Dataset

Object detection model	**Object-based method (object detection)**	**Proposed method (Skin + Human + object detection)**
Mask R-CNN	81.64%	**85.63%**
YOLO	93.87%	**94.06%**
SSD	88.93%	**89.32%**
Cascade Mask R-CNN	**93.43%**	93.39%

B. Pornographic Website Dataset

To evaluate the performance of the pornographic alerting and blocking extension, we collected a website dataset including 100 pornographic websites and 100 safety websites with Vietnamese and English languages. The websites are divided into three categories: Safety, Porn-Type-1, and Porn-Type-2 corresponded with the definitions of extension outcome notifications. The distribution of the pornographic website dataset by category is described in Table I.

VI. Experimental Results

A. Results and Evaluation

On the experiment, we evaluated our method with a manual image dataset along with the public dataset NPDI-800 [7]. The manual image dataset includes 25.000 normal and 25.000 pornographic images, while the NPDI-800 dataset contains 400 pornographic videos and 400 normal (non-porn) videos.

For video evaluation, one frame per second is extracted to determine the pornography of input video. The evaluating strategy of pornographic video is modelized as shown in Fig. 8. As can be seen, extracted frames are classified into two groups, namely pornographic frames, and normal frames, by a classifier. Based on the ratio of porn frames on total frames, the input video is classified as pornographic or normal.

Fig. 9 and Fig. 10 show our experiments on NPDI-800 dataset with a various "porn rate" threshold from 1% to 15%. With that strategy, we achieved the best result when we considered pornographic video contain 1% or more frames detected as pornographic. To assess the effectiveness of our proposed approach, we have done two experiments on the same testing set: (i) using object detection's outputs directly to distinguish pornography from normal images, (ii) feeding the SVM with object detection's output information combining with extracted skin and human segmentation features.

TABLE III. Result on NPDI-800 Dataset

Object detection model	**Object-based method (object detection)**	**Proposed method (Skin + Human + object detection)**
Mask R-CNN	54.38%	**74.13%**
YOLO	94.88%	**94.88%**
SSD	89.00%	**89.13%**
Cascade Mask R-CNN	76.63%	**82.75%**

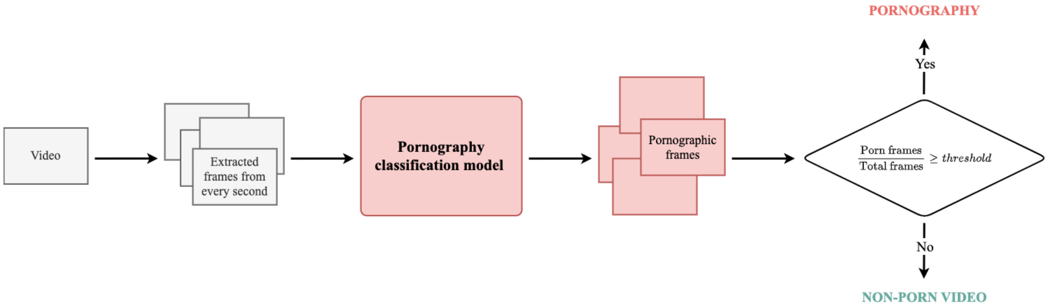

PORNOGRAPHY
Yes
Video
Extracted frames from every second
Pornography classification model
Pornographic frames
$\frac{\text{Porn frames}}{\text{Total frames}} \geq threshold$
No
NON-PORN VIDEO

Fig. 8. Evaluating Strategy on NPDI-800 Dataset.

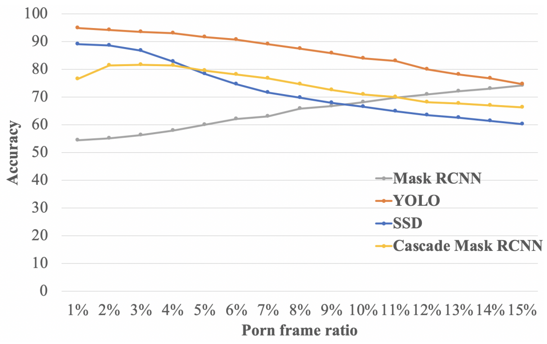


Fig. 9. Using Object Detection Approach to Evaluate on NPDI-800 Dataset with Different Porn Frame Ratios.

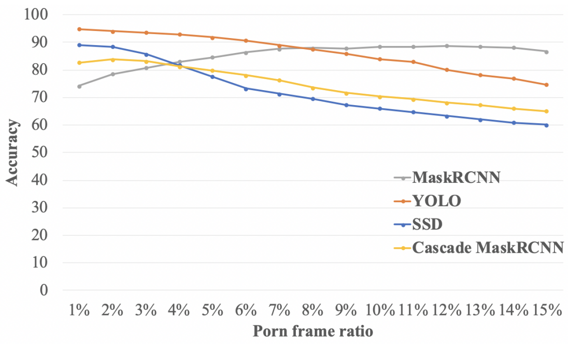


Fig. 10. Using Proposed Method to Evaluate on NPDI-800 Dataset with Different Porn Frame Ratios.

TABLE IV. Compared Results on the NPDI Dataset

Methods	**NPDI-800**
SIFT & Hue-SIFT Decriptor [2]	84.60%
BossaNova + SVM [7]	89.50%
DCNN-based Learning [11]	97.50%
BossaNova Video Descriptor [30]	92.40%
Deep Multicontext Network [3]	85.30%
Our model	**94.88%**

TABLE V. Performance of Extension in Alerting Websites

Website	Ground-Truth	Correct-Predicted	Accuracy
Safety	100	99	99.00%
Porn-Type-1	20	20	100.0%
Porn-Type-2	80	62	77.50%
Total	200	181	**90.05%**

TABLE VI. Performance of Extension in Classifying Websites

Website	Ground-Truth	Correct-Predicted	Accuracy
Non-Porn	100	99	99.00%
Porn	100	100	100.0%
Total	200	199	**99.50%**

The first experiment was done by using the results from sensitive object detection directly to judge an image is pornographic or not. If the model detects at least one of four sensitive objects in an image, this image will be concluded as pornography and vice versa. From the results on the manual dataset, which showed in Table II, the YOLO model achieved the best Accuracy score with 93.87%. While evaluating videos, this approach gives us the highest accuracy of 94.88% when the YOLO model is used to detect sensitive objects on the NPDI-800 dataset. Detail result at 1% of porn rate ratio has been shown in Table III. As shown in Fig. 9, both YOLO and SSD models had the best result at 1% porn frame ratio while Mask R-CNN achieved its best at 15% and this number was 3% of porn frame ratio on Cascade Mask R-CNN.

The second experiment was conducted by adapting our proposed approach, which used an SVM with ten high-level features inputs from sensitive object detection, human presence, and skin extraction modules. Generally, we achieved better results in comparison with the method using sexual object detection algorithms only, which can be observed on Table II and Table III. On the NPDI-800 dataset, Mask R-CNN and Cascade Mask R-CNN's results improved significantly when adapting the proposed approach, from 54.38% to 74.13% and 76.63% to 82.75%, respectively. However, the highest results still are achieved by the YOLO model with 94.06% accuracy on our manual dataset and 94.88% on the NPDI-800 dataset. The comparison with other methods on the NPDI-800 dataset can be observed in Table IV. As can be seen, the performance of our method in the NPDI-800 dataset is quite impressive.

B. Website Extension Experiment

In the experiment with pornography blocking extension, our manual website dataset is adapted to evaluate the effectiveness of recognizing not-safe-for-work websites. The results in prediction on three types of website can be observed on Table V and Table VI.

As can be observed, the extension achieved the outstanding results with 99.00% Accuracy in recognizing Safety websites, 100.0% and 77.50% Accuracy on Porn-Type-1 and Porn-Type-2 websites, respectively. However, the extension recognizes Porn-Type-2 websites with only 77.50% Accuracy as some certain websites are identified into Porn-Type-1. Thus, all the pornographic websites during the experiment are identified by the extension with the absolute 100% Accuracy in recognition, leading to the total result of 99.50% Accuracy in porn/non-porn website classification. Still, we have to notice that there might be some confusing cases, which are websites containing sensitive contents. These kinds of websites might be civil intent, such as sex education lecture/advice or nude photographs for medical purposes. This makes the model quite difficult to determine the suitable decision, which might reduce our extension's performance in the real practical implementation.

VII. Conclusion

This paper proposed a new approach to identify pornographic visual content based on sensitive object detection and skin color. The proposed approach detects four sensitive objects including female breast, male or female genital and anus. Then these sensitive object information are used as high-level features along with human precense and skin extraction information as input of the fusion SVM model. Eventually, the SVM model decides whether the input visual content is pornography or not. Applying the four most notable object detection algorithms, which are Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN, sensitive object features are identified. We achieved the best results on our custom dataset and the public NPDI-800 dataset, which are 94.06% and 94.88% Accuracy respectively, when the YOLOv3 is adapted as the sensitive object detector and RBF is used as the SVM kernel.

In addition, to prove the effectiveness of our proposed method in practical application, an extension for alerting and blocking pornographic website is built. Measuring the accuracy by surfing 200 websites, the extension has shown impressive results: 99.00% of normal websites and 100% of pornographic websites was identified correctly.

Acknowledgment

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-26-02.

References

[1] R. Balamurali and A. Chandrasekar, "Multiple parameter algorithm approach for adult image identification," *Cluster Computing*, vol. 22, no. 5, pp. 11 909–11 917, 2019.

[2] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, and A. de A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *2009 17th European Signal Processing Conference*, 2009, pp. 1552–1556.

[3] X. Ou, H. Ling, H. Yu, P. Li, F. Zou, and S. Liu, "Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 5, pp. 1–25, 2017.

[4] D. C. Moreira and J. M. Fechine, "A machine learning-based forensic discriminator of pornographic and bikini images," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[5] K. Zhou, L. Zhuo, Z. Geng, J. Zhang, and X. G. Li, "Convolutional neural networks based pornographic image classification," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, 2016, pp. 206–209.

[6] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, and A. Sali, "An automated anti-pornography system using a skin detector based on artificial intelligence: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 04, p. 1350012, 2013.

[7] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[8] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space–time," *Forensic science international*, vol. 268, pp. 46–61, 2016.

[9] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neurocomputing*, vol. 210, pp. 283–293, 2016.

[10] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting nsfw images," *Retrieved August*, vol. 24, p. 2018, 2016. [Online]. Available: yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for

[11] Y. Wang, X. Jin, and X. Tan, "Pornographic image recognition by strongly-supervised deep multiple instance learning," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 4418–4422.

[12] C. Tian, X. Zhang, W. Wei, and X. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6629–6645, 2018.

[13] H. A. Nugroho, D. Hardiyanto, and T. B. Adji, "Nipple detection to identify negative content on digital images," in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 43–48.

[14] A. Tabone, A. Bonnici, S. Cristina, R. A. Farrugia, and K. P. Camilleri, "Private body part detection using deep learning," in *ICPRAM*, 2020, pp. 205–211.

[15] Q. H. Nguyen, K. N. K. Nguyen, H. L. Tran, T. T. Nguyen, D. D. Phan, and D. L. Vu, "Multi-level detector for pornographic content using cnn models," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–5.

[16] H. L. Tran, Q. H. Nguyen, D. D. Phan, T. T. Nguyen, D. L. Vu *et al.*, "Additional learning on object detection: A novel approach in pornography classification," in *International Conference on Future Data and Security Engineering*. Springer, 2020, pp. 311–324.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[18] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow https://github.com/matterport," *GitHub*, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[21] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[25] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[26] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4969–4978.

[27] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2 https://github.com/facebookresearch/detectron2," *GitHub*, 2019.

[28] T. A. Dinh, T. B. Ngo, and D. L. Vu, "A model for automatically detecting and blocking pornographic websites," pp. 244–249, 2015.

[29] D. D. Phan, V. T. Nguyen, and D. L. Vu, "Nhan dang trang web co noi dung khieu dam dua tren text va website," in *Fundamental and Applied Information Technology (FAIR) Domestic Conference in Vietnam*, 2019.

[30] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. Guimarães, and A. d. A. Araújo, "A mid-level video representation based on binary descriptors: A case study for pornography detection," *Neurocomputing*, vol. 213, pp. 102–114, 2016.

Deep Learning-based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks

Diego A. Andrade-Segarra[1], Gabriel A. León-Paredes[2]
Grupo de investigación en Cloud Computing Smart Cities & High Performance,
Universidad Politécnica Salesiana,
Cuenca, Ecuador, 010102

Abstract—**Due to TIC development in the last years, users have managed to satisfy many social experiences through several digital media like blogs, web and especially social networks. However, not all social media users have had good experiences with these media. Since there are malicious users that are able to cause negative psychological effects over people, this is called cyberbullying. For this reason, social networks such as Twitter are looking to implement models based on deep learning or machine learning capable of recognizing harassing comments on their platforms. However, most of these models are focused on the use of English language and there are very few models adapted for Spanish language. This is why, in this paper we propose the evaluation of an RNN+LSTM neural network, as well as a BERT model through sentiment analysis, to perform the detection of cyberbullying based on Spanish language for Ecuadorian accounts of the social network Twitter. The results obtained show a balance between execution time and accuracy obtained for the RNN + LSTM model. In addition, evaluations of comments that are not explicitly offensive show a better performance for the BERT model, which outperforms its counterpart by 20%.**

Keywords—***Bidirectional Encoder Representations from Transformers, BERT; Cyberbullying; Natural Language Processing; Recurrent Neural Network + Long Short Term Memory; RNN+LSTM; Sentiment Analysis; Semantics; Spanish Language Processing***

I. Introduction

Nowadays, the constant information and communication technologies (ICTs) development has enabled interpersonal interaction, allowing real experiences to be transferred to a virtualized medium such as social networks (SSNs) [1]. Within this environment, SSNs users can establish real time conversations to exchange ideas [2]. This satisfies the need for affection and integration through positive reinforcers and facilitates socialization among its users [3].

However, there are not only positive reinforcements for the use of social networks, but there are also risks such as, emotional distancing, loss of limits in communication, sexting, cyber addiction, or cyberbullying [4]. In cyberbullying, malicious users take advantage of situations of vulnerability to attack others through offensive comments using digital media. Although cyberbullying does not produce physical injuries, it causes negative psychological effects and disorders in users who are victims of these practices [5]. For this reason, social platforms such as Twitter offer methodologies based on social systems for cyberbullying recognition, such as Social Filter or the Theory of Planned Behavior, have become increasingly popular [6]. These methodologies are able to identify stalkers based on the allegations in their messages and the behavior of the account. However, these techniques tend to be ineffective in the face of occasional or unidentifiable comments.

Similarly, different countries take preventive measures through campaigns or laws that limit and sanction this type of practice. In the case of Ecuador, the Ministry of Education agreed on an operational definition for this type of practices within educational environments, described as violent acts that are frequently carried out intentionally between students of an educational institution, in a relationship of imbalance of power. Through it, a bully seeks to assert superiority in a group. [7]. For its part, the National Agency for Intergenerational Equality establishes that the competent authorities and entities must create an inter-institutional strategy to prevent, detect and address all forms of harassment within educational institutions (*bullying*, *ciberbullying*, sexual harassment) [8]. One of the most common forms of harassment in Ecuador, is cyberbullying which has a 7.46% share of the total, which puts it ahead of the total number of cases where victims were beaten in the same period with 7.02% [7]. It is worth mentioning that these statistics are limited to cyberbullying in educational environments.

II. Background

Following the ideas on [9], [10], [11] they carry out campaigns through images, talks, marches and trends in SSNs on tolerance and social respect to prevent cyberbullying. Despite this, the campaigns carried out, being a preventive method, are not able to act on a perpetuated act or event of cyberbullying. For this reason, there are methodologies based on deep learning and machine learning capable of identifying comments with potential cyberbullying [12]. Based on this resource, it is possible to take the necessary measures against the users involved. In this regard, deep learning offers cyberbullying prediction models based on Support Vector Machine (SVM), Linear Regression (LR) and Naive Bayes (NB) [13], [14], [15], [16].

However, the studies carried out are limited to work in the English language, which results in greater vulnerability for Spanish-speaking user groups in SSNs.

For this reason, works such as [17] and [18] test SVM, LR and NB models to perform sentiment analysis based on Spanish language. Antagonistically, these evaluations are focused on traditional methods for sentiment analysis tasks with NLP. In this sense, traditional methodologies show a good level of accuracy when performing sentimental analysis.

However, methods based on Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Memory Networks (MemNNN), Long Short-Term Memory (LSTM) and Recursive Neural Network (RecNN) it present a more robust and efficient result in the area of Natural Language Processing (NLP) [19], that represent an area of study for cyberbullying cases in Spanish language.

For this reason, in recent years, neural networks are dedicated for sentimental analysis tasks in SSNs; becoming a focus of research interest, offering models capable of classifying various sentimental topics from texts of limited length [20]. At the same time, these works are focused on empirical experiments whose adjustment seeks to avoid overfitting, to allow the implementation of the models in SSNs.

One such model is found within [21], where a sentimental analysis based on a convolutional recursive neural network (RvNN) coupled to pretrained word vectors in the upper layer for English language is proposed. Thus, it uses the RvNN as an alternative to the CNN clustering layer, thus reducing the loss of local information and allowing the model to efficiently capture long-term dependencies based on its tree structure. Consequently, a classification accuracy of 89.1%, as it extracts sentimental representations at word level considering syntactic and semantic relations, similar to what was achieved with techniques such as Latent semantic analysis (LSA) proposed in [22], [23].

On the other hand, [24] proposes a CNN-based text classification model for a dataset of verbal aggression in English language. The experiment combines the LSTM and a CNN to a 2-dimensional embedding Term Frequency Inverse Document Frequency (TF-IDF) layer. The obtained result reaches 91—% accuracy for the LSTM+2D TF-IDF and 92% accuracy for the CNN+2D TF-IDF model. Thus, it increases the accuracy achieved by the LSTM with 72% and CNN with 83% when using a preprocessed embedding layer such as Word2Vec [25].

In 2018, [26] proposes a pre-training method based the deep bidirectionality of BERT. Thus, his model is trained using the ”masked language model” (MLM) but without the ”next sentence prediction” (NSP) task and a left context model that is trained using a standard left-to-right LM (LTR), instead of an MLM. Then, it manages to improve the accuracy of traditional BERT by 2.5%. In this context, [27] proposes a model that uses a single linear neural network layer for classification. In its experimentation, it achieves an accuracy of 81% when using the Wikipedia Corpus.

Also, [20] conducts an evaluation of BERT for the detection of cyberbullying on social media: Twitter, Wikipedia, and FormSpring. The purpose of their evaluation is to refine the parameters of BERT when using the [26] Corpus to achieve better results. Therefore, in evaluating the model, it achieves an accuracy of 90% when using the Wikipedia Corpus. However, its best performance was achieved with a Twitter Corpus with which it obtains 94% detection accuracy. While, an RNN+LSTM model evaluated with the same Corpus only achieves 85% accuracy.

As a Spanish language case, we found that [17] performs the detection of violence against women in SSNs, based on Opinion mining techniques, Document Term Matrix (DTM), such as *bolsa de palabras* (bag of words), together with NB, Decision Trees (DT) and SVM algorithms for the social network Twitter. It´s most relevant result shows an accuracy of 90.35% when using the DT algorithm.

Thus, [18] presents a semi-automatic and presumptive detection system of cyberbullying; based on NB, SVM and LG techniques for the social network Twitter in Spanish language, called Spanish Cyberbullying Prevention (SPC). Thus, [18] uses sentiment analysis, machine learning and NLP to analyze phrases in Spanish from a group of trends and specific users from Ecuador, in order to identify the existence of a semantic set related to cyberbullying with up to 91% of precision, when using SVM.

In this context, we continue the work presented in [18], where a comparison is made between two models of deep learning; the first one based on RNN+LSTM and the second one on BERT for the presumptive detection of cyberbullying in Spanish language. Our goal is to obtain accurate and efficient models in terms of execution time, using bidirectional models with memory capabilities to train the models with a Spanish language Corpus.

The rest of the document is presented as follows: Section III describes the methodology used for the proposed method. Section IV analyze the results obtained. Finally, section V presents the conclusions of this paper.

III. Methodology

After reviewing several proposals of deep learning works for sentimental analysis focused on cyberbullying, we propose the implementation represented in Fig 1, composed of 3 stages: Dataset extraction, deep learning training and testing the trained models.

A. Dataset Extraction

Prior to training deep learning models, it´s important establish a dataset according to the subject matter. Therefore, it is necessary to generate a customized Corpus that fits the needs of the research. For this purpose, Twitter´s API and its advantages were used to form a Corpus through the extraction, processing and analysis of tweets through several implemented scripts; which were subsequently used for the proper training of the proposed models. Scripts with NLP algorithms are used to semi-automatically identify the sentiments of each tweet since its content is related to cyberbullying situations and a correct labeling of the data depends on it.

Main Corpus elaboration starts establishing a set of phrases and keywords that suggest some kind of insult or harassment; as well as others that show a more pleasant type of feeling, as shown in Table I. Thus, a set of weightings is established to determine the choice of an appropriate label for a comment. The label can be between 0 and 1, where the first value indicates that a comment is negative or has harassment and the second one when it is a positive or clean one. In turn, the Corpus used here has been implemented and developed by [18].

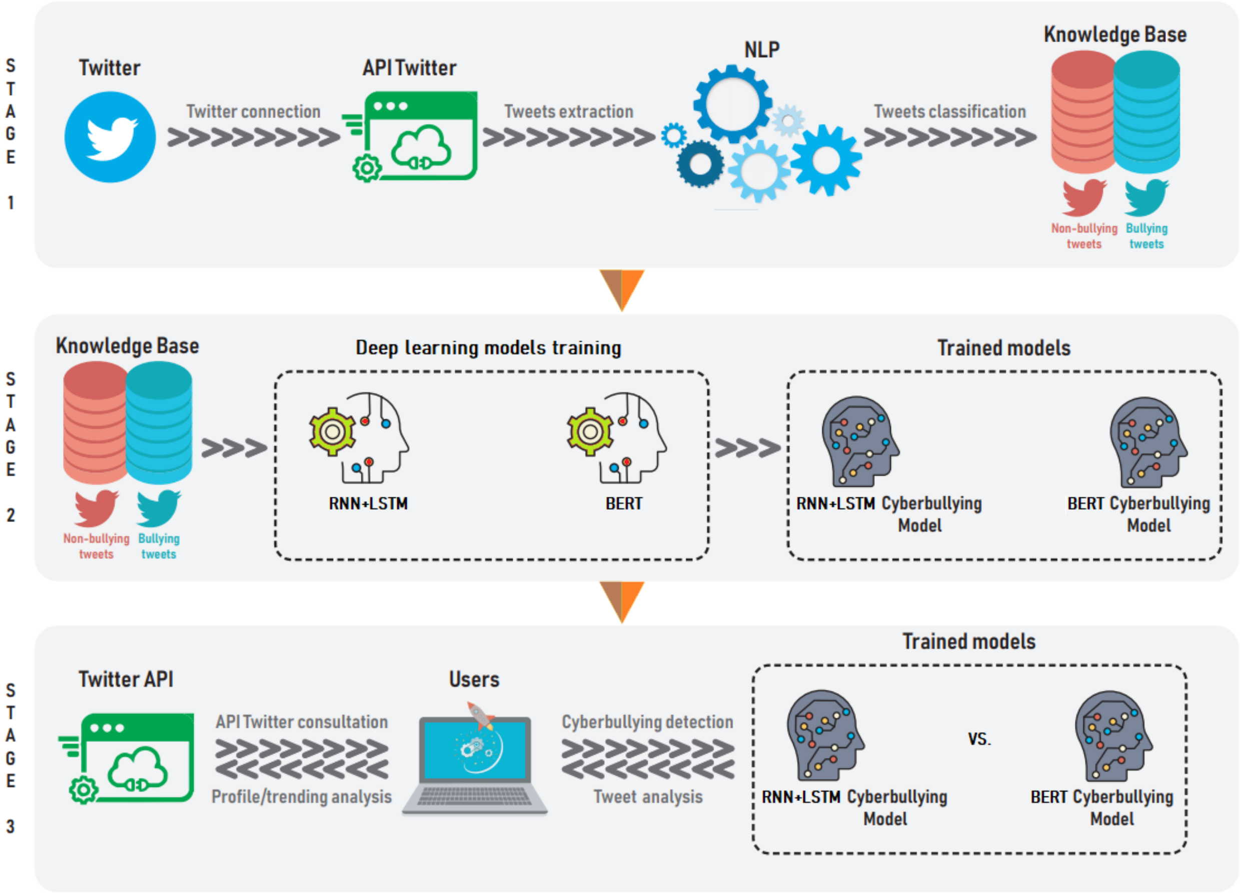


Fig. 1. Implementation Stages of Deep Learning Models.

TABLE I. Sample Corresponding to 6.25% of Phrases used for Tweets´s Automatic Classification According to a Positive or Negative Intent Focus

Frases Positivas	Frases Negativas
amar	desgraciado
eres increíble	gay
genial	hija de put*
gracias	indio
muy amable	vales verg*

However, to establish a margin of error, we have developed a second manually classified corpus. This is how a comparison is established between the effectiveness of an automatic classification method and one with personal criteria. The second Corpus represents 1.36% of information respect to automated version Corpus. The result is a difference of 20% less in the interpretation of the data considered as ”harassment” by the automatic model compared to the manual model. Subsequently, a data cleaning process was applied to remove mentions, emoticons and retweets because they affect the content of the comments and, consequently, could generate false negatives or positives in the system. As a result, a Corpus was formed with 881503 tweets, where 471432 of them correspond to comments with some type of harassment.

It should be noted that, as described in [18], a limitation of the Corpus is the impossibility of analyzing irony, since both verbs and contextual structures are not taken into consideration.

B. Deep Learning Training

To detect comments with presumed cyberbullying, our designed system uses a training corpus composed of 53.48% of comments or phrases with bullying. From this, the use of two models of deep learning is considered to detect and classify negative sentiments in Spanish language.

The proposed system has been developed from two methods: RNN + LSTM and BERT, because they are currently the deep learning models with the highest level of accuracy and development in prediction and classification tasks in the area of sentiment analysis.

1) RNN+LSTM Model: Sequence classification is defined as a predictive modeling problem where we have an input sequence over space or time and our goal is to predict a category where to assign that sequence [28]. Hence, a classification model development starts with the implementation of a block that allows text preprocessing, so that a deep cleaning is applied to the content of the original comments.

The stipulated cleaning includes special characters, punctuation symbols and repeated words elimination.

Also, it is necessary to apply *”Stemming”* or *”Lemmatization”* techniques that allow delete word endings to obtain their base structure, all that to avoid an excessive amount of identifiers, which may result in an inefficient classification system.

On the other hand, word representations are a fundamental component of many NLP systems, so it is common to represent words as indexes in a vocabulary [29]. However, this fails to capture the corresponding lexical structure. Hence, the importance of text preprocessing and the use of techniques that allow a mathematical fitting of a one-dimensional word space to a continuous vector space with fewer dimensions.

For the case of RNN, there is *Word Embedding*, known as a set of techniques for language modeling and NLP learning techniques. Its use represents an efficient way to clean, coding and vectorize words.

Thereby, the technique allows us know the spatial position of a word when it´s part of a vector, its characteristics and surrounding words. Here, a Word Embedding technique *"Word2Vec"* was proposed by [25] to delimit the number of most common words in a dataset, as well as their length provided by this technique. For this reason, with a greater features number defined, a better level of interpretation will be reached during training stage. Also, it should be considered that the system will take as reference only the words defined as the most common ones, since the rest of the words are labeled with a 0. This process is carried out automatically during the Tokenization process, where the larger text strings are divided into smaller parts or Tokens; in this way the larger text samples can be converted into sentences and these in turn can be tokenized into words. From Tokenization, it is possible to analyze the number of times each word appears in the dataset, as shown in Table 2.

TABLE II. Dataset´s Most used Words Sample

Top	Word	Top	Word
1	hac	**11**	grac
2	buen	**12**	amar
3	quer	**13**	sol
4	dec	**14**	mejor
5	put	**15**	dia
6	ir	**16**	pas
7	pod	**17**	verg
8	amig	**18**	tan
9	ser	**19**	gust
10	ver	**20**	hol

Based on the most used words, a dictionary was defined to match the variety of words in the comments with the tokenization of the vocabulary. After that, the comments were transformed into integer sequences. These sequences are truncated or filled in so that they all have the same length, prior to assigning the data to Train and Test blocks. Subsequently, the Train and Test blocks were defined in portions of 25% and 75% respectively. After padding the sequences, two matrices were obtained; the first one corresponding to X-Train of (220375x500) and the second one assigned to X-Test of (661126x 500), where 500 corresponds to the maximum length sequences. On the other hand, the labels in Y-Train and Y-Test are directly extracted from the original data set.

On the other hand, our RNN+LSTM model main feature is to use of a Bidirectional LSTM layer. Its use implies a remarkable improvement compared to a standard LSTM layer. A single LSTM layer is only capable of retaining information from a past state because it only considers one input related to past state. In contrast, a bidirectional LSTM layer can handle past and future states combined and thus give a better understanding context to the model to be trained.

Our RNN+LSTM model is composed as follows: an input layer, followed by an Embedding layer, commonly used in models with text. Next, a Batch normalization layer is placed, to adjust the dimensions of the data in the Embedding layer to the rest of the layers. Subsequently, a Bidirectional layer is applied, to provides our RNN model with a two-way LSTM memory; this adds the BERT´s main feature to the RNN+LSTM, bidirectionality.

2) BERT Model: Preprocessing of the dataset for the BERT model was performed using*stopwords* as a technique for removing special characters and punctuation symbols not relevant to the meaning of a sentence. [30]. Subsequently, as in RNN+LSTM, a tokenization process was applied. However, BERT has it´s own library to perform preprocessing and tokenization tasks called *BERT Tokenizer*. The library, by itself, splits text into tokens and at the same time converts those tokens into tokenizing vocabulary indexes. In addition, it allows to limit or equalize the sentences to equal length and to create an attention mask; where the last parameter is a necessary component for the training of the model. Through this, sentences were constrained into vectors with a maximum length of 64 identifiers of tokens each. Thus, as in the RNN+LSTM model, the dataset was divided into Train and Test blocks in portions of 25% and 75%, respectively.

By using a pre-trained model, BERT presents an efficient prediction accuracy when compared to other methods of deep learning. However, it´s possible increase the model´s accuracy level through Fine-Tunning techniques. This thanks to the *Transformers by Hugging Face* option in the *PyTorch* class. This do a settings, tokenization and architecture optimizing since base model; all thanks because Transformers provide a general-purpose architectures for natural language understanding with about 32 pre-trained models in more than 100 languages.

In addition, BERT is a robust model that consumes a large amount of system resources on which it will be run. One way to optimize its performance during a training and save memory is the use *PyTorch DataLoader* class. On the other hand, *PyTorch*, is a deep learning project development library whose main feature is the use of GPU acceleration. Thanks to this, *PyTorch* allows a dynamic behavior change of a neural network on the go. Finally, for our BERT model, the *BERT base* configuration was chosen for its trade-off between number of parameters vs. execution time. The base BERT model differs from the *BERT large* model by 2.8% and its runtime increases substantially; for this reason it will not be considered in this study.

Base BERT is configured by 768 hidden layers and 12 headers. However, it´s possible set a number of improvements to the model through AdamW optimizers, including a learning rate between $5e^{-5}$, $3e^{-5}$, $2e^{-5}$.

TABLE III. SEVERAL STUDIE-CASES HARASSMENT AND NON-HARASSMENT COMMENTS EXAMPLES

Harassment free Tweets	Harassment Tweets
si te toca estar en mesa. si sólo tienes que ir a votar y desaparecer.	Todos se cagaron, porque empezaron a decir, ”ahí viene la mujer de ese man”, nos escuchó gritar ”cachudo”,
Que viva Guayaquil de mis amores!	jajajaja cerraran las bien puertas y ventanas que pasa el lelo y su mafia.

C. Testing the Trained Models

Once the models are trained, we use different datasets to establish two case studies with which to corroborate their classification accuracy. For this paper, we used two accounts belonging to Ecuadorian users whose contents have different focuses: the first account specializes in hurtful comments (Study Case 1) and the other is focused on political issues (Study Case 2).

Based on this, the operating characteristics are evaluated through accuracy parameters of the trained models.

Thereby, an analysis is made of the phrases or words that, according to the models, suggest some type of harassment or insult.

In addition, the models make it possible determine the percentage of harassment on each analyzed account has in relation to extracted information.

Table III shows one example of a comment with harassment and one without harassment, from which the Study Cases Corpus are composed.

IV. ANALYSIS OF RESULTS

To validate our proposed models, we used the Table IV parameters for RNN+LSTM model and Table V parameters for BERT model.

Both models were trained and evaluated on a computer with 12 processing threads at 3.9GHz coupled with 48GB of RAM. In addition, since the models allow the use of vector graphics processing for their execution, a GPU with 1920 CUDA cores and 6GB of VRAM was used.

TABLE IV. RNN+LSTM MODEL STRUCTURING

Settings	RNN+LSTM
Epochs	6
Embedding neurons	128
Bidirectional Lstm neurons	128, with “batch normalization”
Dropout	128
Batch size	32
Dense layer	1, with “Sigmoid” activation

A. Operating Characteristics Analysis

Fig. 2 shows the curves with operating characteristics (ROC) of RNN+LSTM and BERT models.

TABLE V. BERT MODEL STRUCTURING

Settings	BERT
Epochs	4
Hidden layers	768
Heads	12
Batch size	32
Parameters	768
Learning rate	$5e^{-5}$
Epsilon value	$1e^{-8}$

The curve corresponding to RNN+LSTM model, in blue, shows an area under the curve (AUC) with a value of 0.97. On the other hand, BERT´s curve model, in red, shows an AUC with a value of 0.98.

Thereby, AUC is the probability that a negative random comment is considered as negative and a positive one as positive. Therefore, the value of the AUC ranges from 0 to 1.

Our implemented models provide optimal performance overall, where BERT model outperforms the RNN+LSTM model by 1%.

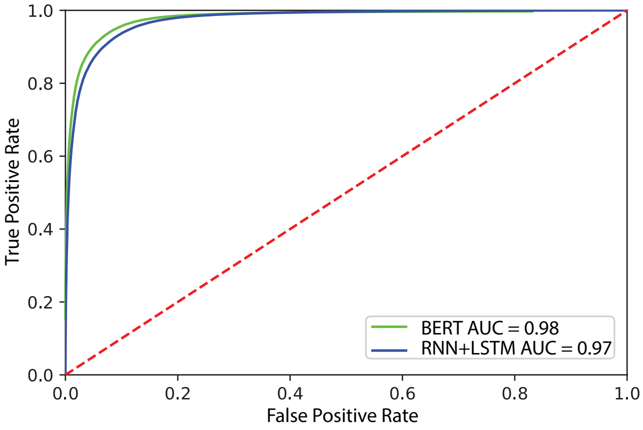


Fig. 2. RNN+LSTM & BERT Model´s Operating Characteristics.

B. Performance Analysis

As shown in Table VI, the BERT model reflects the highest accuracy with respect to its counterpart with RNN+LSTM and even surpasses by far the result obtained with SVM. In [18]. However, the RNN+LSTM model offers very close performance to that of BERT, with only a fraction of its run time.

TABLE VI. RESULTS FROM TRAINED MODELS

Parameter	RNN+LSTM	BERT	SVM	NB	LR
Accuracy [%]	91.82	92.82	87	83	85
Execution time [min]	78	270	-	-	-
AUC	0.97	0.98	-	-	-
Number of Tweets	220000	220000	230000	230000	230000

In terms of Epoch, RNN+LSTM model shows a decreasing trend with respect to the precision value, as shown in Fig. 3. Thus, the higher the number of Epochs the accuracy tends to

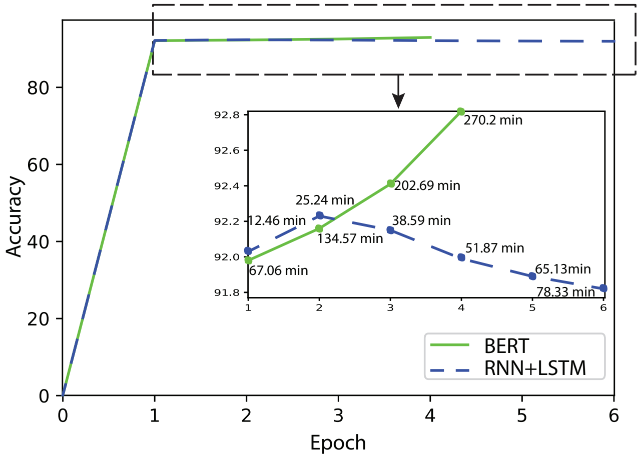


Fig. 3. Execution Time and Accuracy by Epoch.

stabilize at a lower value with respect to its starting point; being the number of Epochs chosen the appropriate one to achieve a stable value close to the one achieved by the BERT model.

On the other hand, BERT model shows an increasing behavior; suggesting that the model can achieve even higher levels of accuracy. However, a higher level of accuracy may represent a model´s overfitting.

At the same time, Fig. 3 evidences a convergence point with an accuracy´s approximate value between the two models at Epoch number 2.

However, this sample shows the difference in the execution time required by each model to achieve this accuracy value; where RNN+LSTM uses only 18% of time used by BERT.

C. Validation Analysis of Acertation

In order to evaluate the trained models, we have used data from two Twitter accounts whose users will be described as @USER_1 (Study Case 1) and @USER_2 (Study Case 2); in order to maintain these accounts anonymous. In turn, the accounts have different content approaches. The analysis are shown in Table VII. At the same time, we use the portion of the dataset intended for Test tasks.

TABLE VII. STUDY CASES ACCOUNTS CYBERBULLYING AVERAGE

Account	Account´s Harassment percentage [%]
@USER_1	86,3
@USER_2	19,27

The model´s results using non-explicitly offensive comments were satisfactory, especially with BERT model. With the dataset test portion, we generate a list of 20 random comments with alleged cyberbullying, BERT was able to identify 85% of them correctly. While, under the same parameters, RNN+LSTM identified 75% of comments correctly.

TABLE VIII. TWEET´S FRAGMENTS WITH ALLEGED HARASSING CONTENT ACCORDING TO THE RNN+LSTM MODEL WITH THE STUDY CASE 1 DATASET

Commentary	Harassment	Real State
Punkeros marihuanos fracasados buenos para nada, y el otro defecto era que cuando ella vivía en Quito se metía hierba, mucha hierba...	✓	✓
A ver te explico. A ver te explico. A ver te explico. A ver te explico. Mira mamaverg*...	✓	✓
Poco a poco esas experiencias de hablar por horas se reducían a minutos, nos volvíamos extraños, como si nunca nos hubiéramos conocido...	✗	✗
Al rato le llega una llamada de cierto hijo de put* preguntándole qué hace y dice: ”nada aquí en el mall, cae”. O SEA BRAAADER, O SEA	✓	✓
Nos fuimos todos a casa y al rato me llama Kar– diciendo que la hermana le dijo que yo le gustaba a ella y se puso a llorar...	✗	✗
’Belleza mis negros de la mini tri, carajo. Put*s para todos!, yo invito.	✓	✗

TABLE IX. TWEET´S FRAGMENTS WITH ALLEGED HARASSING CONTENT ACCORDING TO THE BERT MODEL WITH THE STUDY CASE 1 DATASET

Commentary	Harassment	Real State
Hasta Cuando Nos Vengan a Ver- Peker la Maravilla Ft El Soldadito Broder, esa vaina te llega al soul	✗	✗
Yo andaba en bóxer y la gata hija de put* que estaba tranquila acurrucada, de la nada me mete sus garras en mi virilidad masculina varonil. . .	✓	✗
Admito a veces se me chispotea y escribo algo mal, pero braaaders, hay gente que no tiene esta condición, y escriben con el cul*. . .	✓	✓
Todo el mundo me quedó viendo con cara de: ”Vales verg* pendejo, la hiciste llorar, no se hace llorar a una mujer en público”...	✓	✓
Toda grilla es madrina de un equipo pelotero de la universidad o empresa, y las grillas de más caché son candidatas a reinas de lo que sea. . .	✓	✓
Nos llega la noticia sobre la universidad, a ella le dijeron que estudiaría en Cuenca y justo yo estaba tratando de estudiar otra carrera...	✗	✗

In Study Case 1, 20 random comments are evaluated by RNN+LSTM and BERT they were able to predict 85% of comments correctly; part of these results are shown in Table VIII and Table IX respectively.

However, for Study case 2, BERT demonstrated a higher level of prediction than RNN+LSTM. When generating a list of 20 comments with suspected cyberbullying, BERT got 30% of them right, as seen in Table XI; while RNN+LSTM barely managed to identify 10%, from Table X.

This is due to the comments in Study case 2, which have no insults or hurtful comments as such, unlike Study Case 1. Which don´t contain any insults or hurtful comments as such, unlike Study Case 1. This is where the advantages of using a pre-trained model over a relatively inexperienced one become evident.

In addition, upon visual inspection of the classification results, it can be determined that BERT suggests a better approach to harassment, since it does not take any sentence as harassment even if it shows a potential insult. Likewise, this phenomenon can be better appreciated in the comments of Study Case 2, where RNN+LSTM model generates more false-positive cases.

On the other hand, results show a 16% of comments whose terms suggest some kind of insult without actually being harassment.

This is where implemented models tend to show bad results with respect to a real state of interpretation; taking into account that it can be a subjective parameter depending on the user who interprets it.

TABLE X. Tweet´s Fragments with Alleged Harassing Content According to the RNN+LSTM Model with the Study case 2 Dataset

Commentary	Harassment	Real State
Ministra R— debe estar presa: Es una vulgar delincuente. Terremoto político: Difunden organigrama, elaborado por Dan— M—	✓	✓
Solo por tratarse de Cor— violan las reglas básicas del derecho procesal. Que vergenza!	✓	✓
Los torcidos renglones de este panfleto. Cuánta imaginación malsana!!.Que poca capacidad de análisis político!	✓	✓
PRONUNCIAMIENTO DE LOS EX MIEMBROS DE LA COMISIÓN DE AUDITORÍA DE LA DEUDA Cuestionan la renegociación.	✗	✗
And— Ar— aceptó la candidatura del progresismo: Representamos a todos los sectores del Ecuador.	✗	✗
Caso BOCHORNOS es la suma d todas las arbitrariedades	✓	✓

TABLE XI. Tweet´s Fragments with Alleged Harassing Content According to the BERT Model with the Study Case 2 Dataset

Commentary	Harassment	Real State
La situación en Pichincha es grave. Los esfuerzos para evitar contagios necesitan la responsabilidad de todos los ciudadan	✗	✗
Ningún Asambleísta de está envuelto en sucios y corruptos repartos de hospitales. Eso no sale en la prensa...	✓	✓
Empezaron los recaderos de banqueros! Nos enseña algún sobreprecio? AHORA es que los hay. Ha de ser feo ver a...	✓	✓
Lo mismo se pasaron diciendo los 10 años de mi Gobierno. Recuerdan? La campaña sucia de siempre.	✓	✓
Cuando de pronto empezó a suceder algo raro...	✗	✗
Esta designación lejos de ser únicamente mérito indígena...	✓	✗

V. Conclusions

In this paper, we continue the work presented in [18], evaluating two models of deep learning in NLP tasks for the detection of cyberbullying in Twitter social network. The proposed models achieve at least a 5% improvement in accuracy over the best model of the previous work corresponding to SVM.

Our first model, RNN+LSTM, shows as the most balanced option between execution time with 78 [min] and an accuracy of 91.82%; which means a difference of 1% compared to 92.82% of BERT´s accuracy, but using 82% more run time.

However, RNN+LSTM doesn´t represent the most efficient cyberbulling classification option, due to BERT has a better criterion when detecting harassment, just like Study case 2, where BERT outperforms RNN+LSTM by 20%. Thanks to the fact that BERT is a pre-trained model and that the analyzed account does not contain explicitly offensive comments, which presents a challenge for both models in predicting harassment.

At the same time, RNN+LSTM model requires 33.33% more Epochs to approach the accuracy value of the BERT model. Nevertheless, RNN+LSTM is shown to be a robust option for presumptive detection tasks of cyberbullying in SSNs. For this reason, it is recommended that the model be redesigned to include a more robust Bidirectional layer with a larger number of neurons.

On the other hand, BERT model offers different model architectures with a greater or lesser number of parameters, depending on the desired implementation approach. Thus, there is potential improvement for accuracy if we evaluate the *bert-large-uncased* or *bert-base-multilingual-uncased* models. However, a different BERT model requires higher computational capabilities, which can be limiting factor for its implementation.

Finally, through the case studies, it was determined that the trained models are robust enough to predict whether an account includes cyberbullying comments in its content and at the same time its percentages.

References

[1] E. Hafermalz and K. Riemer, “The question of materiality: Mattering in the network society,” 2015.

[2] R. Ortega Ruiz, R. d. Rey Alamillo, and J. A. Casas Bolaños, “Redes sociales y cyberbullying: El proyecto conred,” *Convives, 3, 34-44.*, 2013.

[3] G. A. Maldonado Berea, J. García González, and B. E. Sampedro Requena, “El efecto de las tic y redes sociales en estudiantes universitarios,” *RIED Rev. Iberoam. Educ. Distancia*, vol. 22, pp. 153–176, 2019.

[4] L. E. Arab and G. A. Díaz, “Impacto de las redes sociales e internet en la adolescencia: aspectos positivos y negativos,” *Revista Médica Clínica Las Condes*, vol. 26, no. 1, pp. 7–13, 2015.

[5] M. A. S. Ruiz and C. Inostroza, “Repercusiones sobre la salud del maltrato entre iguales: coso escolar y ciberacoso,” *Revista de estudios de juventud*, no. 115, pp. 195–206, 2017.

[6] L. P. Bosque Vega, “Detección automática de ciber acoso en redes sociales.” Ph.D. dissertation, Universidad Autónoma de Nuevo León, 2017.

[7] UNICEF *et al.*, “Violencia entre pares en el sistema educativo: Una mirada en profundidad al acoso escolar en el ecuador. unicef. 2017,” *Revista virtual][Fecha de acceso: 08 de Agosto del 2019]. En: https://www. unicef. org/ecuador/Press_Kit__AbusoEscolar_Final. pdf.*

[8] Cnna-Cnii, *Agenda Nacional para la Igualdad Intergeneracional 2013-2017 del Ministerio de Inclusión Económica y Social de Ecuador.*, 2014. [Online]. Available: https://www.ohchr.org/Documents/Issues/OlderPersons/MIPAA/Ecuador_Annex1.pdf

[9] L. C. Díaz, “# nobullying: una acción integral contra el acoso escolar,” *Revista de Estudios de Juventud*, no. 115, pp. 167–191, 2017.

[10] A. Allisiardi *et al.*, “Bullyng y ciberbullying en argentina: el rol de la comunicación en su prevención. guía con orientaciones para campañas de publicidad social.” 2019.

[11] J. Díaz Ospina *et al.*, “Diseño de contenidos digitales para campañas publicitarias de bien social de sensibilización del ciberacoso o cyberbullying en niños de 8 a 12 años estratos 4-6 de manizales,” Master’s thesis, Escuela de Ciencias Sociales, 2018.

[12] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowledge and Information Systems*, pp. 1–51, 2020.

[13] N. Joselson and R. Hallén, “Emotion classification with natural language processing (comparing bert and bi-directional lstm models for use with twitter conversations),” 2019.

[14] M. Andriansyah, A. Akbar, A. Ahwan, N. A. Gilani, A. R. Nugraha, R. N. Sari, and R. Senjaya, “Cyberbullying comment classification on indonesian selebgram using support vector machine method,” in *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017, pp. 1–5.

[15] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, “Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges,” *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[16] S. T. T and B. R. Jeetha, "Cyberbullying detection in twtter using language extraction based simplified support vector machine (ssvm) classifier," vol. 6, no. 3, 2017, pp. 21 – 30.

[17] G. A. Prieto Cruz and E. E. Montoya Vasquez, "Modelo de detección de violencia contra la mujer en redes sociales en español, utilizando opinion mining," 2020.

[18] G. A. León-Paredes, W. F. Palomeque-León, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, L. I. Barbosa-Santillán, and M. M. Paredes-Pinos, "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the spanish language," in *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 2019, pp. 1–7.

[19] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[20] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification," *Multimedia Systems*, pp. 1–8, 2020.

[21] H. Sadr, M. M. Pedram, and M. Teshnehlab, "A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks," *Neural Processing Letters*, vol. 50, no. 3, pp. 2745–2761, 2019.

[22] G. A. León-Paredes, L. I. Barbosa-Santillán, and J. J. Sánchez-Escobar, "A Heterogeneous System Based on Latent Semantic Analysis Using GPU and Multi-CPU," *Scientific Programming*, vol. 2017, p. 19, 2017.

[23] G. A. León-Paredes, L. I. Barbosa-Santillán, J. J. Sánchez-Escobar, and A. Pareja-Lora, "Ship-sibiscas: A first step towards the identification of potential maritime law infringements by means of lsa-based image," *Scientific Programming*, vol. 2019, 2019.

[24] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Computing and Applications*, pp. 1–10, 2018.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[27] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 1096–1100.

[28] A. Huertas Mora *et al.*, "Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas."

[29] M. Vallez and R. Pedraza, "El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines," *Hipertext. net*, 2007.

[30] H. Saif, M. Fernandez, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14)*, pp. 810–817, 01 2014.

Situation Awareness Levels to Evaluate the Usability of Augmented Feedback to Support Driving in an Unfamiliar Traffic Regulation

Hasan J. Alyamani[1], Ryan Alturki[2], Arda Yunianta[3], Nashwan A. Alromema[4],
Hasan Sagga[5], Manolya Kavakli[6]
Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia[1,3,4]
College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia[2]
College of Business, University of Jeddah, Asfan, Jeddah, Saudi Arabia[5]
Faculty of Science and Engineering, Macquarie University, Sydney, Australia[6]

***Abstract*—Driving in an unfamiliar traffic regulation using an unfamiliar vehicle configuration contributes to increase number of traffic accidents. In these circumstances, a driver needs to have what is referred to as 'situation awareness' (SA). SA is divided into (level 1) perception of environmental cues, (level 2) comprehension of the perceived cues in relation to the current situation and (level 3) projection of the status of the situation in the near future. On the other hand, augmented feedback (AF) is used to enhance the performance of a certain task. In Driving, AF can be provided to drivers via in-vehicle information systems. In this paper, we hypothesize that considering the SA levels when designing AF can reduce the driving errors and thus enhance road safety. To evaluate this hypothesis, we conducted a quantitative study to test the usability of a certain set of feedback and an empirical study using a driving simulator to test the effectiveness of that feedback in terms of improving driving performance, particularly at roundabouts and intersections in an unfamiliar traffic system. The results of the first study enhanced the ability of the in-vehicle information system to provide feedback considering SA levels. This information was incorporated into a driving simulator and provided to drivers. The results of the second study revealed that considering SA levels when designing augmented feedback significantly reduces the driving errors at roundabouts and intersections in an unfamiliar traffic regulation.**

***Keywords*—Situation awareness; unfamiliar traffic regulation; augmented feedback; in-vehicle information systems**

I. Introduction

International drivers, especially from an opposite, unfamiliar traffic regulation (UFTR), are involved in a large number of traffic accidents [1], [2]. In this paper, a UFTR represents driving a right-hand drive vehicle in a keep-left traffic regulation for drivers who are only familiar with driving a left-hand drive vehicle in a keep-right traffic regulation. The minority of countries (75 countries), such as Australia and Japan [3] are following a UFTR in their traffic system. In the case of UFTR, and especially for international drivers, driving can be more demanding, resulting in an increased number of vehicle accidents, especially at roundabouts and intersections [4].

Overall, satisfactory driving in these conditions requires drivers to be aware of the environment as well as the configurations of the vehicle. This is known as situational awareness (SA). In general, SA refers to the perception of surroundings with respect to time and space, the comprehension of their meaning, and the projection of their status in the future [5]. Accordingly, SA can be divided into three levels [5]: level 1 perception, level 2 comprehension and level 3 projection.

In-vehicle information systems aim to enhance the road safety by providing the drivers with some feedback that help them to make the proper reactions [6]. However, most of existing works ignore designing the provided feedback based on the SA levels, particularly for information required when driving at roundabouts and intersections in a UFTR. The aim of this paper is to consider the SA levels when designing the feedback provided to drivers by in-vehicle information systems. That is, the feedback is provided in perceptual, comprehensible and projectable manner. We hypothesize that considering the SA levels when designing the target feedback can reduce the driving errors and thus improve the road safety. To evaluate this hypothesis, we firstly conducted a quantitative study to test the usability of a certain set of feedback (i.e., travelling path, turning, signalling and speeding at roundabouts and intersections). The results of the first study enhanced the ability of the in-vehicle information system to provide feedback considering SA levels. Then, we conducted an empirical study using a driving simulator to test the effectiveness of that feedback in terms of improving driving performance, particularly at roundabouts and intersections in an unfamiliar traffic system.

The rest of the paper is organized as follows. Section 2 introduces the situation awareness and its levels in driving domain, the rules of driving at roundabouts and intersections in a UFTR and how these rules differ between familiar and unfamiliar traffic regulation, and some in-vehicle information systems that provide visual feedback at roundabouts and intersections. Section 3 tests the usability of presenting the visual feedback in accordance with SA levels. Section 4 evaluates the effectiveness of that feedback in terms of improving the road safety. Section 5 concludes the paper.

II. Background

A. Situation Awareness in Driving

SA in the driving refers to the perception of the elements in the traffic system within a volume of time and space, the comprehension of their meaning, and the projection of their

status in the near future [7]. Traffic system includes three main components: vehicles on the road, users of the road and road environment. Accordingly, the levels of SA in this domain are:
Level 1: Perception of the relevant elements in the traffic system.
Level 2: Comprehension of the meaning of the perceived elements.
Level 3: Projection of near-future status of traffic system.

In relation to driving in a UFTR, the study [8] discovered the drivers' SA when driving at a UFTR by testing the lane changing performance at straight and curved roads in addition to roundabouts and intersections. The findings of [8] showed that international drivers are likely to have low SA when driving at roundabouts and intersections. Accordingly, the current study focuses on driving at roundabouts and intersections in a UFTR.

B. Driving at Roundabouts and Intersections in a UFTR

Australia is a keep-left country whose traffic system follows a UFTR in this research. In case of no traffic presents in the dual-lane roads, Australian traffic rules are as follows [9]:
1) Approach a roundabout/intersection (a) using the correct approaching lane according to the target direction (i.e. the right, left, and right or left lane when turning right, turning left, and going straight-ahead, respectively), (b) signalling prior to turning left or right and (c) slowing down or completely stopping the vehicle.
2) Enter a roundabout/intersection in the correct direction of traffic flow.
3) Exit a roundabout/intersection using the correct lane.

International drivers who come from keep-right countries must consider the main differences in traffic rules and vehicle configuration when driving at a roundabout/intersection in their home countries and in a UFTR. The differences are as follows:

- In the right-hand drive vehicle (the driver is not familiar with) has its direction indicator stalk on the right of the steering wheel whereas in the left-hand drive vehicle (the driver is familiar with) has its direction indicator stalk on the left of the steering wheel.
- The correct direction when entering a roundabout in a UFTR is clock-wise, not as it is on the keep-right traffic system (i.e. counter clock-wise).
- The correct direction of the traffic flow in a UFTR is the opposite of traffic flow in a keep-right traffic system.

In general, drivers become more aware of the driving environment when they manage to identify the information relevant to the task in rapid patterns to prepare themselves to make appropriate decisions and reactions. This helps the driver understand the relationship between the driving goal and other components of the traffic system at any moment in time [10]. Hence the driver will make fewer errors and avoid vehicle accidents. The information the drivers need can be provided to the drivers by In-vehicle Information Systems (IVISs).

C. Augmented Feedback in Driving

Augmented feedback is a form of technology that enhances the performance of a certain task by integrating computer-generated objects with the virtual or real environment [11]. It can be described as a set of stimuli whereby learners receive information from external sources in order to reinforce their behaviour and learning process [12]. The augmented feedback usually targets a task the operator is performing or going to perform in the future. Augmented feedback can be provided in different modalities, such as video, auditory or haptic in order to enrich the user's experience. In general, augmented feedback shows promising results with applications in various human situation, including driving [13], to enhance a number of related physical tasks. In driving, augmented feedback can be provided to drivers via in-vehicle information systems (IVISs).

IVISs are a form of technology that can support drivers by providing real-time feedback returned to drivers telling them what is going. The driver perceives the feedback from the system and understands it to make a decision, and refine or correct his/her driving reactions [14]. That keeps the drivers in the loop and thereby improving the drivers' SA and safety [15]. However, there is growing concern that IVIS might distract drivers from the primary driving task and thus compromise driving safety, as looking at the provided feedback consumes the driver's visual demand. The drivers are required to move their eyes back and forth between the road and the system. Additionally, there is a potential safety issue if the system provides drivers with unnecessary information, as it reduces driving performance [16], [17].

To reach the potential goal of IVISs, the system should provide only the important information. The driver does not need to capture the information that is irrelevant to the current or upcoming driving task. Additionally, the system should provide perceivable information (represents level 1). That is, the driver should not have to take his or her eyes off the road while glancing at the provided information. Also, the system must provide quick comprehensible information [18] (represents level 2). Unclear information might confuse the driver and affect his or her ability to make the correct reaction within the appropriate time. Moreover, the system should provide the information in a projectable manner (represents level 3). That is, the information helps the driver make a correct decision in relation to the upcoming situation. Understanding the drivers' needs of provided information are necessary to design the most useful system.

In relation to driving in a UFTR, the study [8] extracted the required information to drive safe at roundabouts and intersections. The information included upcoming area of the road to inform the driver about the upcoming roundabout and intersection, travelling path to inform the driver about the correct travelling lane when approaching, entering and exiting a roundabout/intersection, direction indicator to remind the driver of the correct position of direction indicator stalk, and speed to inform the driver about the speed limit of the travelling road. Visual feedback is selected by overseas drivers as the most preferable feedback modality to present that information when driving in a UFTR [19]. Therefore, in this research, we only focus on visual feedback provided by IVISs.

D. Visual Feedback in Roundabouts/Intersections In-Vehicle Information Systems

Becic et al., [20] investigated the effectiveness of an IVIS in reducing the number of crashes at a rural stop-controlled intersection in a keep-right traffic regulation. The system displays the upcoming intersection as arrows cross each others on either right or left display based on the traffic flow in the upcoming intersection. The arrows show the traffic direction without showing the required travelling lane the driver should travel within. Also, the information regarding direction indicator and speed limit is not provided to the drivers.

Zhang et al., [21] designed an IVIS that assists the driver when approaching a cross, "arch-shaped" intersection in a keep-left traffic regulation. The system provides visual feedback in a head-up-display (HUP) on the central-lower part of the windshield. The system shows the traffic status of upcoming intersection, leaving the drivers to slow down or make another proper reaction.

Tran et al., [22] studied driving performance at signalised turn-left intersections in a keep-right traffic regulation. In such a driving scenario, the driver does not have priority to turn left. He/she must cross the traffic flow coming from the opposite direction. Using a 3D HUP, the IVIS system visualises a three second projected path of the vehicle coming from the opposite side of the road. The path is presented in the driver's main field of view. As long as no vehicle is approaching the intersection, the system presents nothing to the driver. In other words, the system does not describe the situation until traffic exists in the upcoming area.

Caird et al. [23] designed a sign IVIS that visually provides two road signs "prepare to stop" and "signals ahead" on a HUD. These two signs inform the driver of an upcoming change in the road, but they do not provide a complete picture of what is coming next, leaving drivers to guess about whether the traffic light might be at an intersection, roundabout, or T-junction.

Alyamani et al., 2019 [24] extended the feedback designed in [19] and provided it to the international drivers when driving at a UFTR (see Figure 1). The feedback is presented on a HUD. The feedback includes up-coming section of road, either intersection or roundabout, the correct travelling path, direction indicator and speed limit to help drivers make proper decision and reaction when driving at a roundabout/intersection.

Table I compares different roundabout/intersection in-vehicle information systems according to feedback the system provides. Two of the works reviewed above support the drivers when driving in a keep-left traffic regulation. All the works reviewed above provide the drivers with feedback related to the upcoming section of the road either prior approaching all upcoming roundabout/intersection or only when traffic exists in the upcoming area. Feedback regarding travelling path is only supported totally in [24] and partially in [22]. None of the above works take into account presenting the direction indicator and speed except [24]. Accordingly, this research focuses on evaluating the usability of the feedback designed by [24] as will be described in the following section.

III. Testing the Usability of Required Feedback

A short questionnaire was developed to collect overseas drivers' opinions about the design of the feedback proposed by [24]. We particularly focused on presenting the feedback in a perceivable (level 1), comprehensible (level 2) and projectable (level 3) way. The proposed location of the target feedback was in the lower-central part of the windshield, using a HUD (level 1), the content included well-known symbols, icons and road signs (level 2) to help the driver make decisions about the required driving performance and behaviour (level 3).

A. Questionnaire Design and Procedure

An open recruitment process was adopted. Participants were recruited via an email that included the survey link. At the beginning of the questionnaire, information and instructions regarding driving at roundabouts and intersections in NSW, Australia, were presented. The questionnaire started by collecting demographic information, such as gender, age and driving experience in both keep-right and keep-left traffic regulations. Then, a GIF image of a vehicle driving in a simulated keep-left road was presented to give participants a better sense of driving in a keep-left traffic regulation, especially for those who were only familiar with a keep-right traffic regulation. Then, the participants were asked to answer the usability questions. Table II shows the list of questions, the target feedback and the rating scale. Each question included an image (see Fig. 2); participants were asked to observe the image and answer the question. The participants could add any comments after answering the question. Questions QA, QBs and QCs addressed the usability of the information in relation to its perceivability (level 1), comprehensibility (level 2) and projectability (level 3), respectively.

B. Participants

We distributed the online surveys to 65 participants, aged 19-52 with mean of 31.7 (SD: 1.1). The participants were only familiar with a keep-right regulation. All had a driver's license issued in their home keep-right country with a mean driving experience of 13.5 years (SD = 1.3) and drove in familiar regulations for an average of 19.7 hours/week (SD = 1.9).

C. Data Analysis

The usability questions were grouped into two categorical variables. Responses of the first question were categorised as 'easy', (including 'easy' and 'very easy') and 'difficult' (including 'very difficult' and 'difficult'). Responses to the second question were categorised as 'clear' (which included 'very clear' and 'clear') and 'not clear' (including 'not very clear' and 'not clear at all'). Responses to the third question were grouped into 'useful', (including 'very useful' and 'useful') and 'not useful' (including 'not very useful' and 'not useful at all'). Undecided responses were excluded. As we had one sample and categorical variables, we conducted a non-parametric test (i.e. one sample chi-square goodness-of-fit test) [25] to determine usability in relation to each question.

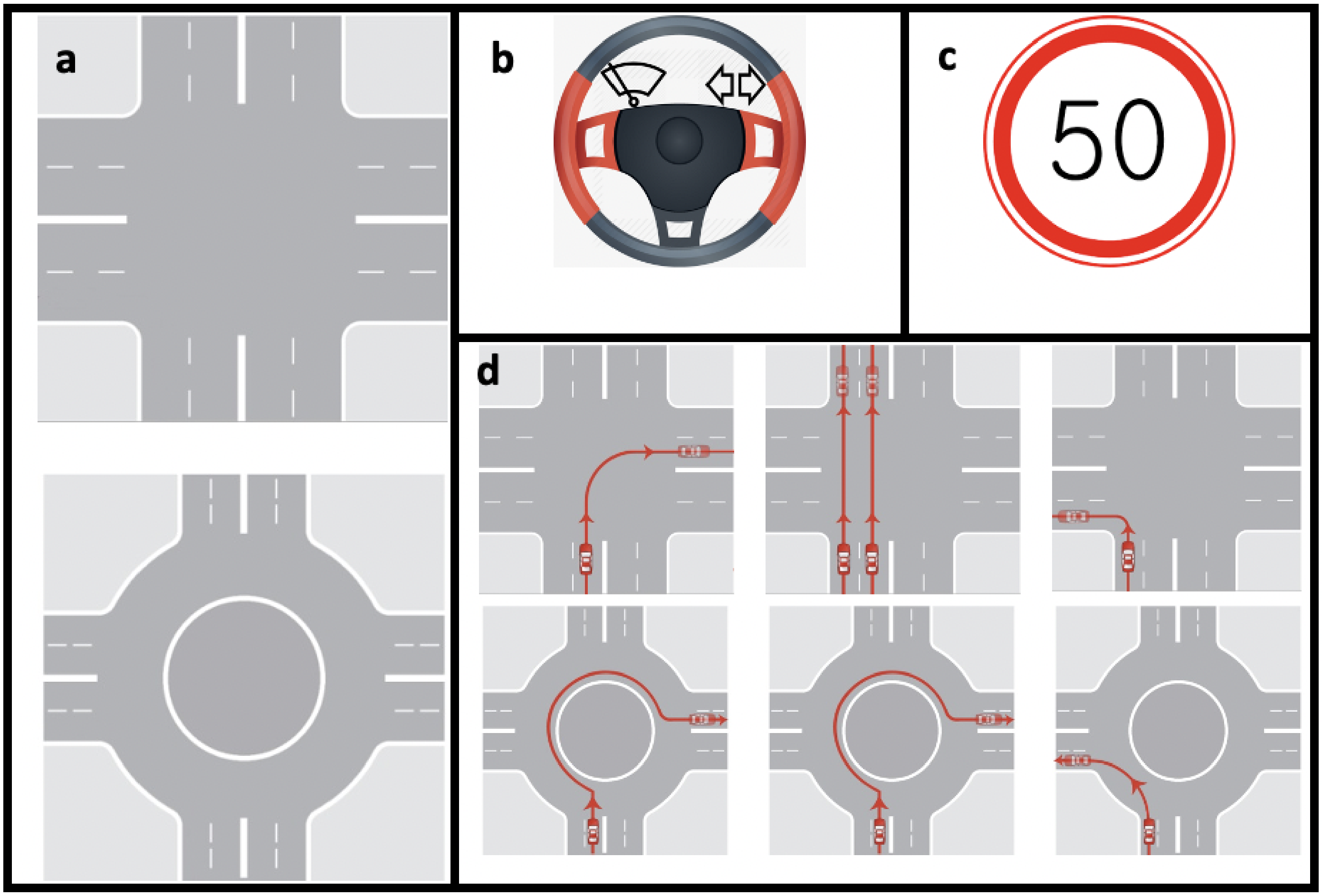


Fig. 1. Extended Design of Required Feedback to Support Overseas Drivers when Driving in Roundabouts and Intersections [24]. The Feedback Include a) the Upcoming Section of the Road, b) using the Direction Indicator,c) Speed Limit and d) Travelling Path).

TABLE I. OVERVIEW OF A ROUNDABOUT/INTERSECTION IVIS RELATED TO TRAFFIC REGULATION AND REQUIRED FEEDBACK. ●: SUPPORTED, ○: NOT SUPPORTED, [●] : PARTIALLY SUPPORTED

Authors	Traffic regulation	Upcoming section of road	Travelling	Using direction indicator	Speed limit
Becic et al. 2012 [20],	KRT	●	○	○	○
Zhang et al. 2009 [21],	KLT	●	○	○	○
Tran et al. 2013 [22]	KRT	[●]	[●]	○	○
Caird et al., 2008 [23]	KRT	●	○	○	○
Alyamani et al., 2019 [24]	KLT	●	●	●	●

D. Results

In relation to the first question, 44 participants thought feedback presented on the lower-central part of the windshield using a HUD was easy to perceive whereas 10 participants thought it was difficult (see Table III). The minimum expected frequency was 27.0. The difference was statistically significant $(\chi^2(1) = 21.407, \rho < .001)$.

The majority of participants (55) thought the upcoming area feedback clearly described the conditions of the upcoming area; only 2 participants did not think it was clearly described. The minimum expected frequency was 25.5. The difference was statistically significant, $\chi^2(1) = 49.281, \rho < .001$. On the other hand, 27 participants thought the upcoming area feedback could help them make a decision regarding the required reaction or performance at the upcoming roundabout or intersection, while 24 participants thought this feedback could not help them do so. The minimum expected frequency was 25.5. The difference was not statistical significant $(\chi^2(1) = .176, \rho = .674)$.

Most participants (56) thought that the combination of upcoming area and travelling path feedback clearly described the situation; only 3 participants found that this combination did not clearly describe the situation. The minimum expected frequency was 29.5. The difference was statistically significant$(\chi^2(1) = 47.610, \rho < .001)$. Similarly, 59 participants thought that presenting upcoming area and travelling path feedback was useful for planning the required reaction at the upcoming section of road; only 2 participants thought it was not useful. The minimum expected frequency was 30.5. The difference was statistically significant $(\chi^2(1) = 53.262, \rho <$

TABLE II. QUESTIONS TO EVALUATE THE USABILITY OF NECESSARY FEEDBACK

Q#	Question	Target feedback	Rating scale
QA	How easy is it to perceive the feedback in the image below?	All	'very difficult', 'difficult', 'normal', 'easy' and 'very easy'
QB1	How well does the feedback describe the situation shown in the clip above?	Upcoming area	'not clear at all', 'not very clear', 'normal', 'clear', 'very clear'
QB2	How well does the feedback describe the situation shown in the clip above?	Upcoming area + travelling path	
QB3	How well does the feedback describe the situation shown in the clip above?	Upcoming area + direction indicator	
QB4	How well does the feedback describe the situation shown in the clip above?	Upcoming area + speeding	
QC1	How useful is the feedback helping you make decisions about the required performance at the upcoming roundabout/intersection?	Upcoming area	'not useful at all', 'not very useful', 'normal', 'useful', 'very useful'
QC2	How useful is the feedback helping you make decisions about the required performance at the upcoming roundabout/intersection?	Upcoming area + travelling path	
QC3	How useful is the feedback helping you make decisions about the required performance at the upcoming roundabout/intersection?	Upcoming area + direction indicator	
QC4	How useful is the feedback helping you make decisions about the required performance at the upcoming roundabout/intersection?	Upcoming area + speeding	

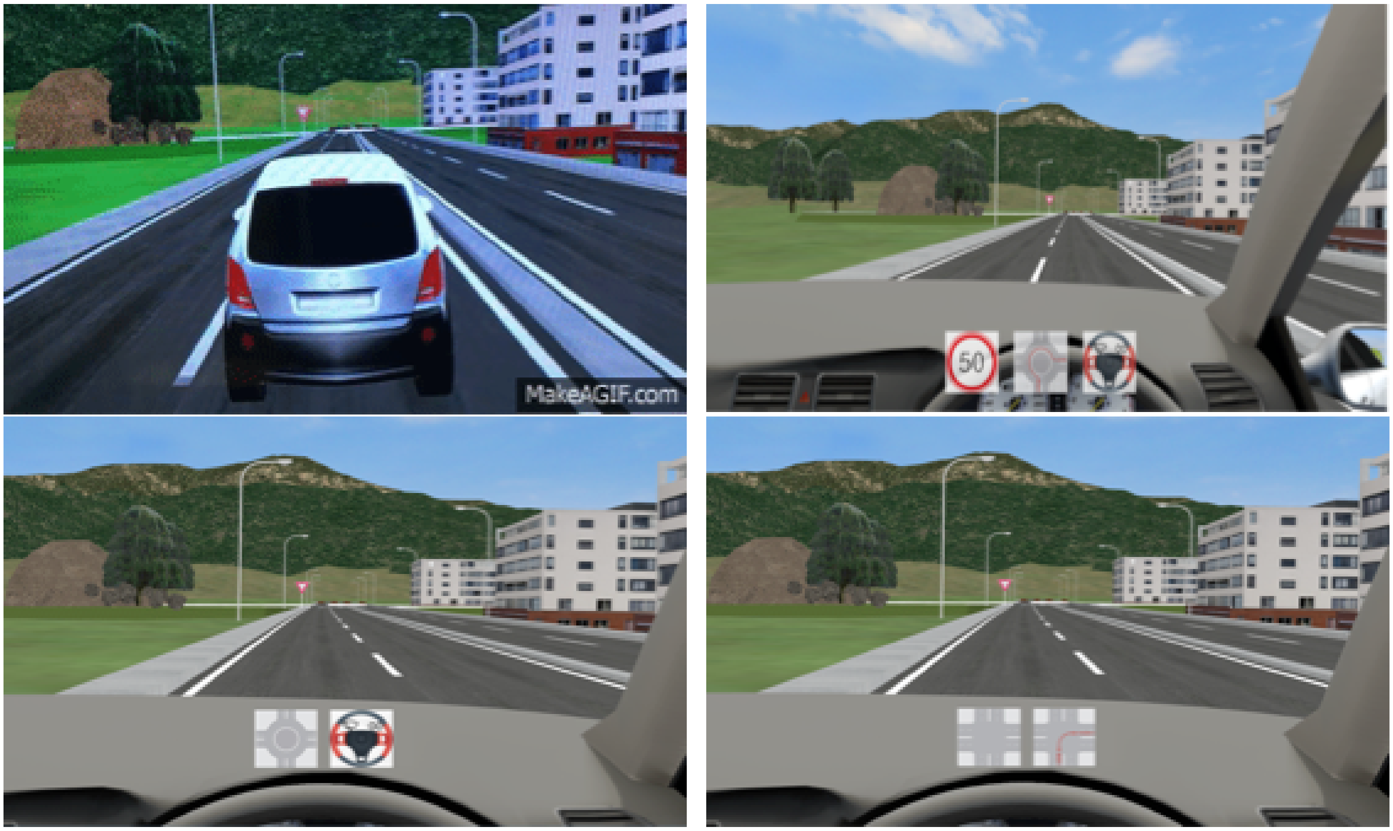


Fig. 2. Some Images used in the Usability Questionnaire.

TABLE III. POSITIVE AND NEGATIVE RATINGS FOR EACH QUESTION

Question	# of 'easy' responses	# of 'difficult' responses	$\chi^2(1)$	p
QA	44	10	21.407	$<$ .001*

Question	# of 'clear' responses	# of 'not clear' responses	$\chi^2(1)$	p
QB1	55	2	49.281	$<$ .001*
QB2	56	3	47.610	$<$ .001*
QB3	47	10	24.018	$<$ .001*
QB4	57	5	43.613	$<$ .001*

Question	# of 'useful' responses	# of 'not useful'	$\chi^2(1)$	p
QC1	27	24	.176	=.674
QC2	59	2	53.262	$<$ .001*
QC3	39	12	14.294	$<$ .001*
QC4	62	0	-	-

* Statistically significant.

.001).

In relation to upcoming area and direction indicator feedback, 47 participants thought this combination clearly described the situation, whereas 10 participants thought it did not. The minimum expected frequency was 28.5. The difference was statistically significant, $\chi^2(1) = 24.018, \rho < .001$. In addition 39 participants thought it was useful to plan the required reaction using both upcoming area and direction indicator feedback together, whereas 12 participants thought it was not useful to do so. The minimum expected frequency was 25.5. The difference was statistically significant ($\chi^2(1) = 14.294, \rho < .001$).

In relation to a combination of upcoming area and speeding feedback, 57 participants thought presenting upcoming area and speeding feedback together clearly described the situation, while 5 participants thought it did not. The minimum expected frequency was 31.0. The difference was statistically significant ($\chi^2(1) = 43.613, \rho < .001$). 62 participants thought it was useful to plan the required reaction using that combination of feedback; no participant thought it was not useful to do so. The chi-square goodness-of-fit test could not be performed as the variable was constant (there was no value for the 'not useful' group).

E. Discussion

A significant percentage of participants (67.7%) reported that feedback displayed on a HUD was easy to perceive. Therefore, we should present the target feedback on a HUD. A significant percentage of participants (84.6%) also thought that providing only upcoming area feedback (i.e. static feedback) clearly described the upcoming situation. However, there was insignificant difference between the percentage of participants (41.5%) who reported that upcoming area feedback was useful for directing the driver to make the proper decisions at the upcoming section of road and those who did not (36.9%).

Participants were confused about the required reaction due to their unfamiliarity with the keep-left traffic regulation, even the driving rules at roundabouts and intersections were explained at the beginning of the questionnaire. On the other hand, a significant percentage of participants thought that combining upcoming area information with any of the other forms of dynamic feedback (i.e. travelling path, direction indicator and speed) clearly described the situation and was useful for making appropriate decisions at an upcoming section of road. In fact, 86.2%, 72.3% and 87.7% of participants found that the combination of upcoming area and travelling path, upcoming area and direction indicator, upcoming area and speed, respectively, clearly described the situation of the upcoming section of road. Also, 90.8%, 60.0% and 95.4% of participants found the combination of upcoming area and travelling path, upcoming area and direction indicator, upcoming area and speed, respectively, was useful for making appropriate decisions at an upcoming section of road. Hence, we recommend providing static feedback supported by dynamic feedback. The static feedback functioned to describe provided the upcoming situation while dynamic feedback supported decision making.

In relation to a combination of upcoming area and travelling path feedback, four participants (6.2%) who responded that this combination was clear and useful recommended that we use only travelling path instead, as travelling path was shown on the roundabout and intersection. Two participants (3.1%) who were undecided about the clarity and usefulness of this combination found the simultaneous presentation confusing. One participant (1.5%) who had a negative response regarding the clarity and usefulness of that combination commented: "I do not know if I will drive through one or two roundabouts" and "should I follow the line at first and second roundabout?" As a result of the wider potential for such confusion, we would follow the recommendations we received from participants and only provide travelling path feedback when it was necessary.

IV. EFFECTIVENESS EVALUATION

The study evaluates the effectiveness of the feedback mentioned above for improving the road safety, particularly for drivers who are not familiar with driving in a UFTR.

A. Method

1) Participants: The study involved twenty international participants. The participant age range was 20-35 years with a mean age of 24.2 years ($SD = 2.7$). All participants were unfamiliar with a keep-left traffic regulation and right-hand drive vehicle. They each had a driving license of their home country with a mean driving experience of 5.9 years ($SD = 3.3$). They drove in average 14.7 hours/week ($SD = 9.0$).

2) Apparatus: The experiment was conducted on the Forum8 drive simulator, which is a fixed-based simulator. It represents a right-hand drive vehicle with three 42-inch monitors, creating a 150 degree horizontal and 30 degree vertical field of view. The monitors display the central and peripheral visual fields to the driver. UC-win/Road was used as software to model the traffic system in addition to locating the visual feedback on of the monitors. The driving data (e.g., speed, brake and steering wheel angel) was recorded for future analysis.

3) Driving Track, Scenario and Tasks: Using Forum8, the participants drove in a simulated keep-left traffic system (see Fig. 3- (a)). The driving track was similar to the track designed in [24]. The track included dual-lane roads crossed each other in three roundabouts and intersections. Participants were asked to drive in a certain direction, which forced them to drive through all roundabouts and intersections, each time with a different direction (left, right, and forward). Participants were asked to drive appropriately at those roundabouts and intersections following the Australian rules (see Section II-B).

4) Procedure: The experiment had three sessions - pre-experiment, preparation and two driving tests. In the pre-experiment session, participants were asked to complete an initial computer-based questionnaire. It was used to collect demographic information and driving experience data. In the preparation session, the driving simulator, Forum8, was introduced. Participants were verbally informed about the upcoming session and the Australian traffic rules, focusing on the rules for driving at roundabouts and intersections. In addition, they received a quick explanation of differences in vehicle configuration, such as the position of the wiper and direction indicator stalk, whereas the direction indicator stalk of a right-hand drive vehicle is placed the other way around. Then, participants were driving in a built-in test scenario for 10 minutes in order to familiarise themselves with the driving simulator. This scenario was supported by Forum8 and represented a keep-left driving environment. The scenario followed a route that differed from the test route to minimise the learning effect. During the familiarisation test, the researcher answered any questions participants might have. Participants had a 5-minute break after the familiarisation test while the researcher prepared the simulation for the driving test.

The third session (i.e. driving tests) had two driving trials - “Control” (no feedback offered) and “Experimental”, (providing feedback). The researcher randomly divided the 20 participants into two similar sized groups (10 participants each). Both groups participated in both trials. However, the order of the two trials was randomised between them in order to reduce the learning effect. Whereas the first group started with the control trial, the second group started with the experimental trial. There was a one-hour break between the two trials. Each driving test took around seven minutes to complete. The session started with providing the participants with a map of the driving track that indicated the direction the driver should follow (see Fig. 3- (b)). Participants took approximately three minutes to study the map. Then, they were asked to complete the driving tasks of each trial without making driving errors. In the experimental trial, participants did not receive any information regarding the feedback they would receive.

5) Data Collection and Analysis: Two log files (video and CSV) were generated by the simulator for each participant and for each trial. Both log files had data on driving performance. For each trial, we observed the following driving errors:

Error 1 Driving in an incorrect lane when approaching the roundabout and the intersection.
Error 2 Not indicating the target direction when approaching the roundabout and the intersection.
Error 3 Speeding while approaching the roundabout and the intersection.
Error 4 Driving in the incorrect direction of traffic flow inside the roundabout/intersection.
Error 5 Driving in an incorrect lane when exiting the roundabout and the intersection.

The number of the above errors that occurred at all roundabouts and intersections was calculated. The Wilcoxon signed-rank test was run to compare the number of each error in the experimental trial versus the control trial.

B. Results

Overall, the total number of errors among participant of each error type decreased when the feedback was provided. The difference in number of each driving error with and without providing feedback among participants was not normally distributed (see Table IV). For instance, the difference of driving in an incorrect lane when exiting the roundabout/intersection (Error 3) had a skewness of 1.032 (0.512) and a kurtosis of -0.230 (0.992). Thus, a non-parametric test (i.e. Wilcoxon signed-rank test) was employed to investigate differences between the driving errors before and after providing feedback.

Wilcoxon signed rank test (see Table V) showed that the numbers of Error 1, Error 2, Error 3 and Error 4 when providing augmented feedback were statically significantly fewer than than the numbers of those errors when augmented feedback was not provided ($Z = -2.648$, $p = .008$), ($Z = -1.998$, $p = .046$), ($Z = -2.167$, $p = .030$), and ($Z = -2.266$, $p = .023$), respectively. On the other hand, Wilcoxon signed rank test indicated that the number of Error 5 when providing augmented feedback was insignificant fewer than the number of Error 5 when the augmented feedback was not provided to participants ($Z = -1.642$, $p = .101$.).

C. Discussion

In general, presenting perceivable, comprehensible and projectable feedback significantly assisted the drivers who are not familiar with driving in a keep-left regelation to approach and enter roundabouts and intersections in a UFTR.

One participant mentioned indicating the target direction was very difficult as the direction indicator stalk is located on the right side of the steering wheel not on the left as it is on the vehicle that he used to drive in my home country. However, presenting perceivable, comprehensible and projectable direction indicator feedback significantly reduced the number of using improver direction indicators when approaching roundabouts and intersections ($Z = -1.998$, $p = .046$). One participant commented: ”it was easier for me to look at the indicator feedback instead of looking at the real stalks to find out which is the correct stalk to use, the right or the left”. Another participant mentioned that direction indicator feedback corrected him when he used the wiper indicator instead of direction indicator.

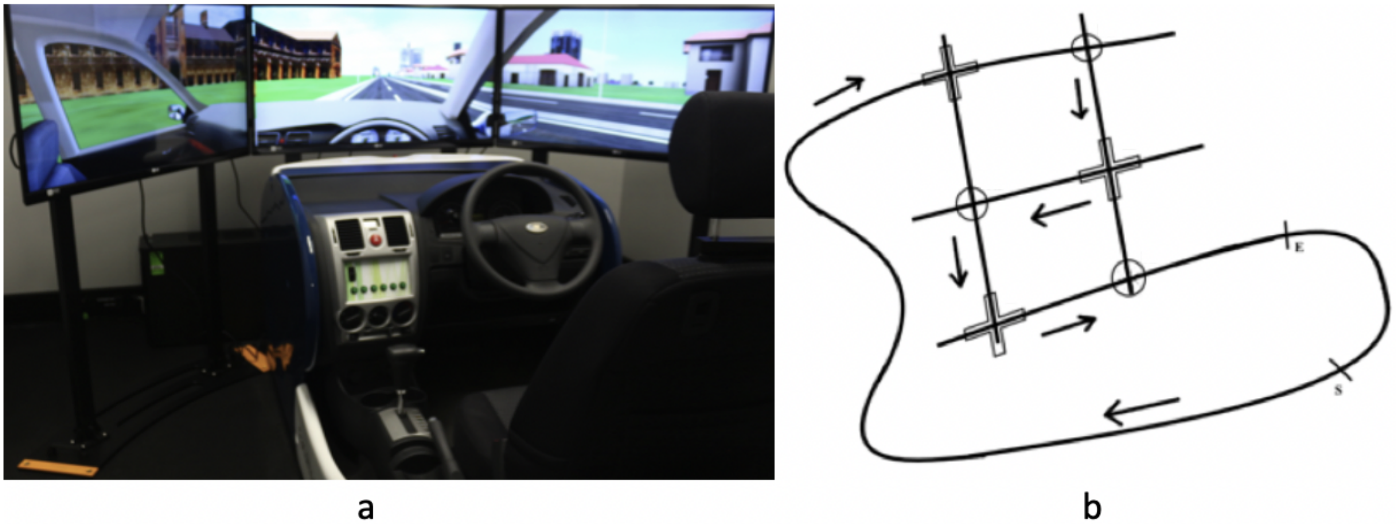


Fig. 3. (a) Forum8 (b) The Start and End Points and Required Direction on the Driving Track. Adapted from [24].

TABLE IV. Results of Normality Tests for the Difference in Number of Each Error Type with and without Feedback

Error	Shows (SE)	Kurtosis (SE)	Error	Shows (SE)	Kurtosis (SE)
Error 1	-2.573(0.512)	9.386(0.992)	Error 4	1.032(0.512)	-0.230(0.992)
Error 2	0.992(0.512)	1.712(0.992)	Error 5	-0.120(0.512)	-0.088(0.992)
Error 3	0.991(0.512)	0.263(0.992)			

TABLE V. Number of Driving Errors Grouped by Error Type with and without Feedback Support

Error	Number of errors (Control trial)	Number of errors (Experimental trial)	Difference	Z	p
Error 1	20	5	-15	-2.648	.008*
Error 2	17	11	-6	-1.998	.046*
Error 3	34	18	-16	-2.167	.030*
Error 4	15	1	-14	-2.266	.023*
Error 5	39	28	-11	-1.642	.101

* Statistically significant.

The correct direction when entering a roundabout in a UFTR is clock-wise, not as it is on the keep-right traffic system. In addition, the correct direction when entering an intersection in a UFTR is the opposite side of the road that the driver is familiar to use when driving in a keep-right traffic system. That might lead the drivers to make improper decision when approaching roundabouts and intersections in a UFTR. For instance, in case of approaching a roundabout, if the driver does not realise those changes in traffic rules, the driver might approach the roundabout using improper travelling lane. Also, the driver might enter the roundabout from the wrong direction. However, presenting perceivable, comprehensible and projectable travelling path significantly reduced the cases of approaching the roundabouts and intersections from incorrect lane ($Z = -2.648$, $p = .008$) and entering roundabouts and intersections from the incorrect direction ($Z = -2.266$, $p = .023$). One participant mentioned that travelling path feedback was very clear and understandable and that helped him to recognize the correct travelling path.

On the other hand, controlling the speed of the vehicle is a task that is not changed when driving in a familiar and unfamiliar traffic regulation. That is, the driver does not require to adapt his or her speeding behaviour when driving in a UFTR. However, presenting perceivable, well-known and projectable speeding feedback helped the drivers to significantly reduce their speed when approaching the roundabouts and intersections ($Z = -2.167$, $p = .030$). This result might be influenced by providing other information next to speed feedback. Other studies [20], [21], [22], [23] managed to reduce the drivers' speed without presenting speed feedback. More research is needed to explore the exact feedback that assisted the drivers to slowdown.

V. Conclusion

In this paper, we presented the results of a quantitative study that subjectively investigated the usability of augmented visual feedback that designed to help the drivers when driving at roundabouts and intersections in an unfamiliar traffic regulation (i.e. a keep-left traffic regulation using a keep-right drive vehicle). The usability was evaluated based on the situation awareness levels. That is, the information should be designed in a perceivable, comprehensible and projectable manner. Sixty-five participants who were not familiar with an Australian traffic regulation answered the online survey. The results enhanced the ability of the driver to capture the provided information, understand it and plan appropriately for the required driving reaction.

Then, another twenty participants who were not familiar with an Australian traffic regulation participated in an empirical study. The participants drove with and without providing the feedback in a simulated keep-left traffic environment using a driving simulator (i.e. Forum8). The study aimed to evaluate the effectiveness of the perceivable, comprehensible and projectable feedback in reducing the driving errors and thus improving the road safety. The results confirmed our hypothesis that considering the SA levels when designing the feedback can significantly reduce the driving errors when reaching and entering roundabouts and intersections and thus enhance road safety in a UFTR.

The current work focused mainly on driving at roundabouts and intersections in a UFTR. Thus, the work was limited by feedback that provided for a particular driving task. Another limitation was that presenting the visual feedback in accordance with SA levels was evaluated in a specific driving condition (a fine weather with no traffic movement).

The current work can be extended to cover the following points:

- Evaluate the usability of further feedback that designed to assist the drivers in different driving tasks, such as lane-changing and parking.
- Evaluate the effectiveness of the feedback in other driving conditions, such as rainy and traffic jam.

References

[1] S. J. Page and D. Meyer, "Tourist accidents: an exploratory analysis," *Annals of Tourism Research*, vol. 23, no. 3, pp. 666–690, 1996.

[2] J. Wilks, B. Watson, and J. Hansen, "International drivers and road safety in queensland, australia," *Journal of Tourism Studies*, vol. 11, no. 2, pp. 36–43, 2000.

[3] "Which countries drive on the left - a handy guide!" 2017. [Online]. Available: https://www.rhinocarhire.com/Customer-Services/Privacy-Policy.aspx

[4] Ministry of Transport, "Overseas driver crashes (including matched crash and visitor arrival data)," Ministry of Transport, Wellington, New Zealand, Tech. Rep., 2016.

[5] M. R. Endsley, "Situation awareness global assessment technique (sagat)," in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, vol. 3, 1988, pp. 789–795.

[6] J. Piao and M. McDonald, "Advanced driver assistance systems from autonomous to cooperative approach," *Transport reviews*, vol. 28, no. 5, pp. 659–684, 2008.

[7] L. Gugerty *et al.*, "Situation awareness in driving," *Handbook for driving simulation in engineering, medicine and psychology*, vol. 1, pp. 265–272, 2011.

[8] H. J. Alyamani and M. Kavakli, "Situational awareness and systems for driver-assistance," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[9] NSW Roads and Maritime Services, "Road users' handbook," 2015.

[10] N. A. Stanton, A. Dunoyer, and A. Leatherland, "Detection of new in-path targets by drivers using stop & go adaptive cruise control," *Applied ergonomics*, vol. 42, no. 4, pp. 592–601, 2011.

[11] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, "Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review," *Psychonomic bulletin & review*, vol. 20, no. 1, pp. 21–53, 2013.

[12] R. Schmidt and C. Wrisberg, *Motor learning and performance: A problem-based learning approach*, third edition ed. Champaign, IL: Human Kinetics, 2004.

[13] S. De Groot, J. C. De Winter, J. M. L. García, M. Mulder, and P. A. Wieringa, "The effect of concurrent bandwidth feedback on learning the lane-keeping task in a driving simulator," *Human factors*, vol. 53, no. 1, pp. 50–62, 2011.

[14] N. A. Stanton and P. M. Salmon, "Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems," *Safety Science*, vol. 47, no. 2, pp. 227–237, 2009.

[15] T. O. Heijer, H.-l. Wiethoff, M. Boverie, S. Penttinen, M. Schirokoff, A. Kulmala, R. Heinrich, J. Ernst, A. Sneek, and N. Heeren, "Action for advanced drivers assistance and vehicle control system implementation, standardisation, optimum use of the road network and safety advisors deliverable d1/2.1 v1: Problem identification, user needs and inventory of adas (advanced driver assistance systems): Final report," Contract No. DGTREN GRD1 2000–10047, Commission of the European Communities, Directorate-General for Energy and Transport, Brussels, Tech. Rep., 2002.

[16] P. Dalton, P. Agarwal, N. Fraenkel, J. Baichoo, and A. Masry, "Driving with navigational instructions: Investigating user behaviour and performance," *Accident Analysis & Prevention*, vol. 50, pp. 298–303, 2013.

[17] J. B. Van Erp and H. A. Van Veen, "Vibrotactile in-vehicle navigation system," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 7, no. 4, pp. 247–256, 2004.

[18] M. Baumann, A. Keinath, J. F. Krems, and K. Bengler, "Evaluation of in-vehicle hmi using occlusion techniques: experimental results and practical implications," *Applied ergonomics*, vol. 35, no. 3, pp. 197–205, 2004.

[19] H. J. Alyamani, A. Hinze, S. Smith, and M. Kavakli, "Preference feedback for driving in an unfamiliar traffic regulation," in *Service Research and Innovation*. Springer, 2018, pp. 35–49.

[20] E. Becic, M. P. Manser, J. I. Creaser, and M. Donath, "ntersection crossing assist system: Transition from a road-side to an in-vehicle system," *Transportation research part F: traffic psychology and behaviour*, vol. 15, no. 5, pp. 544–555, 2012.

[21] J. Zhang, K. Suto, and A. Fujiwara, "Effects of in-vehicle warning information on drivers' decelerating and accelerating behaviors near an arch-shaped intersection," *Accident Analysis & Prevention*, vol. 41, no. 5, pp. 948–958, 2009.

[22] C. Tran, K. Bark, and V. Ng-Thow-Hing, "A left-turn driving aid using projected oncoming vehicle paths with augmented reality," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2013, pp. 300–307.

[23] J. Caird, S. Chisholm, and J. Lockhart, "Do in-vehicle advanced signs enhance older and younger drivers' intersection performance? driving simulation and eye movement results," *International journal of human-computer studies*, vol. 66, no. 3, pp. 132–144, 2008.

[24] H. J. Alyamani, M. Kavakli, and S. Smith, "Vehand: an in-vehicle information system to improve driving performance in an unfamiliar traffic regulation," *International journal of human factors and ergonomics*, vol. 6, no. 4, pp. 355–389, 2019.

[25] N. L. Leech, K. C. Barrett, and G. A. Morgan, *SPSS for intermediate statistics: Use and interpretation*. Psychology Press, 2005.

An Interactive Tool for Teaching the Central Limit Theorem to Engineering Students

Anas Basalamah
Department of Computer Engineering
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—The sole purpose of this paper is to guide students in learning the introductory statistical concepts, such as, probability distribution and the central limit theorem (CLT) in an intuitive approach through an interactive tool. When a used data has different probability distributions, this paper intends to clarify the notions of the CLT and the use of samples in the hypothesis testing of a population by demonstrating step-by-step procedures and hands-on simulation approach. This paper discusses the relationship between the sample size and the nature of the sampling distribution, which is a vital element of the CLT, in different population distribution using the developed interactive tool. Finally, the impact of the developed interactive tool is measured via a survey experiment that illustrated the success of the developed tool in teaching the CLT.**

***Keywords*—Probability distribution; CLT; population; interactive tool; sampling distribution**

I. Introduction

The fundamental concept of statistics field in which students more often face difficulties to understand is the central limit theorem (CLT). The CLT provides an overview of using random sampling method for making an inference about any population. Therefore, it is considered as a vital concept in inferential statistics, a critical knowledge for any statistician, and one of the foundation concepts of any statistics course. In [1], authors highlighted to focus on developing the ideas of central statistics before moving on to the set of tools and procedures. The CLT is one of the central ideas of statistics. While learning it, students get puzzled more often to interpret the theory and implication of this important concept. This paper aims to create an engaging way of teaching and learning the theory and implication of the CLT for both teachers and students through explanation, simulation, and visualization. We built an web application (interactive tool) using html, css, and javascript for generating and visualizing the uniform, normal, positively skewed, and negatively skewed distribution with corresponding sampling distribution.

Students can collaborate by selecting their desired distribution, sample size, and number of samples during the simulation process in the class for giving them a part of the simulation process and creating an engaging environment in the class. Students can select different sample size and number of samples for a particular distribution that enables them observing the changes of sampling distribution more closely with the changes of these parameters. It helps the student to understand the impact of sample size on the sampling distribution. The sampling distribution is the probability distribution obtained from a large number of samples, which are drawn from a particular population. It is the distribution of frequencies of all possible outcomes that could occur for a statistics of that particular population.

The pedagogical approach taken in this paper for teaching CLT is different comparing to the other approaches because of the features of the developed interactive tool, which creates an engaging environment for the students and ensures students participation alongside to make the understanding of the CLT clear to the students.

The remaining part of this paper is organised as the following: Section II discusses the background and motivation of this research topic along with the literature review, Section III studies the concept of the central limit theorem and different probability distributions, Section IV describes the importance of the central limit theorem, Section V shows the empirical demonstration of the CLT with the help of visualization using the interactive web-application tool, Section VI investigates the impact of the developed tool in teaching the CLT in the classroom, and Section VII concludes the paper.

II. Background

To make valid statistical interference about the concept of the CLT and sampling distribution, students must get the opportunity to draw multiple samples [2]. This paper is highly motivated by the recommendation given in [2], hence it incorporates multiple samples in simulation to clear the logic of the CLT to the students. The simulation provides a realistic scenario to the students for intuitively understanding these concepts. Although, simulation is not exempt of issues [3], but it is the best way of teaching the CLT to the students.

In [4], [5], the authors investigated that the concept of CLT is not only difficult for the non-math major students, but also for math major students. They added that the central limit theorem is the most vital result conveyed theorem in the introductory statistic course since it includes many of the statistical inferences that are required for the later part of the course. Although, students are lost more often in understanding the logic behind the CLT. Therefore, it becomes a necessity presenting the concept in a intuitive way through simulation such that it can be easily understandable to the students.

Interactive learning experience has always been considered as the most effective pedagogical tools [6], [7], [8], [9], [10], [11] and the use of interactive environment is considered as the effective way of learning and visualizing a complex theory. There are a large number of simulation tools and online applets for learning the CLT as well [12], [6]. Most of these simulated environments has complex interface, requires knowledge of programming, and has complex procedures, which makes students less attractive to these simulation applets. The time spent to these applets for understanding different distributions and the CLT is comparatively higher and less engaging. The sampling distribution of any given population always follows normal distribution. To clear this concept, this paper brings all the distributions along with various sample sizes together in a single interface without showing the background codes that helps students get the idea easily in the shortest time. This paper intends to bring a simulated interactive teaching applet, which can be used both online and offline by the students and can meet all the above criteria.

In sum, the above stated context motivates us working on the same problem and providing a more interactive method of teaching CLT to the students. The interactive tool designed for the students has an interactive interface that allows the students to take different size and different number of samples iteratively for plotting their respective sampling distribution. As a result, students get an overview on how the estimated means of different samples with a particular sample size get centralized and form a bell-shaped curve, which is the principle of the CLT. The following section investigates different probability distribution and the concept of the CLT along with its properties.

III. Population Distributions, Sampling Distribution, and the CLT

A. Population Distribution

The central limit theorem states that the sampling distribution of a given population forms a bell-shaped curve or follows normal distribution regardless the variable's distribution in the population. The distribution of a variable can follow different probability distributions, i.e. uniform distribution, normal distribution, positively skewed distribution, and negatively skewed distributions (see in Fig. 1).

The distribution of a variable is the distribution of the random sample that is drawn from the population. The CLT acts on all the probability distribution that have a finite variance. Therefore, the CLT cannot be applied on Cauchy distribution because of having infinite variance. In addition, the CLT acts on independent and identically distributed variables, where the value of a particular variable does not depend on the values of other variables. But the distribution of all these variables must remain constant throughout the measurement process.

B. Sampling Distribution

The sampling distribution is the mean of the randomly drawn samples from the population. For example- lets assume that someone draws a sample with a fixed sample size from a given population, then calculates its mean and plots it on a histogram. If this process in repeated many times, the produced histogram displays the distribution of the sample mean. This mean distribution is considered as the sampling distribution in statistics.

The shape of the sampling distribution depends on the sample size and number of samples. For different sample size the shape of the sampling distribution differs. Similar observation found with the number of samples. The following sections discusses these phenomena in detail.

C. The Central Limit Theorem

The central limit theorem states that the sampling distribution of sufficiently large size of samples drawn with replacement from a given population that has the mean μ and standard deviation σ, will form an approximated normal distribution curve. This statement is true regardless the distribution of the population. Usually, a sufficiently large sample size is the samples having size of 30 or more ($n \geq 30$). If the population distribution is normal, the CLT is true for smaller size of samples as well. If the population distribution is strongly skewed, it may require larger size and number of samples. The relationship between the sample size and the shape of the sampling distribution is clarified in section V.

Also, the CLT statement holds correct for the binomial population distribution, provided that $min(np, n(1-p)) > 5$, where p is the probability of success and n is the sample size. As a result, normal probability distribution can be used to quantify uncertainty for making inferences about a population mean.

1) Properties of CLT: There are two attribute of any distribution, namely- the mean (μ) and the standard deviation (σ). The sampling distribution converges the normal distribution, when its mean is equal to the population mean and the standard deviation is $\sigma/\sqrt{n}$. The standard deviation σ decreases by $\sqrt{n}$ with the increment of the sample size n.

In sum, the sampling distribution approximates the normal distribution with the increment of sample size and the spread of the distribution suppressed. These properties have significant implications, which will be discussed in the later sections of this paper.

IV. Importance of the CLT

The central limit theorem is important because of the following to reasons, namely- normality assumption and precision estimates.

A. CLT and the Normality Assumption

The normality assumption is very essential in statistics for parametric hypothesis testing of the mean, i.e. t-test. The CLT supports the assumption as it states that the sampling distribution of any population can approximate a normal distribution. This critical implication of the CLT allows the hypothesis testing even if the data is non-normally distributed. However, the testing is allowed if and only if the sample size is large enough. Because, a non-normally distributed data also behaves like a normal distribution for larger sample size.

Moreover, parametric tests of the mean are robust from the normality assumption when the sample size is sufficiently large. This is also a contribution of the central limit theorem.

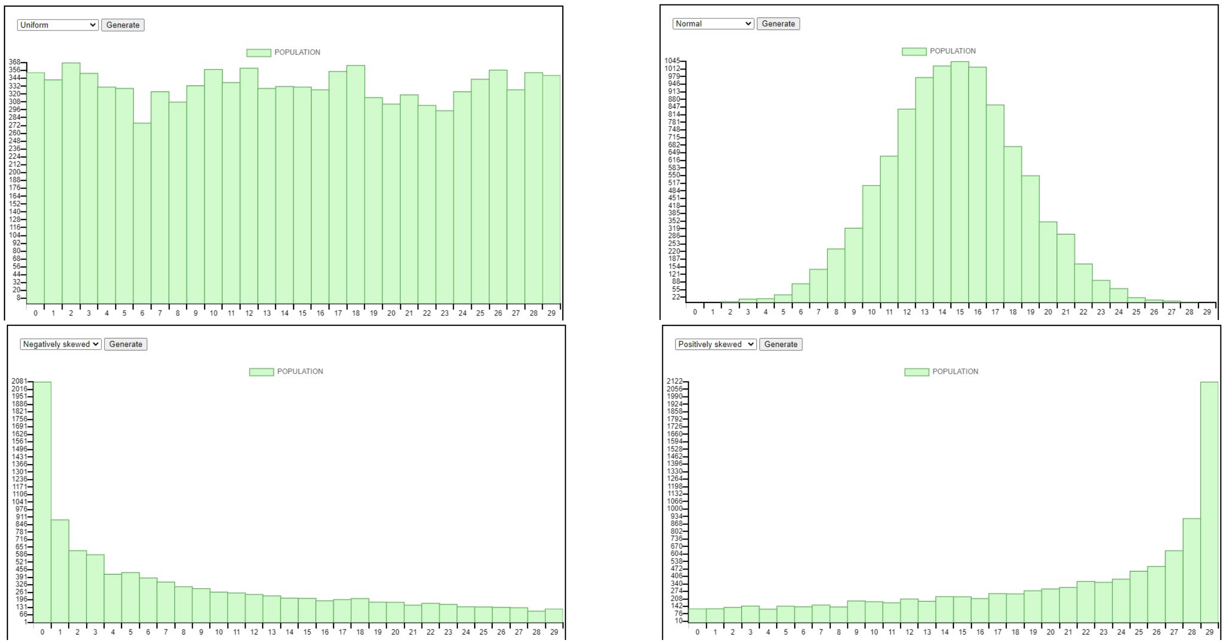


Fig. 1. Different Population Distributions, Namely- uniform, Normal, Negatively Skewed, and Positively Skewed Distribution.

B. Precision of Estimates

In all the figures, the sampling distribution of the mean clusters around the population and becomes denser as the sample size increases (see in Fig. 2, 3, and 4). This property of the Central Limit Theorem is relevant when using samples to estimate the mean for the entire population. The larger the sample size, it is more likely that the sample mean will be closer to the actual population mean. In other words, the estimate is more accurate.

On the other hand, the sampling distribution of the mean is much wider for smaller sample sizes. When the sample size is small, it is not uncommon for the sample mean to be away from the actual population mean. In this case, the estimate will be less accurate. Finally, understanding the Central Limit Theorem is important when relying on the validity of the results and assessing the accuracy of the estimation. We should use a large sample size to meet the normality assumptions and get more accurate estimates even if the data is non-normally distributed.

V. Empirical Demonstration of CLT

This section demonstrates the central limit theorem with respect three different different population distribution, namely- uniform distribution, normal distribution, and severely skewed distribution. For illustrating the impact of CLT on severely skewed distribution, we consider the negatively skewed distribution. All the population distributions taken for the demonstration are shown in Fig. 1.

A. The CLT with Uniform Distribution

At first, we consider the uniform distribution shown in Fig. 1. We take the sampling frequency distribution for the sample size less than 30 and greater than 30 in Fig. 2, respectively. We observe that the obtained sampling distribution for the sample size of 10 is ranging from 6.9 to 22.7, whereas the sampling distribution for the sample size of 30 ranges from 6.9 to 22.7. Besides, the shape of the sampling distribution is more normal using the sample size 50 comparing to the sample size 10.

Therefore, we can conclude that if we increase the sample size, the sampling distribution gets more tighten and centralized in uniform distribution.

B. The CLT with Normal Distribution

Similarly, we check the behaviour of sampling distribution using a normal distribution that is depicted in Fig. 1. We consider sample size of 10 and 50, respectively, and observed the similar conclusion as uniform distribution. The demonstration of CLT using normal distribution is depicted in Fig. 3. We observe that the obtained sampling distribution for the sample size of 10 is ranging from 11.1 to 17.9, whereas the sampling distribution for the sample size of 30 ranges from 12.68 to 16.48.

Hence, it is clear that the sampling distribution for larger sample size is more centralized comparing to the smaller sample size. We further observe that sampling distribution for the normal population using sample size of 10 is more tighten and centralized comparing to the sampling distribution of uniform population using sample size of 50 (see in Fig. 2 and Fig. 3).

C. The CLT with Skewed Distribution

Finally, we take a highly skewed population distribution, i.e. negatively skewed distribution depicted in Fig. 1, for testing the sampling distribution behaviour and investigating the CLT. The same process is followed for the skewed population analysis as before. Two sample size of 10 and 50 are considered. The obtained sampling distribution figures for both the sizes

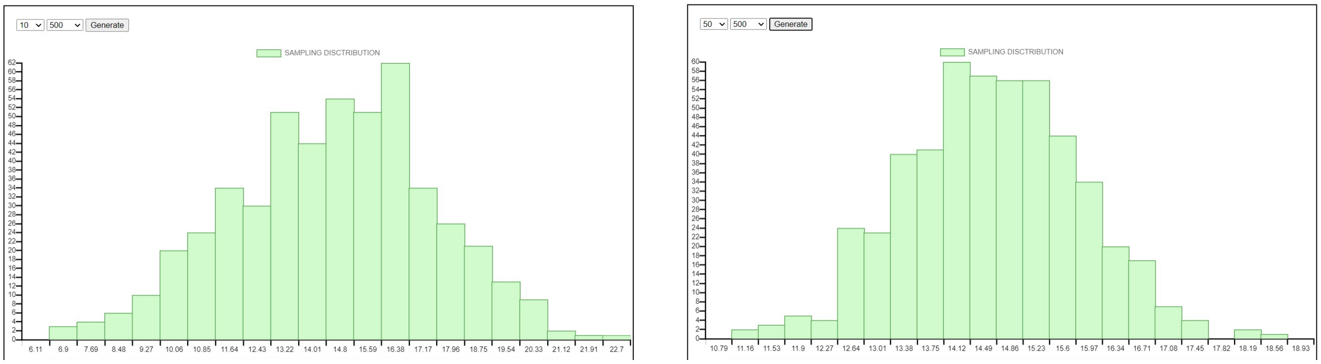


Fig. 2. The Central Limit Theorem Explained with Uniform Distribution.

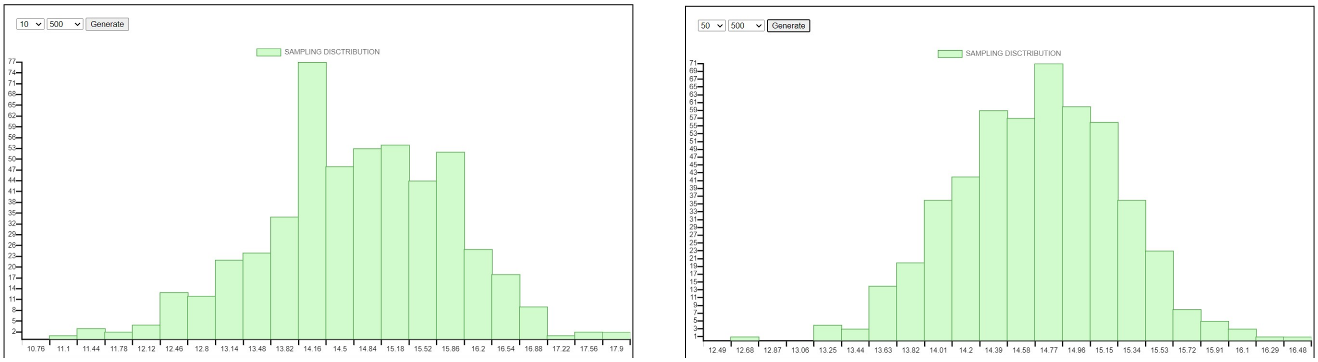


Fig. 3. The Central Limit Theorem Explained with Normal Distribution.

are depicted in Fig. 4. Similar observation is found, i.e. if the sample size increases the obtained distribution gets tightened and the shape approximates a normal bell-shape.

The range of sample mean obtained for the size 10 is from 1.6 to 16.6 and for the size 50 is from 4.92 to 11.92, which supports the previous statement and justifies the central limit theorem.

VI. Impact

We have developed an interactive teaching applet using HTML5, CSS, and Java-script. The interactive tool is a dynamic, responsive, and device independent web-application that can be opened by double clicking on the index html file. Students only need to have a web-browser installed in their device. The teaching applet has a attractive and simple user interface, where user can generate different population distributions, can select random samples of different sizes, and can plot their sample mean. Student can add a single sample to the sampling distribution iteratively or can add large number of samples at a time, which enables them to visualize the changes occurred in the sampling distribution after adding each sample to the sampling distribution. Students can easily absorb the motto of the CLT by watching how the sampling distribution becomes tightened after the addition of new sample means to the distribution.

We observed that all the students got the gist of the CLT easily after taking the class with the help of the developed interactive teaching applet. Therefore, we can claim that the interactive tool developed by us can successfully be used during the teaching/learning process of the CLT.

VII. Conclusion

This paper intends to provide a simulation based overview of the central limit theorem, which intuitively helps the students for cracking the concept. The relationship of the sample size and the sampling distribution is well illustrated throughout the empirical analysis section, which is an essential concept for the students as they go deep into their statistics course. The last section shows the impact of this study and the developed interactive tool about how well the students were able to absorb the concept after using the developed tool.

References

[1] P. Cobb and K. McClain, "Principles of instructional design for supporting the development of students' statistical reasoning," in *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 2004, pp. 375–395.

[2] T. Hodgson and M. Burke, "On simulation and the teaching of statistics," *Teaching Statistics*, vol. 22, no. 3, pp. 91–96, 2000.

[3] M. E. Brussolo, "Understanding the central limit theorem the easy way: A simulation experiment," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, no. 21, 2018, p. 1322.

[4] M. L. Lunsford, G. H. Rowell, and T. Goodson-Espy, "Classroom research: Assessment of student understanding of sampling distributions of means and the central limit theorem in post-calculus probability and statistics classes," *Journal of Statistics Education*, vol. 14, no. 3, 2006.

[5] A. E. Watkins, A. Bargagliotti, and C. Franklin, "Simulation of the sampling distribution of the mean can mislead," *Journal of Statistics Education*, vol. 22, no. 3, 2014.

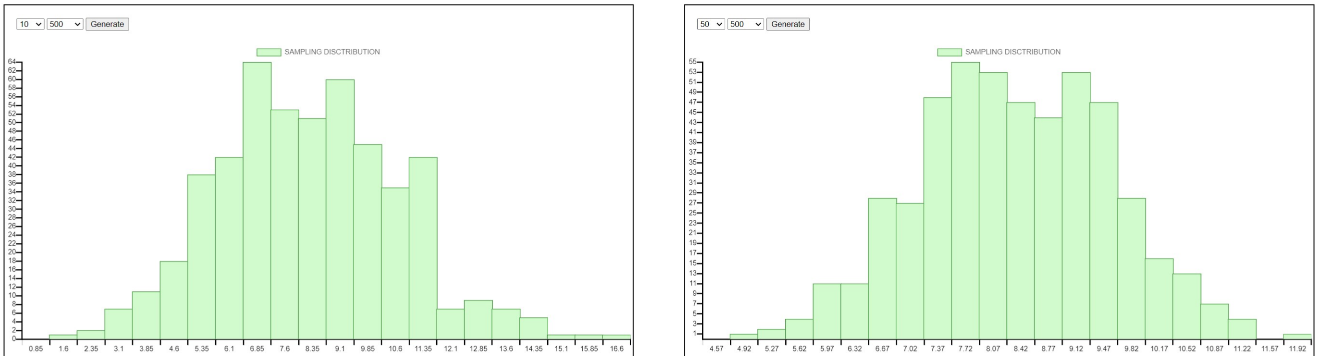


Fig. 4. The Central Limit Theorem Explained with Negatively Skewed Distribution.

[6] S. P. Gordon and F. S. Gordon, “Visualizing and understanding probability and statistics: graphical simulations using excel,” *PRIMUS*, vol. 19, no. 4, pp. 346–369, 2009.

[7] J. Garfield, “Teaching statistics using small-group cooperative learning,” *Journal of Statistics Education*, vol. 1, no. 1, 1993.

[8] K. A. Carlson and J. R. Winquist, “Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach,” *Journal of Statistics Education*, vol. 19, no. 1, 2011.

[9] T. J. Pfaff and A. Weinberg, “Do hands-on activities increase student understanding?: A case study,” *Journal of Statistics Education*, vol. 17, no. 3, 2009.

[10] R. K. Steinhorst and C. M. Keeler, “Developing material for introductory statistics courses from a conceptual, active learning viewpoint,” *Journal of Statistics Education*, vol. 3, no. 3, 1995.

[11] D. M. Jamie, “Using computer simulation methods to teach statistics: A review of the literature,” *Journal of Statistics Education*, vol. 10, no. 1, 2002.

[12] E. Ruggieri, “Visualizing the central limit theorem through simulation,” *PRIMUS*, vol. 26, no. 3, pp. 229–240, 2016.

Improving the Effectiveness of e-Learning Processes through Dynamic Programming: A Survey

Norah Alqahtani[1], Farrukh Nadeem[2]
Department of Information Systems,
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—E-learning has been widely adopted as an important tool for distance education, especially in these days of pandemic Covid-19. However, several problems/challenges have been reported in different processes of e-learning that need to be addressed for effective use of e-learning. These problems/challenges include development of student focused contents, giving learner partial control, addressing different learning styles, etc. Recently, several efforts have been made to solve e-learning process problems using dynamic programming techniques. Dynamic programming techniques divide a problem situation into several subproblems and dynamically solves each sub-problem based on student needs. Thus it allows student focused customization at each step and provides adaptive e-learning to support students with different capabilities. The objective of this study is to review different e-learning problems and challenges and how those can be addressed using dynamic programming techniques. We conclude by highlighting the importance of different dynamic programming techniques for different processes of e-learning.**

Keywords*—*e-learning; e-learning challenge; dynamic programming

I. Introduction

Developments in information and communication technology (ICT) have introduced E-learning programs which can be employed efficiently as an alternative to allow users to learn and teach new technologies and courses. E-learning refers to knowledge that uses electronic technology beyond the traditional classroom to access curricula. It can be applied to a course, program, or diploma that is entirely online [1]. E-Learning online courses are available 24/7, regardless of the time of day, accessible and reliable educational resources to students in geographically spread regions and time zones. This type of delivery enables the portability of training (i.e., tablet/iPad, laptop, cell phone for files or video assessments and links to webinars, etc.), allowing for easy learning on the go, particularly with the growth of networks, mobile use, and learning via computers or laptops [2]. There are numerous e-learning systems and methods (known as learning management systems) that can be used to conduct online courses. Given the right tools, various processes can be automated, such as scoring or viewing reading materials. E-learning is often the most efficient way to adapt the learning to people's busy lifestyles, allowing individuals to gain new skills and develop their careers. In recent days, the importance of e-learning has been recognized, as the COVID-19 pandemic crisis has made e-learning mandatory in all institutions of learning such as schools, colleges and universities throughout the world [3]. In fact, The problems of e-learning can be summarized in five major issues such as: struggle for adaptability, technical problems, computer literacy, time management, and self-motivation. Therefore, Many methods have been used to improve the e-learning process that has helped to improve this area and increase its level of effectiveness; one of these methods is Dynamic Programming (DP). Dynamic programming techniques which are ideal for embedded systems are increasingly comparable to adaptive control techniques for dynamic systems. A growing number of researchers are inspecting and studying dynamic programming (DP) algorithm-based learning systems to solve and fix stochastic ideal control problems, arguing that DP provides the proper basis for combining planning results into real-time reactive dominance techniques, as well as for learning these techniques when the managed system is incompletely understood. This paper reviews how the DP is contributing to solving some of the problems in the field of e-learning through various studies. This paper's structure is as follows: as an introduction, we first present the definition of the DP and then provide a description of the DP. Next, the role of the DP in improving learning is discussed. The fifth section is a comparative analysis that concludes the studies presented in this paper, and the final section concludes the article.

II. Dynamic Programming

Dynamic programming allows to overcome an optimization problem (e.g. optimizing, decreasing, or finding the total number of ways to do something). It is defined as a quantitative technical analysis that has been used for fundamental and sophisticated problems that require a series of decisions to be made [4]. In [4], the authors defined DP as one of the algorithms used to find the ideal value of a problem. This problem's solution is decomposed into several steps (subproblems) until the solution of a required problem can be considered a succession of interdependent decisions.

A. Steps of Dynamic Programming

DP problems require four phases [4]:

- Break the basic problem into subproblems known as phases.
- Accomplish the last problem stage for all potential states.
- Returning from the last phase, resolve every middle phase. This is completed by defining ideal policies from that point to the terminus (final stage) of the problem.

- Get the ideal solution to the original problem by sequentially solving all the stages.

Using the DP, sub-problem results are stored. Therefore there is no need to recalculate these results the next time they are requested. This is an alternative to simple recursion, which requires repeating the solution operation each time the sub-problem is encountered. In (Programming n.d.) address these phases as following: description,definition, computation and construction. knabsack problems and shortest paths are two problems that predominate in studies and research on computer algorithms [5]. The issue of knapsacks goes back more than a century, i.e., we need to calculate the amount of each item to be included in a collection for a set of objects, each with a weight and a value whereby the overall weight should be smaller or equal to a specific limit. The overall value is set as high as possible [5]. The shortest path's problem consists of obtaining a path between two vertices in a diagram or such that the total of the weights of its edges decreases, described in [6]. In this paper also we introduce the longest common subsequence problem and illustrate how to enhance learning area.

III. The Role of Dynamic Programming in e-learning

DP is a useful technique for optimizing problems, which pursue the maximum or minimum solution since it allows to browse all possible sub-problems without defining a specific answer. It ensures that the methods used to solve or approximate algorithms are accurate and efficient. Futhermore, it provides a general means of modeling and solving such consecutive learning problems. This formulation offers value in three respects. First, the formulation of the DP allows the optimal solution to be calculated. This is usually only true for small problems [7]. Second, the formulation of the DP sometimes allows structural results to be shown in a theoretical way, which either provides insight into the problem and the behavior of the optimal policy or provides a characterization of the optimal system that can be directly exploited to find an optimal approach, as in sequential hypothesis testing. Third and finally, the DP sometimes provides heuristics that are beneficial for large-scale and complex learning problems.

DP approved his role to enhance and optimize problems in learning during the following studies:

A. Shortest Path

Shortest path is the problem of getting the lowest rout in a graph from one vertex to other. In e-learning systems, a course is considered as a graph where each node represents a unit of knowledge. The shortest path's challenge is to find a routing between the peaks of the graph mentioned above so that the overall total of the edge weights is minimal. The Bellman-Ford algorithm (BFS) is one of the algorithms that may be used to solve this problem.

In [8] a DP approach is employed to choose the shortest and the optimale path in virtual learning environments (VLEs). This environment catches a large number of students to allow them to learn and study everywhere. The intention is that the location of the student is considered to be no longer a band [9]. In [8], the author applied an analytical hierarchy process (AHP) method for transforming parameters from qualitative to quantitative ones. The system of VLE is suggested on the basis of an in-depth personal learner profile. The objective is to select the best route compatible with a learner's attention and qualitative properties like attitude, motivational capacity, learning style, and knowledge level. The user in the system, is confronted with service providers who prop VLEs. Every provider of service provides learning subject and internships that are educated by different teachers. The layers are presented in Fig. 1. [8]

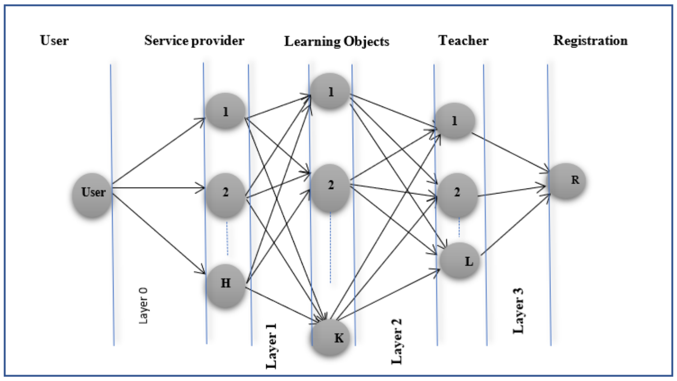


Fig. 1. The Suggested Network [8].

In the above figure, a network has been presented in which a dynamic process is implemented to choose the shortest route for each learner. The selection of the user depends on the qualitative criteria of every route. These paths must have a quantitative value to implement the DP. To achieve this, the qualitative criteria are converted into numerical values according to the user's preferences (the numerical values are shown in the table in the study paper). Then, the path preferences are related with numeral values. Three matrices are constructed in this study to calculate the ratio of the numerical values of pathways, which are estimated by the learning styles, attitudes, abilities, motivations, and knowledge levels of users based on the numbers of preferences. In this manner, the weights are calculated, and the ideal pathway for learners will be specified.

In [10], the author notes that they aim to estimate the prosperity and/or validation of education systems with a specific pipeline, unlike successful e-learning architectures. He suggests an evaluation procedure that consists of applying the DP process to model the user profile's ideal path problem and applying reliability to measurement the interconnections between users in a network of e-learning. This study has endorsed the effectiveness of DP through a comprehensive example. It is identical to the previous study, resulting from this example presented in Fig. 2 below [10]. In below Fig. 2, we can see the service providers number are five, the number of presented course are five and there are five teachers

According to the calculation of numeric values, the ideal path as presented in Fig. 2 is 0-3-9-12-16 that means the group included of the 3rd provider of service, the 4th learning object, and the 2nd educator is identified to become the shortest and the optimal. DP contributes to solving Markov decision process to enhance the learning path of learns in [11]. The atmosphere for adaptive learning is random. Therefore, to establish an

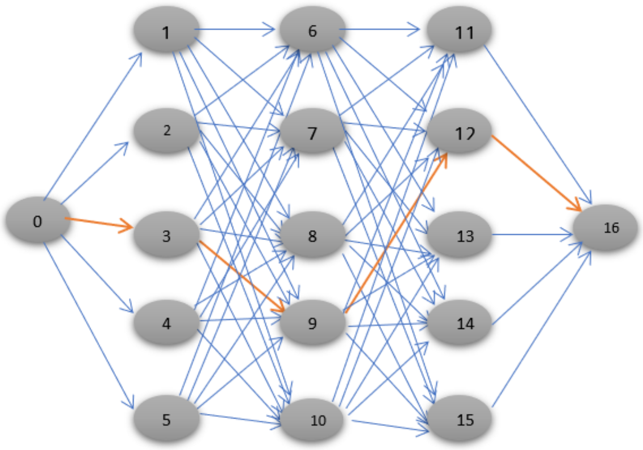


Fig. 2. Finding the Best Path [10].

effective learning path, the Markov Decision Process (MDP) model is suitable. Besides, the best activity in a sample space [12] is observed. Therefore, the MDP is recommended to provide a structured format to describe a multi-stage decision-making process in a probabilistic context. DP approach utilized to resolve the MDP. The issue in this study, selecting the bestead learning path with the highest reward, this problem an optimal substructure it can be split into subproblems, for that DP is the most proper way of solving it [13]. It is almost similar to that of a traveling salesman problem who wants to find the shortest way to increase his bonus. The learning objects it will consider as nodes, and learners' learning style as routes between nodes (in this study, used the Felder-Silverman Learning Style). According to Markov's process, the selected path will offer the highest cumulative remuneration [14]. After constructing the DP code by MATLAB program, the shortest path among two learning materials (learning objects) are computed for all successive LO. Fig. 3 illustrates the result of this study [10].

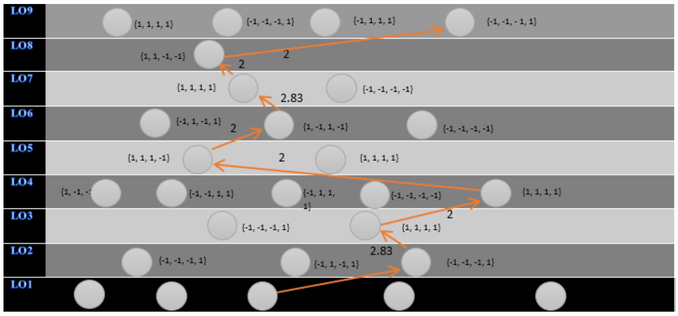


Fig. 3. Optimal Learning Path for Learner with Style 1, 1, -1, 1[11].

B. 0/1 Knapsack

Knapsack problem is considered as a combinatorial improvement problem in which each item is related with a profit and a weight. This problem aims at maximizing the number of items in a knapsack by matching capacity of knapsack.
The 0/1 knapsack problem can be declared as: offered a limited number of items and a knapsack, find the total revenue under the constraint that the overall weight of all selected items at maximum the weight boundary. However, present learning material is uniform and constant during each grade without concentrating on individual trait's or group differences. This problem makes materials of educational not attractive to some students. Establishing a progressive direction in compiling student books is a playful way of meeting students' demands and enhancing education efficiency. The study [15] In 2012 discussed the problem when the learner meets with a vast amount of content through the process of learning while he/she has a fixed time, and for this, he/she may sense disappointed and frustrated. The study proposed a novel procedure to aid learners get learning content more personalized to their characteristics and background in a playful manner; it used the 0/1 knapsack problem to choose the most suitable adaptive learning materials in a limited amount of time specified by the learner. There are various content categories in learners' books, like texts, exercises, examples, etc. These are known as elements. For each of these items or components, a specific part of the time allotted for teaching a book is set aside.

Students take the regular exam first, and their learning rate is evaluated. Then, the outcome of this test is converted into a coefficient. The time required for each of the upper elements to be taught determined by multiplying the original time allocated by each element's coefficient. For all the features, we can also measure time [10]. To provide the most appropriate learning objects to the learner in a restricted period, the 0/1 Knapsack problem has been applied. This problem is one of the most known combinatorial improvement problems. It is known as a classical NP-hard, which has an excellent search space.

The 0/1 knapsack problem (KP) is appropriately known as follows: We are given n items and a knapsack. Item i weights wi, and the knapsack has a weight boundary C. If object i is put into the knapsack, we will gain a profit p. The issue is to increase the overall profit under the constraint p that the sum weights of all selected objects are at most C. For that, the knapsack problem can be formulated as:

$$\begin{array}{ll} \text{maximize} & z = \sum_{j=1}^{n} p_j x_j \\ \text{subject to} & \sum_{j=1}^{n} w_j x_j \leqslant c \\ & x_j \in \{0,1\}, \quad j \in \{1,\ldots,n\} \end{array} \tag{1}$$

Where p denot to the profit ,w refer to the weight and x is a binary variable indicating whether object i is selected or not, in the above Equation, p the constraint that requires to be convinced and realizes the profit of a feasible n-tuple, the following Fig. 4 illustrates this problem [10]. Each student has his/her account and personal profile. When coming to the application for the first time, many questions were faced by each learner to reach the personalities of learner, characteristics, and learning styles. This application, which is established in the suggested way, showed a significant effect on adaptive learning.

Dynamic programming has proven its role in universities through scholarship programs. Bursaries serve as an important motivation and play a vital role in inspiring students to work hard at the universities, exemplify, and strive in their academic career in higher education [16]. It is often viewed from the teaching perspective as an essential factor. Many research findings have shown that bursaries positively impact college

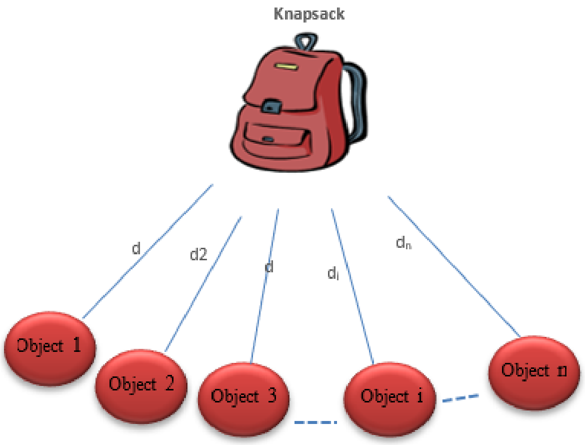


Fig. 4. The Description of Knapsack Problems [10].

performance and inspire students to pursue further education [17]. Various scholarships are sponsored by multiple donors, such as entrepreneurs, companies, citizens, etc. responsible for deciding the number of awarded persons and the cost allocated to any individual beneficiary. The university is involved for awarding the scholarships to applicants by the rate on departmental results. Yet university administrators are not allowed to split the current scholarships further into new scholarships with lower rewards.

In [16], they proposed a challenging DP approach to change the conventional process of decision-making for awarding scholarships. The aim is to figure out the ideal scholarship allocation with the greatest justice while taking into account practical channels and the requirement for fairness. This study explores a department's scholarship allocation process. A variety of scholarships are available to graduate students from a department who have done exceptional work in the course of their university studies. In two ways, multiple types of scholarships are different the number of the offered ones and the money allocated for these scholarships. For various scholarships, students can apply free of charge. The outcomes of the scholarship assessment depend on the scores received by the students in the previous year. A student's total score consists a score named basic, which is calculated based on coursework and some additional study findings provided by the academic accomplishments, such as journal papers. The score mentioned above is calculated in a scale to simplify student work assessment, whereas various forms of achievements are estimated. Also, to turn a particular work into standard ratings, a standardization procedure is implemented. For example, the score of a published paper can be measured in two components: Rating the impact factors of journals and the order of the authors. The study assumed that the department was responsible for the allocation of scholarships and the scholarship was distributed according to the rank of the applicants' final scores. To prevent unfairness in distributing scholarships, we add the limitation that at least no less scholarship should be given to students with a higher score.

Otherwise, the research takes into account other realistic constraints. For example, multiple students cannot share a single scholarship; they can only be rewarded no more than once for a specific form of scholarship, various grants; students should be assigned all the grants. The ideal assignment is the method that guarantees that the sum of a student's scholarship is corresponding to the scale of the overall score.

TABLE I. SCHOLARSHIP DETAILS (HUANG ET AL. 2018)

ID	1	2	3
Scholarship value	8000	5000	10000
Number of awardees	2	3	1

TABLE II. CANDIDATES DATA (HUANG ET AL. 2018)

ID	1	2	3	4	5
Score	25	20	30	35	15

The scholarship assignment problem was viewed as a join of two sub-problems. One is seeking the available assignments schemes that pass the constraints and the primary justice demand together [16]. The second, lies on the quantification of each sensible scheme's justice, then, choose the ideal strategy with the top justice. To handle the first subproblem, a way formed of knapsack sub-problems is advanced to allocate in sequence students to scholarships. In the knapsack sub-problem, every student's overall mark is considered the knapsack, and the valent mark of every scholarship is considered as the element placed in the knapsack [16]. The Gini coefficient to calculate unfairness of each feasible solution is added to resolve the second sub-problem. The equity of welfare distribution can be quantified by the Gini coefficient (variation among the real and ideal sum of scholarship earned by every single student) [16]. This study explained the proposed method through a small case: three scholarship kinds and five candidates as an example, as we see in the below Tables I, II and III.

Then, we can calculate the coefficient of Gini of the individual plans, as mentioned in Table IV which is indicated in the schedule IV: As the Scheme 1 Gini coefficient is the smallest, Schema 1 is the one with the highest equity as shown in the Table V. In this numerical example, scheme 1 is therefore selected as the last scheme of scholarship assignment. Dynamic programming helps universities in its strategic management for whole university or departments

TABLE III. RESULTS OF THE IDEAL(HUANG ET AL. 2018)

ID	1	2	3	4	5
Amount of scholarship	8200	6560	9840	11480	4920

TABLE IV. DERIVED EFFECTIVE SCHEMES OF ASSIGNMENT OF SCHOLARSHIP (HUANG ET AL. 2018)

	ID	1	2	3	4	5
	Score	25	20	30	35	15
Scheme 1	Amount of scholarship	8000	5000	10000	13000	5000
	Type of scholarship	1	2	3	1,2	2
Scheme 2	Amount of scholarship	8000	5000	13000	15000	0
	Type of scholarship	1	2	1,2	1,3	/
Scheme 3	Amount of scholarship	5000	5000	13000	18000	0
	Type of scholarship	2	2	1,2	1,3	/
Scheme 4	Amount of scholarship	5000	5000	13000	18000	0
	Type of scholarship	2	/	1,2	1,2,3	/

TABLE V. GINI COEFFICIENTS FOR EVERY SCHEME (HUANG ET AL. 2018)

Gini coeffient	0.334	0.351	0.475	0.540
Scheme	1	2	3	4

inside it through multiple knapsacks. It popularizes the basic knapsack problem (KP) from a singular knapsack to many knapsacks with (probably) various capacities. The goal is to allocate each element to at most one of the knapsacks so that none of the capacity restrictions are assaulted and to maximize the overall profit of the elements placed in the backpacks [18]. When the university develops her strategic administration to plan and apply optimization procedures, this enables the university to be ranked in the education list [19]. In 2019, Dorota et al. in [19] proposed a model which is quantitative for university managing strategy as an entire or university section. This study's methodology is gathering qualitative methods utilized in the administration strategy and the quantitative multiple knapsacks.

This study covers only the building model and decision-making operation related to administration in small Polish universities but not yet implemented. This gives us the attraction standards of the department chosen by students.

C. The Longest Common Subsequence (LCS)

Finding the longest subsequence in the same order in the two sequences given,is consider the goal of LCS. This means determining the most extended sequence obtained by removing some items from the first original sequence and removing other items from the second original sequence. This Problem also has a different application in text editing and data compression [20].

Among the main tasks of e-learning is the customization of learning. As mentioned earlier, the interest of numerous researchers has been drawn to the customization of learning objects (LOs) due to the immense significance of the LO in the learning process [21]. LCS contributed to the adaption of learning through personalizes LOs for learners. Mahdi and Fattaneh in [21] see Students and educational resources are the two key learning entities, so educational programs should give both these two entities high priority. They have developed a robust algorithm which is genetic for educational slides, thus adapting as a particular material in the e-learning resources. This algorithm potentially implemented in each type of LO. Initially, a default LO series will be assigned to the learners. This series is diagram of LOs as node or vertex and relationship among them as an edge. This hypothetical sequence is defined with the aid of a skilled and then required from the learners to rearrange these materials, and reasons will be presented for each change.

Below, the action is taken to determine the rearranging of the sequence of LOs and create a new series of them.

- Insert a new LO or deleting from the LOs sequence default (stretched to add or remove a sequence of LOs).
- To interchange the two LOs with each other.

TABLE VI. RESULT OF FIRST CASE(DUNN ET AL. 2011)

Groups	The sequence of commonly used learning objects						
Group 1	1	2	3	4	5	6	7
Group 2	1	2	3	4	5	6	7

TABLE VII. RESULT OF SECOND CASE(DUNN ET AL. 2011)

Groups	The Frequently Occurring Sequence of Learning Objects						
Group	1	3	2	4	5	6	7

This study introduced a diversity of relationships among LOs in the following.

First Case: (independent relationship): Since the LCS considers a sequence in the default sequence that is not necessarily sequential; the sequences suggested by students can be placed in groups with a specific order. For instance, as we see in Fig. 5, the LOs default sequence is displayed, and independent relationship and prerequisite are indicated in it; two LOs (4 and 5) are not linked, and for of them can be educated after LO 3 and if finished from 4 and 5, LO 6 can be taught [21] If we assume that the sequences are as follows for two sets of students (Table VI): As shown in Table VI, a general pattern is found in both groups with a hypothetical LO sequence, and the variance is on the independents (LO 4 and LO 5); therefore, it is possible to examine the reactions of the groups on their LO sequences, and after the educational experts, the endorsement chooses a better sequence. It might alternatively be conveyed that both sequences are right, and any of the sequences can be used in developing the preferred method. For instance, the best sequence may be chosen (a more similar group) after evaluating the new individual's similarity with each group. Then it sequence suggested to the person. Any calculation employed to determine the sum of similarity between entities and the groups can be a measure of similarity [21].

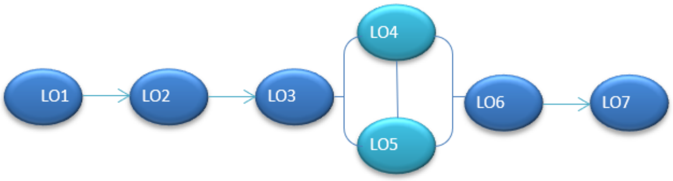


Fig. 5. The Green Line Shows the Independent Relation and the Blue Line Shows the Precedence Relation [21].

Second Case (Precedence relationship): Let's assume the above example and that one group has determined the prevalent sequence as follows (Table VII): Fig. 5 shows that LO 2 is a precondition for the LO3, but the group in Table VII does not reveal this prerequisite relationship. This may be because the group has been unable to find the group has established the prerequisite relationship among LOs or the fake prerequisite relationship among LOs [21].

Third Case (Container relationship): if there are some LOs, everyone has much smaller LOs. The largest LOs have the required connection, but the interior ones are independent. In Fig. 6, For instance, the First and last LOs are selected. As shown in Fig. 6, the lesson's three primary topics are (Part 1, Part 2, and Part 3). Many sequences of LOs in each

TABLE VIII. RESULT OF THIRD CASE (DUNN ET AL. 2011)

Groups	Proposed sequence							
Group 1	S 11	S 12	S 14	S 13	S 31	S 33	S 34	S 32

subject are not linked to each other, but the following part is expected to be articulated before starting. Both classes did not completely recognize the sequences in the following table. We were unable to find any LCS from 21, 22, 23, and 24 in the first category, for example, due to the fact that its members are suggested sequences which are sufficiently different [21]. The proposed sequence groups are shown in Table VIII and the final result of case 3 displayed in Table IX. Case 4 (relationship unknown): The effect of dissimilar granularity levels of LOs is investigated. Most popular strategies teachers apply to prepare instructional material is to choose many sections from distinct marital learning such as leaflets, handbooks, or books. Because each of these various tools explores the topic from different viewpoints, there are often no standard sections between selected parts; it is a crucial and time-consuming job to design a series of these instructional materials [21] as shown in Fig. 7. To overcome this problem, we suggest to the student to be able to offer or suggest a sequence for these sections (chapters), next we created the sets based on this proposition. For each set, we need another sequence for each chapter's parts to be suggested by its member (such as case 3), and novel groups are created in ones that already in the past stage [21].

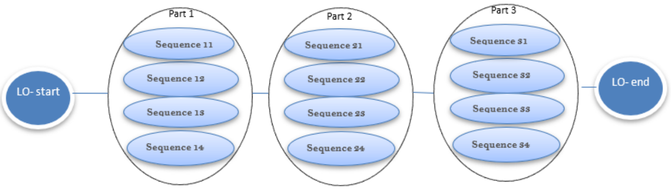


Fig. 6. The Relationship to the Container is Indicated by the Grey Circle [21].

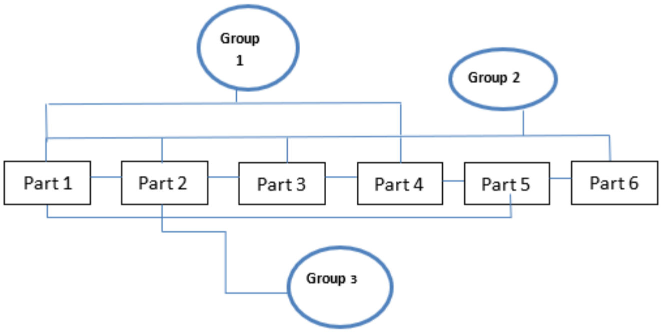


Fig. 7. Extended third Case[21].

In this study, LCS helps define the fitness function; this function for each set is equal to how much the organ is similar to others in that set. This propinquity gains from computing the LCS of student suggested sequence for LOs with other groups' organs. The findings suggest that this algorithm is efficient, and its findings could be used in different learning activities [21].

Another application of the dynamic programming using the LCS is to detect the plagiarism from a specific text or document, this method was proposed by Mahalakshmi S. et al. [22] in which they presented a system based on a symbolic execution next to the reasoning of the weakest requirement in order to find the distinction of paths and to grasp the semantics of execution paths. Indeed, the proposed method uses the LCS to provide a line-by-line comparison between the text-based search and the string by passing the file that is the object of this comparison through the following steps, starting with reading and converting the file, and then finding the shortest path between the data in the file and the data in the repository. Finally, extract the similarity report using the LCS algorithm based on recursivity. The proposed method proved to be very efficient compared to the existing methods.

Alexey Sorokin in [23] has applied the LCS method differently to predict the word forms. In fact, LCS is used to extract summary paradigms from input pairs, inflected word forms and suffix/prefix characteristics to automatically predict the paradigm. The proposed algorithm is based on a mixture of affix characteristics and ngram character models, which significantly improves performance, especially for language processing. In [23], A. Sorokin has employed several tasks to perform his method starting by the encoding stage which is consist in encode the input sequence with some label and trying to predict these labels using machine learning algorithms to convert the input into paradigm abstract. Then, a finite automata is used to extract the LCS. Furthermore, they perform the automatic inflection since the proposed method includes in the feature set the lemma suffixes and the prefixes. Finally, they use the ngram score enhancement the model score.

The LSC was also used to identify the sentence parallelism in student essays, W.Song et al. [24] proposed a robust method to automatically detect and identify the sentence parallelism using machine learning algorithms with the help of LSC. In their work, they build an annotated essay dataset. Furthermore, they construct a feature vector by using the strategies of the combination of generalized word alignment and the alignment measures between word sequences. W.Song et al. [24] started by collecting a dataset of 500 essays made by students, then they annotated the dataset to make the classification and to generate the classification report. The features extracted from the annotated dataset will be fed to a machine-learning algorithm to generate a prediction model. They made a study on various alignment measures, because the parallelism is quite similar to the alignment, to assess the alignment of sentences. They employed a several strategies to generalize the word alignment based on two features which are the semantic and the syntactic properties. Interactions of alignment metrics with word alignment strategies produce the features to show the alignment between sentences. Furthermore, The LCS is used to solve the problem of the parallel sentences by the use of dynamic programming. To assess the efficiency of the suggested method, they tested several algorithms on the basis of dynamic programming, and they reported that they achieve an f-score up to 82% in the best cases.

An interesting application of the LCS was proposed in [25] by A.Flores-Mendez et al. [25], they proposed a new method to verify the dynamic signature by the use of LCS and genetic algorithm. In their approach, they based in two major steps :

TABLE IX. FINAL RESULT OF THIRD CASE(DUNN ET AL. 2011)

Effective groups	1	1	1	1	2	2	2	2	1,2	1,2	1,2	1,2
Result	S 31	S 33	S 34	S 32	S 22	S 21	S 24	S 23	S 11	S 12	S 13 / S 14	S 14 / S 13

learning and verification by the use of the LCS. The learning process is made by solving a polynomial equation and by measuring the distance between the current signature and a set of data to verify if they are similar or not. In the other hand, the verification process was about calculating the shortest path to be assigned with a very similar signature based on the learning stage. Furthermore, the genetic algorithm was used to build a knowledge model to make the process inspired by the natural mechanism. The authors reported that the accuracy of the proposed method was outstanding, and it achieves more than 90%.

Multiobjective Dynamic Programming (MODP), which based basically on traditional DP, is advanced as a mechanism for resolving issues (problems) that involve incompatible goals that obey to features of DP. Significant progress in multiobjective dynamic optimization has been made in the last two decades. MODP considers One of the most provocative topics, is A normal expansion of DP is its utilized in conjunction with sets of fuzzy [26]. one of the applications of mode is resource allocation, this also can help e-learning systems to allocate its resource [26].

IV. COMPARATIVE ANALYSIS

DP is considered one of the methods that contribute to improve e-learning through different problems. The DP technics are used in many fields to solve several issues in different case of studies. In the present paper, we discuss three major issues (shortest path, knapsack and LCS) that help improve the learning sector. Based on the studies that we have discussed before, we find that the LCS is used generally to perform a classification process. Indeed, the LCS technique allows finding the nearest samples by finding the shortest paths based on the semantic, which is different from the shortest path, which is based only on the distance. In the other way, the knapsack is used to optimize the choice of the sample; thus, it employed always to maximize the gain. Table X summarizes the studies introduced above.

V. CONCLUSION

Dynamic programming is a method of optimization depends on the precept of optimality realize by Richard Bellman. It is considered one of the solutions and methods that have contributed to solving many problems in various fields, especially in e-learning. This survey explained how DP contributed to supporting education through three types of problems (shortest path, knapsack, and long common sequence), where during the shortest path problems DP supports adaptive e-learning through generating the suitable path of different learners with different learning style and provide the appropriate learning materials as we shown in first three studies. knapsack problems contribute in supporting higher education through the scholarship assignments and assist to select the suitable learners to gain scholarship in a fair way as introduced in fourth study and knapsack problems aid universities in strategic plan and administration like in fifth study. Last studies illustrates the role of LSC problem to detect the most suitable sequence of Learning Objects from the ones introduced by the learners to personalized the learning materials in effective way and this enhancing adaptivity in e-learning. Also, LCS contribute in some application that help learners such as plagiarism detection and verifying the dynamic signature. All studies have proven how DP has contributed to improving and supporting e-learning.

REFERENCES

[1] A.-P. Pavel, A. Fruth, and M.-N. Neacsu, "ICT and E-Learning – Catalysts for Innovation and Quality in Higher Education," *Procedia Economics and Finance*, vol. 23, no. October 2014, pp. 704–711, 2015.

[2] A. O. Wong and K. Sixl-Daniell, "The Importance of e-Learning as a Teaching and Learning Approach in Emerging Markets," *International Journal of Advanced Corporate Learning (iJAC)*, vol. 10, no. 1, p. 45, 2017.

[3] R. Radha, K. Mahalakshmi, V. S. Kumar, and A. R. Saravanakumar, "E-Learning during Lockdown of Covid-19 Pandemic: A Global Perspective," *International Journal of Control and Automation*, vol. 13, no. 4, pp. 1088–1099, 2020.

[4] B. Render, R. M. Stair, and M. E. Hanna, *Quantitative Analysis For Management ELEVENTH EDITION*, 2012.

[5] N. Voloch, "Optimal paths of knapsack-set vertices on a weight-independent graph 2 Problem Formulation- The knapsack weight-independent graph," vol. 16, pp. 163–171, 2017.

[6] I. Abraham, A. Fiat, A. V. Goldberg, and R. F. Werneck, "Highway Dimension, Shortest Paths, and Provably Ecient Algorithms."

[7] P. Frazier, "Learning with Dynamic Programming," 2011.

[8] R. G. Selvanathan, "A Dynamic Programming Approach to Identifying the Shortest Path in Virtual Learning Environments," *Journal of Management and Sustainability*, vol. 5, no. 1, pp. 89–96, 2015.

[9] A. Rivas, A. González-Briones, G. Hernández, J. Prieto, and P. Chamoso, "Artificial neural network analysis of the academic performance of students in virtual learning environments," *Neurocomputing*, no. xxxx, 2020.

[10] H. Fazlollahtabar, "Reliability-based dynamic programming for E-learning user profile assessment," *International Journal of Information and Communication Technology Education*, vol. 8, no. 3, pp. 13–21, 2012.

[11] N. Alqahtani, M. Kamel, and M. Saleh, "Enhancing the Adaptive E-learning Environment by using the Markov Decision Process (MDP)," vol. 05, no. 9, pp. 43–46, 2018.

[12] M. N. Moghadasi, A. T. Haghighat, and S. S. Ghidary, "Evaluating Markov decision process as a model for decision making under uncertainty environment," *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, vol. 5, no. August, pp. 2446–2450, 2007.

[13] D. Gries and F. B. Schneider, *Programming Challenges: The Programming Contest Training Manual*, 2011, vol. 34, no. 2.

[14] G. Durand, F. Laplante, and R. Kop, "A Learning Design Recommendation System Based on Markov Decision Processes A Learning Design Recommendation System Based on Markov Decision Processes," 2011.

[15] N. Z. Atashbar and F. Rahimi, "Optimization of Educational Systems Using Knapsack Problem," *International Journal of Machine Learning and Computing*, vol. 2, no. 5, pp. 552–555, 2011.

[16] D. Huang, Y. Gu, H. Wang, Z. Liu, and J. Chen, "An Incentive Dynamic Programming Method for the Optimization of Scholarship Assignment," *Discrete Dynamics in Nature and Society*, vol. 2018, 2018.

TABLE X. COMPARATIVE ANALYSIS OF SURVEY

Study No.	**authors**	**Problem statement**	**Method**	**Result**
1	HAMED FAZLOLLAHTABAR 2008	E-Learning has mainly been embraced as a favorable resolution by multiple communities to provide opportunities of learning to personal learners to reduce training time and cost.	Applying DP to detect the shortest route in the education environment for learners and use the approach of analytical hierarchy process to transfer the parameters of qualitative to quantitative ones	An ideal learning path was generated with the help of dynamic programming and parameters quantitative conversion of qualitative parameters
2	Hamed Fazlollahtabar 2012	From the reliability of each direction, choose the best path aligned with the concern of the student. Of course, depending on the decision maker's preference, other requirements may also be introduced.	It proposes an evaluation technique that uses a DP approach to model an ideal user profile route and use reliability to calculate user interconnections in an e-learning network.	The system helped the user find the optimum course and approved that during the illustrated example.
3	Norah Alqahtani , Mahmod Kamel, Mostafa Saleh 2018	There are individual differences for learners, and each learner needs to get educational materials appropriate to his learning style.	Develop adaptivity in the education process by using the Markov Decision Process (MDP) with dynamic programming (shortest path problem).	With the help of DP, MDP contributes to generating the appropriate e-learning path for different learners with different learning styles.
4	Hans Wang Yu Gu ,Jun Chen , Zhiyuan Liu , and Di Huang 2018	The traditional way to scholarship assignments consumes much time and effort in application operations conducted by students to gain university scholarship.	The repeatedly solving a chain of knapsack subproblems and altering the monetary a unit scores value. The ideal deployment plan can then be eliminated by performing the coefficient of Gini for quantifying the equity of each proper schema.	To clarify the applicability of the suggested procedure, an illustrative example is acknowledged in this study, and the result shows the proposed method addresses the fairness demand that pupils who do better should receive awards higher than those who do worse.
5	Dorota Kuchta Radoslaw Rynca 2019	Several approaches in the literature may be beneficial in university strategic administration. Some of them are linked to the sustainability element in balancing the grade of complishment of multiple, often overlapping, goals that need to be taken into account when constructing strategies. Such approaches include portfolio product/service ones. Their utilized, however, is mostly axiomatic and disconnected from management's quantitative sides. This study want to provide a proposal to alter the portfolio methods.	Gathering among non-quantitative portfolio approaches with the quantitative multiple knapsack dilemma used in strategic management.	For the first time in the literature, a connection among portfolio procedures and a quantitative improvement approach for university administration objectives was introduced in this study. For all manner of educational institutions, portfolio approaches seem to be theoretically useful. The chosen criteria make it possible to identify different items (departments, colleges, services, topics, etc.) and imagine the present status and plan enhancement. The entire strategy makes it potential to simulate different methods for the learning institution's growth, and take the ideal decision.
6	M. Mahdi Barati, Fattaneh Taghiyareh 2013	finding the most suitable sequence of LOs from the ones proposed by the students to personalized the learning materials in effective way	genetic-based algorithm with help LCS for adapting the learning slides as a significant learning material in e-learning environments	The needed data has been gathered with questionnaires from thirty-six learners that register in the course of Bio Computing. The results analysis displays that the suggested model has been so effective in adapting the slides of educational.
7	Mahalakshmi S, and Kavitha S. 2016	The suggested model aims to discover the similarity and the plagiarism between the input document/text and the data in the repository	The authors apply an LCS method to find the high similar text by using five steps process to the input document,in which they employ the LCS as a recursive function.	The proposed method outperformed the existing algorithm in the literature, which are based on a simple distance calculation.
8	Alexey Sorokin 2016	The proposed method aims to predict the words form by transforming the input data into labels and extracting the lemma suffixes and the prefixes to construct a feature set. Finally, a classifier is performed to predict the form.	The authors apply a finite automata is used to extract the LCS. Then, they perform the automatic inflection to construct the final feature set to be subject for a classifier to find and foretell the words form	The method has been proposed shows a high accuracy since it uses additional information which is the ngram score enhancement the model score.
9	Song et al. 2016	Automatically detect and identify the sentence parallelism using machine learning algorithms with the help of LSC	They employed a several strategies to generalize the word alignment based on two features which are the semantic and the syntactic properties. Furthermore, The LCS is used to solve the problem of the parallel sentences by the use of dynamic programming.	they tested several algorithms based on dynamic programming, and they reported that they achieve an f-score up to 82% in the best cases.
10	Flores-Mendez and Bernal-Urbina 2010	An automatic way to verify the dynamic signature by the use of LCS and genetic algorithm	They use the LCS combined with the genetic algorithm to verify and classify the dynamic signatures	The authors reported that the accuracy of the proposed method was outstanding, and it achieves more than 90%.

[17] S. L. DesJardins and B. P. McCall, "The impact of the Gates Millennium Scholars Program on college and post-college related choices of high ability, low-income minority students," *Economics of Education Review*, vol. 38, pp. 124–138, 2014.

[18] H. Kellerer, U. Pferschy, and D. Pisinger, "Knapsack problems," 2004.

[19] D. Kuchta, R. Rynca, D. Skorupka, and A. Duchaczek, "The use of the multiple knapsack problem in strategic management of a private Polish university: Case study," *International Journal of Educational Management*, vol. 33, no. 2, pp. 335–358, 2019.

[20] F. S. Pribadi, A. E. Permanasari, and T. B. Adji, "Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS)," *Education and Information Technologies*, vol. 23, no. 6, pp. 2855–2866, 2018.

[21] A. M. Dunn, O. S. Hofmann, B. Waters, and E. Witchel, "A Longest Common Subsequence based Genetic Algorithm for Courseware Design," pp. 395–410, 2011.

[22] Mahalakshmi S and Kavitha S, "Software Program Plagiarism Detection Using Longest Common Subsequence Method," *International Journal of Computer Techniques* -–, vol. 3, no. 4, pp. 35–40, 2016.

[23] A. Sorokin, "Using longest common subsequence and character models to predict word forms," no. 2002, pp. 54–61, 2016.

[24] W. Song, T. Liu, R. Fu, L. Liu, H. Wang, and T. Liu, "Learning to identify sentence parallelism in student essays," *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, pp. 794–803, 2016.

[25] A. Flores-Mendez and M. Bernal-Urbina, "Dynamic signature verification through the longest common subsequence problem and genetic algorithms," *2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 IEEE Congress on Evolutionary Computation, CEC 2010*, 2010.

[26] M. A. Abo-Sinna, "Multiple objective (fuzzy) dynamic programming problems: A survey and some applications," *Applied Mathematics and Computation*, vol. 157, no. 3, pp. 861–888, 2004.

The Feasibility of Implementing a Secure C2C Credit Scoring Platform

Mariam Musa Al- oqabi[1], Dr. Wahid Rajeh[2]
Department of Information Technology
University of Tabuk
Faculty of Computers and Information Technology
Saudi Arabia, Tabouk

***Abstract*—The continuous development of social media and online commerce, which permeates all aspects of our lives, leads to the need for a similar mechanism similar to the financial credit score in traditional business. Also, a realistic classification of users through social media to be used in all aspects of the relationships between users and some of them or between them and organizations is needed. In this article a new metrics to classify users according to their creditworthiness of the transactions that take place through the Internet is established. The object from this aricle design a social credit system model (SCSM) based on these new metrics. How to deal on the Internet, attacking people on social media, violating the privacy of people and others. Also Buying and selling operations, executing purchase and sale orders, paying amounts of money easily and quickly, and so on. These data and their degree of importance were determined according to several questionnaires directed to several segments of society. This creditworthiness can be used in banks, Uber, Online transactions and so on.**

***Keywords*—Social media; online commerce; social credit system; creditworthiness**

I. Introduction

The increase in credit fraud in the information age has disrupted economic market activities and could cause significant damage to global economic systems. This lead researcher to study a social credit and assess credit risk by scoring each user based on objective measures.

This study is mainly done to build a SCSM model which gains importance from the need to facilitate all operations that take place over the Internet.

Social media can be defined as a Web based information system that allows users to create profiles, support communication between users, and create relationships [1]. Loans, financial transactions and credit cards need certain conditions to approve these transactions. These conditions are known nowadays as people's creditworthiness. Companies and banks rely on home address, income, job data and so on to determine the creditworthiness of people. Also, the way to pay its past payments will greatly affect its creditworthiness, which allows or does not allow it to take a loan or get a credit card. Traditional data in our time is insufficient to build creditworthiness, Using social data is one such option [2]. It is needed in all online transactions. The individual's profile data on social networks such as Facebook, largely reflects the real data of the person who can be used to build creditworthiness [3]. Building a social credit system based on these ideas can improve all Traditional and online financial transactions. This system will add confidence to all online processes. There is various example that may benefit from social credit scoring, for instances you can accept a friend on Facebook based on his social credit. There are multi ways people interact with the consumer credit system, more specifically, the way they access credit and the way they are held accountable for their debt [4]. Microfinance has been taken care of and how to ensure credit evaluation. In addition to traditional credit rating methods such as sociodemographic and credit data, other data from Facebook (like group, friends, and so on) has been extracted and used in a new model to automate the credit scoring process for microfinance [5]. The impact of using metrics based on social networks has been analyzed to give a score to customers is illustrated in [6]. Goel and Gold stein in [7], discussed using the available data through social media to show people's trends in purchasing, clicking on ads and registering in different events. They rely on friend-to-friend relations to make the decision and categorize people. This relationship on social media predicts a user's behavior by tracking a friend's behavior in purchases, clicking and any online transaction. The Chinese system of social credit has been designed to target many fields [8], [9]. The social credit scores are routinely incorporated to include most information about the people, so it's affect on get job, even dating or marriage, it become active to structure the people's life [10], [11]. These are aimed at judicial and other credibility in the fight against fraud. There is also another goal of building social community loyalty to increase trust among individuals. Social credit information is collected by specific parties, ranging from administrative sanctions, payments, volunteering, web searches, social media comments, GPS tracking data and most online transaction data. This system expands the information collected and the goals in which this system is used. This data is comprehensive and not important to our primary goal of building creditworthiness for online transactions. Researchers in [12], relied on a collection of social media data collected through the most popular Chinese website Weibo. This data mainly includes demographics, tweets, and people-to-people relationships. They apply some machine learning techniques to predict the credit score of a person. In [13], how to rely on people-to-people relationships on social networks in the distribution of new products has been studied. The point was also made that with more information on social networks, the marketing process would be accurate. Infer the degree of credit for people from their online social data studied at [14]. A model for learning about online user behavior, especially via Twitter, has been suggested as illustrated in Fig. 1. In it, the tweets are analyzed and the words discovered to classify people credit. For example, as described in the figure, the same drunken habits and watching football matches are associated with sending late-night tweets. Accordingly, the system can predict certain behavioral habits of users, and on the basis of these habits can be credited through social media. In Research [15], [16], work is done to identify the most influential social media users in the activities of others in order to draw greater advertising and administrative attention to them. In 2014-2020 the Chinese government intend to build social credit system (SCS) to better collect and evaluate citizens' creditworthiness, and grant rewards and punishments based on one's social credit. They built the SCS depend on collect information from national development and perform commission, Supreme People's Court, other Central Governmental Agencies, and the other Government Agencies Bank of China and others benefit from this information to judgment on the people from are this people in redlist or discredited judgment debtor [17]. China's Internet market was examined in [18], [19], where social-media data were used to describe borrowers' behavior and how credit rating systems can influence social behavior in china's social credit systems, it's research in concepts of an ideal citizen, the state and associated private enterprises can reinforce trust within society by issuing social credit scores to everyone, so everyone will likely aim to achieve a high social credit score, but the government state need to consider the implications of their policies and consider avenues that individual for low social credit scores could turn towards for assistance.

This article was based on the development of multiple models to compare the accuracy of customer scores obtained based on social networks or without. It relied on the efforts of lawyers and data scientists who extract credit worthiness from the big data available online in most social network sites. Legal problems have also been analyzed and solutions are found for all these problems to ensure transparency and trust. From the previous, we noted a lack of processing some concepts that must be dealt with to produce a robust system that can reflect an individual's true creditworthiness. Violating the privacy of people and other social media events and linking social media to e-commerce sites to assess people's transactions is our goal. In addition, a model will be built in which the type of data is determined and how it is collected, analyzed, and used to build the creditworthiness. Also, this article different from others, specific data will not be analyzed for creditworthiness, but will be designed a model adjust what data can be collected or added and how it will be used to accurately grant people credit. This model collect data from multi sources such as social networks, online sales and purchases data about the individual, in order to classify individuals in terms of creditworthiness.

This article organizes as follows: In section introduction, we talked about the literature review and display the difference between this article and others and the features this article has. The methodology used in collect the data are discussed in section Methodology. The analysis the questionnaire content and get the results are explained in results and discussion section. In section design SCSM, talked about the design the model. In section conclusion, displayed the conclusion we got from this article.

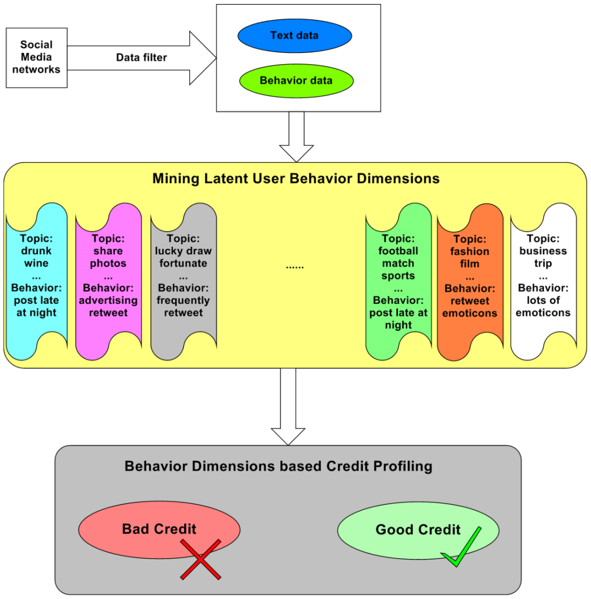

Social Media networks
Data filter
Text data
Behavior data
Mining Latent User Behavior Dimensions
Topic: drunk wine ... Behavior: post late at night
Topic: share photos ... Behavior: advertising retweet
Topic: lucky draw fortunate ... Behavior: frequently retweet
......
Topic: football match sports ... Behavior: post late at night
Topic: fashion film ... Behavior: retweet emoticons
Topic: business trip ... Behavior: lots of emoticons
Behavior Dimensions based Credit Profiling
Bad Credit
Good Credit

Fig. 1. Credit Score Based on Social Data.

II. Methodology

This article depend on the questionnaire. So, in order to gather needed information from the society, two-part questionnaire is designed using Google Docs and URL. The questionnaire was distributed online by WhatsApp, Instagram, and Telegram social media users. The questionnaire distributed to a number of specialists on banking, service providers and other peoples. It will be taken into consideration that the subscribers familiar with regulations, laws and banks and a fair distribution in terms of age and gender. The questionnaires designed to three groups:

1) The first for ordinary internet and social users.
2) The second for companies and service providers via the Internet or in traditional ways.
3) Finally, for those familiar with the work of banks.

By using Likert questionnaires, the data can be collected quickly and this data can also be interpreted and relied on it [20]. Here this type of measure will be used to collect the exact respondent's opinion on specific points. It allows us to build our model accurately. The advantages of the Likert scale questions in surveys is that:

- It is an accurate way to collect data, through which you can analyze and understand data well.
- It can be used with quantitative data.
- It is easy to conduct it as it does not take time.
- Multiple options, allowing accurate answers from respondents.

In this article a 6-point measures will be used as illustrated in table I, as they increase the accuracy of the answer. The scale combines answers that vary from weaker (Strongly disagree) to strongly agree support. To increase the accuracy of the questionnaire results, this scale does not have a neutral answer

to obtain weighted answers from all participants. Another type of question (yes / no) is used in this article to specifically measure the participants' opinion on some issues related to financial transactions over the Internet and the legality of tracking users' information over the Internet. The volunteers were requested to answer the question are listed in Appendix A.
total of 600 respondents answered the survey questions. In this dataset, rows represent different type of events and column different criteria for analysis. Each data record represents the value of a particular analysis criterion for a type of event. Based on the data obtained from the survey, a dataset are organized using excel tool. The next step was sampling the data in order to prepare them for further analysis. Based on the results of this questionnaires and the previous literature review, a SCSM will be designed.

TABLE I. A Sample 6-point Likert Scale

	likert scale	
Srtongly Disagree	Disagree	Slightly Disagree
Slightly Agree	Agree	Strongly Agree

III. Results and Discussion

This section contains the results and discussion of all questionnaire.

1) Results and discussion of the first questionnaire: Table II, explains responses obtained from respondents; the percentage response was calculated in line with the 6-likert scale.

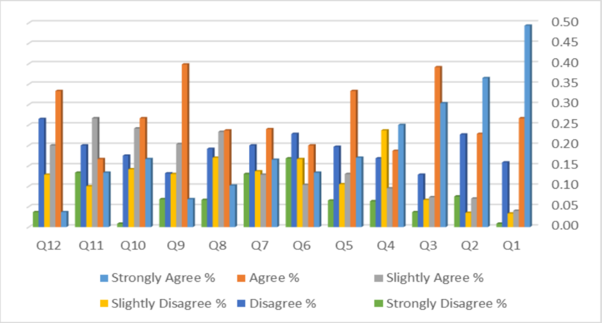


Fig. 2. Percentage Results of Internet and Social Users Questionnaire Likert Part.

As shown in fig. 2, for the first question, a large percentage of people agreed that they use the Internet a lot, where the percentage of Strongly Agree was 49% for this question. The overall approval rate was approximately 80%, which supports us in providing the proposed model for classify internet users. Also, even if it was less, the respondents agreed in the second question that they use social media frequently a lot, which will support our proposed model based on social media for extracting data on individuals. The results for the third question are Strongly Agree 30%, Agree 39%, Slightly Agree 7%, Slightly Disagree 7%, Disagree 13%, Strongly Disagree 4%. Regarding this question, the high approval rate indicates the future of Internet commerce and that it will expand rapidly and continuously. Participants have been responding to the fourth question with percentage Strongly Agree 25%, Agree 19%, Slightly Agree 10%. These ratios make us predict that the information available from users' interactions through the Internet will be rich and sufficient to build our model. The approval rate (63%) for the fifth question opens up different methods for us to obtain accurate data on individuals to be accurately evaluated through our proposed model. Since the use of such applications as the Uber application provides us with more information on the behavior of individuals. The sixth and seventh questions are concerned with studying the degree of problems that individuals face in buying and selling through the Internet. The percentage of problems encountered when selling transaction by individuals on the Internet is higher than in purchasing transaction, and this mainly results from the lack of experience of individuals in selling via the Internet. Generally, these questions indicate the existence of actual problems when applying transaction via the Internet: Strongly Agree 13%, Agree 20%, Slightly Agree 10% for question six and Strongly Agree 17%, Agree 24%, Slightly Agree 13% for the seventh question.

Questions from 8 to 10 are concerned with studying the existence of verbal and psychological attacks, privacy violation, and data theft. The rate of expectation of a privacy violation problem is higher than the other two problems, where the overall approval rate is 61%. Respondents' answers for these three questions support the need for the model presented in this article.

The eleventh question measure the prevalence of banking transactions that take place over the Internet. Of course, there is a percentage that has a weight that does not use these services, due to fear of problems with online transactions. Finally, the twelfth question is one of the most important questions, as it measures the opinions of the participants regarding the possibility of using the transactions that take place through the Internet and social media in building a creditworthiness system similar to that found in banks.

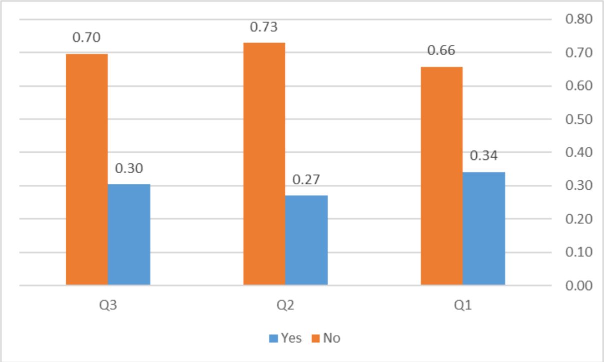


Fig. 3. Percentage Results of Internet and Social Users Questionnaire Yes/No Part.

Fig. 3, illustrate the three (Yes / No) questions that display Internet users in general. Table III, contains the respondent's answers. it given to users to measuring their opinion specifically about the problems they face when selling (such as eBay) or buying. It is noticeable in Fig. 3, that the product quality problems were approved by a number of subscribers, close to 34%. Also, the problems of buying and selling via the

TABLE II. RESULTS OF INTERNET AND SOCIAL USERS QUESTIONNAIRE LIKERT PART

Questions	Strongly Agree	Agree	Slightly Agree	Slightly Disagree	Disagree	Strongly Disagree	Total Response	Results
Using Internet	296	160	24	20	95	5	600	Accepted
Using Social Media	219	137	42	21	136	45	600	accepted
Buying Product	182	235	44	40	77	22	600	accepted
Purchases and Sales	150	112	57	142	101	38	600	Accepted
Using Uber	102	200	78	63	118	39	600	Accepted
Problem During Sales	80	120	62	100	137	101	600	Not Accepted
Problem During Buying	99	144	77	82	120	78	600	Accepted
Verbal and Psychological Attacks	61	142	140	102	115	40	600	Accepted
Violating Privacy	41	239	122	78	79	41	600	Accepted
Stolen Personal Data	100	160	145	85	105	5	600	Accepted
Banking Transactions Online	80	100	160	60	120	80	600	Accepted
Extract Information for Creditworthiness	22	200	120	77	159	22	600	Accepted

TABLE III. INTERNET AND SOCIAL USERS QUESTIONNAIRE – YES/NO PART

Questions	Yes	No	Total Response
Quality Problem	205	395	600
Financial Problems at Selling	162	438	600
Financial Problems at Buying	182	418	600

Internet had support, but less than the quality problems. These problems are mainly found because there is no real evaluation that reflects users' behavior online.

An important question was asked to Participants about how to pay the value of goods purchased over the Internet. The classification of the answers is shown in Table IV, Also, Fig. 4, shows the percentage of each choice. As shown, payment on receipt is the predominant feature with a good percentage that uses the visa in its online purchasing transactions. This result is due to the fear of a large percentage of participants from using their credit card over the Internet. Also, this supports our need for the proposed model in our article.

TABLE IV. RESULTS OF INTERNET AND SOCIAL USERS QUESTIONNAIRE CHOOSE PART

Question ID	Visa	when receiving	Total Response
Payment Method	12	18	30

2) Results and discussion of the second questionnaire: The questions directed to companies, that apply their transaction through the Internet, measure the extent of the difficulties they face, whether there is a need to make rating of Internet users and what are the basic data that can be used in this rating process. Table V, contains the answers of 30 participants in this questionnaire.

As shown in Fig. 4, the first question measures the extent to which the company uses the Internet to apply their commercial transactions. Answers show that approval responses slightly outweigh disagreement. This is a result of the fear of conducting these transactions, as there is no accurate rating of the person you are dealing with. The answer to the third question may be related to the first question. As is the case with a number of participants who are not using online business transactions, they are likely to see no income from the Internet.

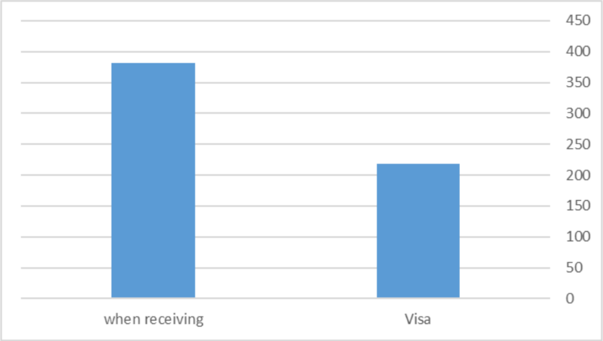


Fig. 4. Percentage Results of Internet and Social Users Questionnaire Choose Part.

Returning to the second question, the respondents 'answers fluctuate between approval and disapproval (Agree 30%, Slightly Agree 20%, Slightly Disagree 20%, Disagree 30%) Equally. This indicates that there are multiple problems in dealing with suppliers via the Internet for companies. This encourages us the importance of having a model that manages the rating process for all users over the Internet.

Questions 4 to 6 are concerned with measuring the extent of problems caused by online buyers. Such as users not completing their orders, refusing to receive product, and using the product return process a lot. The percentage of those who agree to the fourth question is high, and this is due to a large number of individuals who do not complete their requests for various reasons. This behavior does not significantly affect the design of our model, but put a little weight on it in the rating of individuals. As for the fifth and sixth question (refusal to receive, return of products), even with a low percentage in approval of them, it indicates its importance within the proposed model. As these two process represent a major inconvenience to companies.

For questions 7 to 9, it measures the percentage of respondents' approval or disapproval, gradually from Strongly Agree% to Strongly Disagree%. These questions measure the quality of data exchanged over the Internet and its relevance in building a model that represents a user's creditworthiness.

The high approval rate for the ninth question indicates that the majority of respondents agree that the user's credit worthiness based on the study of his behavior over the Internet increases individuals' confidence in completing their electronic transaction.

with respect to the yes/no questions, as explained in Fig. 5, the first and second questions are related to the companies' dealings with suppliers. The responses of the participants confirmed that there are problems with the quality of the products received, as well as financial problems with the suppliers, but in a not large way as it is equal to 27% of the total percentage. Also, the third question is related to financial problems with corporate buyers, and the percentage has risen slightly to 30%. I think even with low percentages that support the existence of actual problems for companies, the existence of a rating for users through internet will contribute to reducing this percentage significantly while providing protection for these companies.

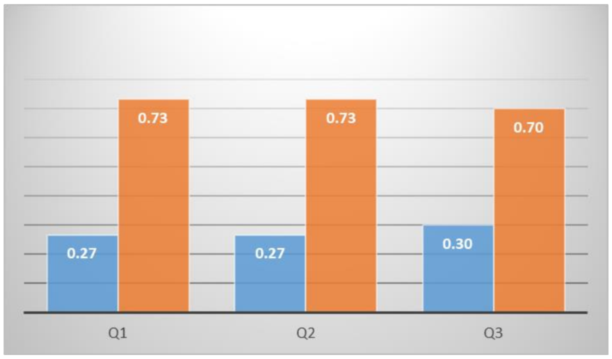


Fig. 5. Percentage Results of Companies and Online Service Providers' Questionnaire Yes/No Part.

3) Results and discussion of the third questionnaire: The third questionnaire related to banking workers or specialists in the field of regulations and laws was applied by 30 of the participants. Table VI, shows the results obtained, also Fig. 6, shows the percentages of these results for each question. As shown in the Fig. 6, the percentages of the answers to the first question are slightly tilted to refuse to track individuals 'transactions through the Internet. I believe that the percentage of approval also has weight because of the need to track individuals' transactions to prevent crimes. Using of this data was rejected more strongly in the answers to the second question from the participants, due to the sensitivity of not using this data institutionally. From observing the results, the use of this data can be permitted based on strict restrictions on its use. Finally, the answers to the third question tend to deny permission to use this data in the absence of a legal umbrella for this use.

with respect to yes/no questions, Table VII, illustrates the answers of 30 participants. Fig. 7, explain clearly the large percentages of the answer were yes. Where the three questions focus on the importance of the existence of laws and legislation to allow tracking, use and publication of individual transactions over the Internet. Naturally, this segment of the participants was encouraged to issue regulations and legislation that regulate these operations.

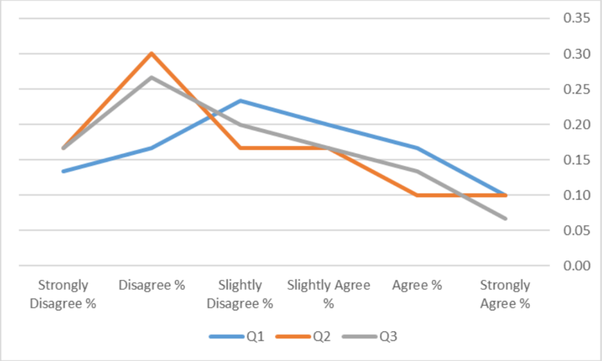


Fig. 6. Percentage Measure for Participants that are familiar with Regulations, Laws and Banking.

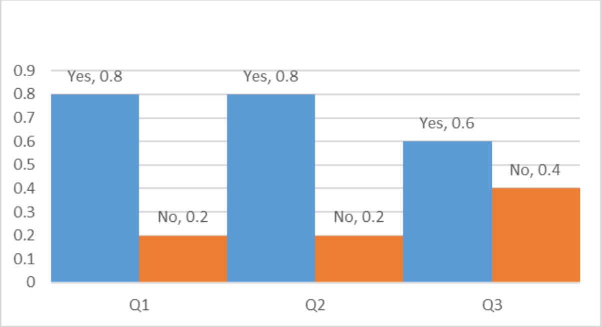


Fig. 7. Percentage Results of Those Are Familiar with Regulations, Laws and Banking Questionnaire Yes/No part.

Based on the results of the third questionnaire, it is clear that it is possible to benefit from individuals' transactions on the Internet to design the proposed model that gives a score for each individual, provided that there are legislations that support and protect the existence of this model.

Based on the results of the third questionnaire, it is possible to benefit from individuals' transactions on the Internet to design the proposed model that gives a score for each individual, provided that there are legislations that support and protect the existence of this model.

IV. Design SCSM

This model was designed based on the results and discussion of previous questionnaires. Most of the survey results support the existence of the proposed SCSM model, as the results see electronic commerce expanding at increasing rates in the future. The proposed model (SCSM) is illustrated in Fig. 8. First, with regard to our need for regulations and laws, the results of the questionnaire propose that it is better to have controls for tracking, using and disseminating individuals 'data on the Internet. Therefore, a module was added in proposed model that is responsible for monitoring and controlling the use of this data in terms of its size or who is authorized to use it. This module offers the business community a great opportunity to complete its operations with confidence and under a legal umbrella.

Based on the results of the questionnaire, there are problems in buying and selling between individuals and some

TABLE V. RESULTS OF COMPANIES AND ONLINE SERVICE PROVIDERS' QUESTIONNAIRE LIKERT PART

Questions	Strongly Agree	Agree	Slightly Agree	Slightly Disagree	Disagree	Strongly Disagree	Total Response	Results
Commercial Transactions	9	6	3	3	6	3	30	Accepted
Problem with Suppliers	0	9	6	6	9	0	30	Not Accepted
Income Increasing	3	6	15	0	6	0	30	accepted
Incomplete Transaction	6	9	6	0	6	3	30	Accepted
Order Items Refusing	3	3	3	3	6	15	0	30	Not Accepted
Product Returning	0	6	3	6	15	0	30	Not Accepted
Uber History Transaction	3	9	3	3	9	3	30	Not Accepted
Social Interaction	6	9	3	3	6	3	30	Accepted
Measure of the User's Credit Worthiness	9	10	5	0	6	0	30	Accepted

TABLE VI. RESULTS OF THOSE ARE FAMILIAR WITH REGULATIONS, LAWS AND BANKING QUESTIONNAIRE LIKERT PART

Questions	Strongly Agree	Agree	Slightly Agree	Slightly Disagree	Disagree	Strongly Disagree	Total Response	Results
Tracking of Electronic Transactions	3	5	6	7	5	4	30	Not Accepted
Use of Tracked Information	3	3	5	5	9	5	30	Not Accepted
Without Facing Litigation	2	4	5	6	8	5	30	Not Accepted

TABLE VII. INTERNET AND SOCIAL USERS QUESTIONNAIRE YES/NO PART

Questions	Yes	No	Total Response
Law to Tracking Users' Transaction	24	6	30
Law to Use Users' Information	24	6	30
Law to Publishing Users' Information	18	12	30

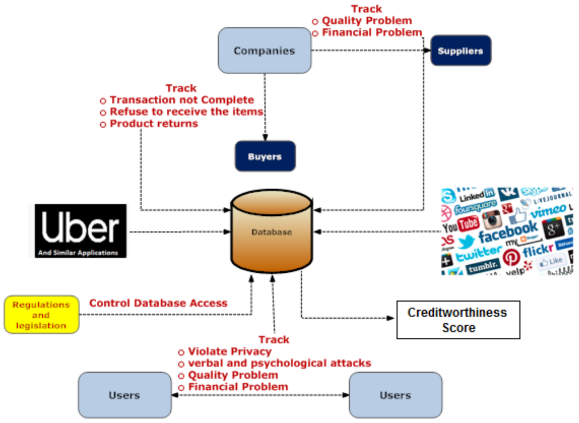


Fig. 8. Designed Model.

of them, which lose the credibility of the transaction that take place over the Internet. Therefore, this data was tracked and added to the model database in order to be used in the creditworthiness of individuals because of its weight according to the analysis of the questionnaires.

There are Another important data that needed to track and add to the model database. For example, verbal and psychological attacks, privacy violation, and data theft. This data has great weight because of its impact on accurately describing the behavior of individuals. Its presence also reduces or eliminates the fear of individuals from online transactions because it adds safety and reliability for the individuals available for transactions.

Today's social media data is very informative. Most people of different ages have multiple social media accounts. With the multiplicity and diversity of this data, our model can rely on it to give an accurate assessment of creditworthiness.

On the other hand, the use of such applications as the Uber application gives important information that reflects the behavior of individuals, whether the individual is the service provider or the beneficiary of it. Tracking this data and recording it in the form database (from discussing the survey results) is very helpful for the final decision describing the creditworthiness of each individual. Therefore, the data for these applications were added to the model as shown in Fig. 8.

V. CONCLUSION

a new concept was introduced to classify individuals in terms of creditworthiness and confidence in dealing with the Internet. This concept was developed to introduce SCSM model in order to be used to increase the reliability of transactions that take place over the Internet. The design of this model relied on collecting multiple data such as social media data, online sales and purchases data, and so on. These data and their degree of importance were determined according to several questionnaires directed to several segments of society.

The first was directed at regular users of the Internet and social media to study the extent of their confidence in transactions that take place over the Internet and what problems they face. As for the second, it was directed at the business community from companies that carry out transactions over the Internet. In this questionnaire, companies' problems with customers and suppliers are studied. Finally, there was a questionnaire directed to banking workers or specialists in the field of regulations and laws. This questionnaire was directed to study the extent of our need for laws regulating the tracking and circulation of information.

These questionnaires were analyzed and the results of the analysis were used to determine the data that is important to incorporate into the model. Finally, the SCSM was introduced.

A. Future Work

Developing the SCSM model to make it compatible with the huge development in communications and data transmission will be the focus of our attention in the future. It will be made more flexible to accept and rely on new data that arise from social media and other new services on the Internet.

References

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication, 13, 210-230*, vol. 13, pp. 210–230, 2008.

[2] Y. Wei, P. Yildirim, C. Van den Bulte, and C. Dellarocas, "Credit scoring with social network data," *Marketing Science*, vol. 35, no. 2, pp. 234–258, 2016.

[3] S. De Cnudde, J. Moeyersoms, M. Stankova, E. Tobback, V. Javaly, and D. Martens, "What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance," *Journal of the Operational Research Society*, vol. 70, no. 3, pp. 353–363, 2019.

[4] A. Rona-Tas and A. Guseva, "Consumer credit in comparative perspective," *Annual Review of Sociology*, vol. 44, pp. 55–75, 2018.

[5] S. D. Cnudde, J. Moeyersoms, MarijaStankova, E. Tobback, Vinayak-Javaly, and D. Martens, "What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance," *Journal of the Operational Research Society, 70:3, 353-363*, 2010.

[6] C. V. d. B. Yanhao Wei, Pinar Yildirim, "Credit scoring with social network data," *Marketing Science 35(2), 234–258*, 2016.

[7] S. . D. G. Goldstein, "Predicting individual behavior with social networks," *Marketing Science*, vol. 33, pp. 82–93, 2014.

[8] M. Blomberg, "The social credit system and china's rule of law," *Mapping China Journal 2: pp. 77-113*, 2018.

[9] C. Liu, "Multiple social credit systems in china," *Economic Sociology: The European Electronic Newsletter 21 (1): 22–32*, 2019.

[10] M. Fourcade and K. Healy, "Classification situations: Life-chances in the neoliberal era," *Accounting, Organizations and Society*, vol. 38, no. 8, pp. 559–572, 2013.

[11] A. Rona-Tas, "The off-label use of consumer credit ratings," *Historical Social Research/Historische Sozialforschung*, pp. 52–76, 2017.

[12] G. GUO, F. ZHU, E. CHEN, Q. LIU, L. WU, and C. GUAN, "From footprint to evidence: An exploratorystudy of mining social data for credit scoring..acm transactions on the web," *Research Collection School Of Information Systems 10, (4), 22:1-38*, 2016.

[13] M. Haenlein and B. Libai, "Targeting revenue leaders for a new product," *Journal of Marketing 77 (3), 65–80*, 2013.

[14] GuangmingGuo, F. Zhu, E. Chen, L. Wu, Q. Liu, Y. Liu, and MinghuiQiu, "Personal credit profiling via latent user behavior dimensions on social media," *In PAKDD 2016. 130–142*, 2016.

[15] B. R. Trusov M, Bodpati AV, "Determining influential users in internet social networks," *Marketing Res. 47(4): 643—658*, 2010.

[16] S. Sevignani, "Surveillance, classification, and social inequality in informational capitalism: The relevance of exploitation in the context of markets in information," *Historical Social Research/Historische Sozialforschung*, pp. 77–102, 2017.

[17] C. Liu, "Multiple social credit systems in china," *Economic Sociology: The European Electronic Newsletter*, vol. 21, pp. 22–32, 2019.

[18] M. H. . J. Adebayo, "Credit scoring in the era of big data," *Yale Journal of Low. & Technology, 18*, 2017.

[19] N. Raghunath, "A sociological review of china's social credit systems and guanxi opportunities for social mobility," *Sociology Compass*, vol. 14, no. 5, p. e12783, 2020.

[20] D. Nemoto, T. & Beglar, "Developing likert-scale questionnaires," *JALT2013 Conference Proceedings. Tokyo: JALT*, 2014.

Appendix

TABLE VIII. INTERNET AND SOCIAL USERS QUESTIONNAIRE LIKERT PART

	Questions	Srtongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	I use the internet a lot	1	2	3	4	5	6
2	I use the social media sites a lot						
3	The internet is useful for buying products at a good price						
4	I'm frequently used internet to make purchases and sales processes						
5	I'm frequently used application of technology companies like Uber						
6	It is possible to encounter too many problems during sales through internet						
7	It is possible to encounter too many problems during buying through internet						
8	It is possible to encounter some verbal and psychological attacks on social						
9	I think My privacy can be violated on social media media						
10	I think My personal data can be stolen on internet						
11	I'm frequently executed my banking transactions online						
12	I believe that transactions via the Internet and social media can extract information that can be used to build a creditworthiness system similar to that found in banks						
13	Total Scores						

TABLE IX. INTERNET AND SOCIAL USERS QUESTIONNAIRE – YES/NO PART

	Questions	Yes	No
1	Have you encountered problems with the quality of the product received?		
2	Have you faced financial problems when selling on the Internet?		
3	Have you faced financial problems when buying from the Internet?		

TABLE X. INTERNET AND SOCIAL USERS QUESTIONNAIRE – CHOOSE PART

	Questions	1	2	3	4
1	What is payment method you prefer when you buy something from the internet	visa	when receiving		

TABLE XI. COMPANIES AND ONLINE SERVICE PROVIDERS' QUESTIONNAIRE LIKERT PART

	Questions	Srtongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	I'm frequently performed commercial transactions through the internet	1	2	3	4	5	6
2	Dealing with suppliers through the internet has increase the transaction problem						
3	The internet has increased my income						
4	Most of my internet buyers don't complete their transaction						
5	Most of my internet buyers refuse to receive the items						
6	Most of my internet buyers use Product returns process						
7	I think we can build on people's history of how they treat tech companies apps like Uber to rank a new customer						
8	I see that we can build on the history of people who interact negatively or positively with social media in rating a new user customer						
9	I think that the existence of a measure of the user's credit worthiness based on the study of his behavior over the Internet increases confidence in electronic transactions						
10	Total Scores						

TABLE XII. COMPANIES AND ONLINE SERVICE PROVIDERS' QUESTIONNAIRE – YES/NO PART

	Questions	Yes	No
1	Have you encountered problems with the quality of the product received from suppliers?		
2	Have you encountered financial problems when dealing electronically with suppliers?		
3	Did you encounter financial problems with buyers?		

TABLE XIII. Those are Familiar with Regulations, Laws and Banking Questionnaire Likert Part

	Questions	Srtongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	I think it is possible to allow tracking of electronic transactions of users over the Internet	1	2	3	4	5	6
2	I think it is possible to allow the use of this information gathered online						
3	I think it is possible to use this information without facing litigation in the future						
4	Total Scores						

TABLE XIV. Those are Familiar with Regulations, Laws and Banking Questionnaire – Yes/No Part

	Questions	Yes	No
1	We need laws that manage the process of tracking users' electronic transactions		
2	We need laws that manage the use of information about users		
3	We need laws that manage the process of publishing information about users		

Extended Graph Convolutional Networks for 3D Object Classification in Point Clouds

Sajan Kumar[1], Sai Rishvanth Katragadda[2], Ashu Abdul[3], V. Dinesh Reddy[4]
School of Engineering and Sciences
Department of CSE, SRM University,
Amaravati, AP, India

***Abstract*—Point clouds are a popular way to represent 3D data. Due to the sparsity and irregularity of the point cloud data, learning features directly from point clouds become complex and thus huge importance to methods that directly consume points. This paper focuses on interpreting the point cloud inputs using the graph convolutional networks (GCN). Further, we extend this model to detect the objects found in the autonomous driving datasets and the miscellaneous objects found in the non-autonomous driving datasets. We proposed to reduce the runtime of a GCN by allowing the GCN to stochastically sample fewer input points from point clouds to infer their larger structure while preserving its accuracy. Our proposed model offer improved accuracy while drastically decreasing graph building and prediction runtime.**

Keywords*—Object classification; graph convolution networks; non-autonomous driving

I. Introduction

Autonomous vehicles are becoming the future of mobility, supported by advances in deep learning techniques. Point cloud learning has lately attracted increasing attention due to its wide applications in many areas, such as computer vision, autonomous driving, and robotics. Recent advances in graph convolution networks suggest that graph representations could provide better features for point cloud processing. Graphs are one of the most common data structures in the analysis and storage of the real-world data-modeling of social networks, roads, etc. However, the amount of work devoted for developing the neural network models to process graphs has not been proportional to the amount of data available for such analysis. In the past couple of years, some researchers have looked at generalizing current neural network models to process arbitrary graphs, some of which we will briefly summarize later in this section. Thomas et al. [1] offers a brief overview of GCNs and the architecture is shown in Fig. 1. In essence, most GCNs have a universal architecture and their convolutional nature arises from sharing filter parameters over all graph locations.

The objective of a GCN is to learn a function of features for a given graph G, which takes the inputs:

- X, an $N \times D$ feature matrix which summarizes a feature description x_i for every node i (where N is the number of nodes, and D the number of input features),
- A descriptor of the graph's structure in matrix form, usually as an adjacency matrix A.

This produces an output Z, which is a node-level $N \times D$ feature matrix (F is the number of output features per node).

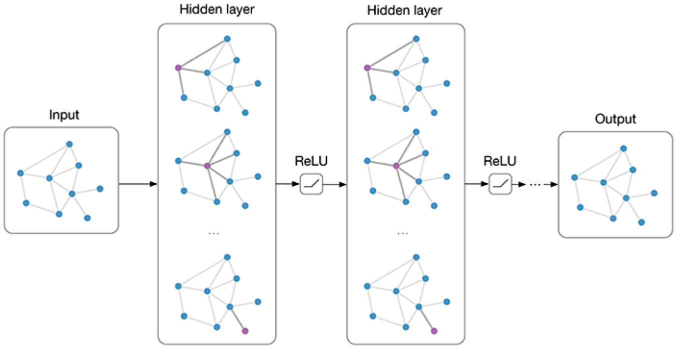


Fig. 1. Overview of a Graph Convolutional Network (GCN) and its Associated Layers [1].

Each layer H^l out of the total layers L in the GCN can then be modeled as a nonlinear function:

$$H^{(l+1)} = f(H^{(l)}, A)$$

, where

$$H^{(0)} = X \text{ and } H^{(L)} = Z. \quad (1)$$

Most GCNs only differ in how the nonlinear function f is parametrized and chosen.

The rest of this paper is organized as follows. Section II presents motivation as well as related works. Limitations and contributions are presented in Section III. Section IV elaborates the pointGCN model and the proposed work is presented in Section V. Experimental studies are then presented in Section VI, followed by a discussion on the results in Section VII. Finally, conclusions are provided in Section VIII.

II. Related Works

Existing research on processing the arbitrary graphs proposed the usage of the neural networks for extracting the information from the arbitrary graphs. Bruna et al. [2] used the graph-based analogs with the convolutional neural networks (CNN) to obtain an efficient architecture by reducing the number of parameters, relying on hierarchical clustering of the graph and by analyzing the spectrum of the graph's Laplacian matrix. David et al. [3] introduced a CNN approach with an end-to-end pipeline that could operate directly on arbitrarily sized graphs to generalize molecular feature extraction. Jain et al. [4] used the high-level modeling ability of spatio-temporal graphs to improve a recurrent neural network architecture to model the sequential computer vision tasks (like human or object interactions) more effectively. Mich et al. [5] proposed a true generalization of CNNs to work on any graph structure by developing a model based in spectral graph theory, which affords the same linear computational and learning complexity as traditional CNNs. This work constitutes a big jump forward in the field in terms of modeling GCNs. Moreover, Kipf et al. [1], created a GCN model based on spectral graph convolutions that scale linearly in the total number of graph edges. The model layer representations encoded the specific features of nodes as well as the local structure of the graph around a given region. Hence it is clear that GCNs offer a better scheme for the analysis of data that is generally arbitrarily grouped or sparse in nature, such as the structure seen in point clouds. In terms of the application of CNNs to the processing of 3D point clouds, PointNet [6] is the first major DL method for 3D classification and segmentation. The network directly takes point clouds as input and allows for efficient and effective classification, segmentation and scene semantic parsing. The model proposed by kiran et al. [7] uses 3D prior maps to reduce the computational requirements on 3D point clouds. Zar zar t al. [8] proposed the PointRGCN for using the point cloud inputs for vehicle tracking. The model uses a residual GCN and a contextual GCN for refining a 3D object from a point cloud using a graph representation of the object.

We propose to extend the PointGCN model developed by zang et al. [9]. We down sample the graph with localized convolutions to obtain the latent features for describing the local structures of the input point cloud. Our approach attempts to leverage the flexibility of GCNs in dealing with unstructured input with the inherent nature of point clouds. From our experiments, we see that proposed model reduces the computational requirements on 3D point clouds and improves the performance.

III. Current Limitation of GCN Models

Most of the research in the field of point cloud computation involves converting of 2D image to a point cloud to extract information. The existing research focusses on utilizing the depth information collected via a LiDAR for classification or segmentation (as done in PointNet). We know that such methods are accurate individually, but do not necessarily make the best use of the data at hand. The GCN models which we reviewed are effective but not necessarily specialized for real-world object classification tasks. PointGCN [9] does not appear to target any specific application and hence is not specifically optimized for any purpose. Likewise, PointRGCN [8] mainly works on bird's-eye view detection and classification tasks rather than typical views from a camera on the ground. It also uses a variety of different models to obtain its final result. Through our work, we would like to extend [9] and [8] to fit other kinds of point cloud datasets without restricting to autonomous driving. We will be looking into using some of the datasets at [10] to train on and will attempt to generalize the model to gain good predictive performance on these varied datasets.

The main contributions of the paper are:

- Extending and generalizing a GCN model to work on other kinds of LiDAR point clouds such as objects in urban environments or the interiors and exterior of buildings, without suffering a significant performance loss.
- Optimizing and generalizing current GCN model such that it can run on small computing processors such as arm cortex A7.

Our work will attempt to combine useful features from the current state of the art (for 3D object classification) into an innovative approach that will hopefully match current performance in the field at a reduced time cost. The models currently used in research (processing on a point cloud with PointNet and RGB feature/depth extraction) have their own merits and demerits. 2D RGB images are better for feature extraction and segmentation at a reduced time cost. However, the 2D RGB images lack the depth information, which leads to poor capturing of the relationship between the subject and the objects. In our approach we use the point clouds to bridge this gap for extracting the relationship between the objects and the subject. We note that we can generalize the current methods in the field to adapt with different types of data. In our approach, we adapt a method without suffering the feature loss. This method will be important for the researchers working on autonomous driving, cave mapping, home modeling, and video game designers, to name a few.

IV. Architecture of the PointGCN Model

The PointGCN appears to have the potential for flexibility in inputs and did not require the processing of its data through multiple models [9]. It was also extremely lightweight in terms of the code involved, and so would be easier to modify.The following is a brief overview of the model architecture for a Graph Convolutional Network for the purpose of 3D object classification as used in [9]. Similar to most simple CNNs, the GCN consists of a convolution layer, a pooling layer, and a fully connected layer. The graph convolutional layer allows the GCN to encompass the structural information of the object to be able to discriminate between them. Graph Laplacians of the input point cloud data are generated. The graphs are normalized to keep them to a uniform spectrum range. As described in [11], the ChebyNet Graph-CNN performs better on homogenous graphs for prediction tasks such as image classification. Hence, a similar transformation is applied to the heterogeneous graphs generated from point clouds. To obtain a single level of a feature transformation, the model applies Chebyshev polynomial filters, which keep the learned feature maps localized.

A Rectified Linear unit (ReLu) nonlinear activation function is applied at the output of each graph convolutional layer. Furthermore, global max pooling is performed after these activations are applied. This allows for the computation of the global statistics of all the output points. When multiple graph convolutional layers are used, the statistics of each layer are used together for the final probability computations.

The version of the model we used two graph convolutional layers, and the Chebyshev polynomials are of order = 4 and order = 3 for the two layers respectively. The rest is similar to as used in [9], with a 40-nearest neighbor graph for each point cloud object for graph convolution and global pooling. The intermediate outputs of the graph convolution layers are passed through a final fully-connected linear layer with softmax activation as a flattened vector. The number of output nodes in the output layer corresponds to the number of labels in the dataset, which in our case, we kept at 40 labels. Hence, the output is a vector where each index corresponds to the probability of the input belonging to that label.

V. Proposed Approach - Modified PointGCN

PointGCN was geared to be used with the ModelNet40 dataset. Therefore, in order to use the model with different datasets and different labels, we had to make the model label-agnostic, and make sure the data format we wanted to use worked everywhere in the model.

We also approached the problem from another angle. The initial implementation in [9] used 2048 point cloud examples per test/train set, and each point cloud consisted of 2048 points. This amounted to an extremely large amount of input data and meant the creation of the nearest-neighbor graphs from the point cloud data was computationally expensive and extremely memory-intensive. It often formed a bulk of the actual runtime of the network where data was fed into it. Therefore, we tried running the model and testing its accuracy with fewer training examples and points per cloud instead, which would significantly reduce runtime if successful.

We then also attempted to make some changes in the PointGCN model itself, by introducing link prediction and an entity classification layer, which helps in the recovery of missing points or facts from the dataset, and recovery of attributes corresponding to the missing entity, which was motivated by work in [12]. Hidden representation of GCN of ith layer is computed by:

$$h_i^{l+1} = \sigma\left(\sum_{jN_i} \frac{1}{c_i} W^l h_j^{(l)}\right)$$

If we introduce the link prediction and entity layer in the above function to utilize this hidden representation we can compute the edges more freely by:

$$h_i^{l+1} = \sigma\left(W_0^l h_i^l + \sum_{rR} \sum_{jN_i} \frac{1}{c_{i,r}} W_r^l h_j^{(l)}\right)$$

where $N_{r,i}$ denotes the set of neighbor indices of node i under relation rR and ci,r is a normalization constant. In the entity classification layer the GCN uses $c_{i,r} = |N - ri|$. This method helps determine a node's relational encoding, aids in nearest-neighbor finding, and helps reduce the computation cost. Though the there is an increase in the number of hyperparameter's the paper [13]

We decided to try and generalize the PointGCN model to achieve a good accuracy over two different datasets:

- Princeton ModelNet40 Dataset [14]: A volumetric representation of 3D objects across 40 categories, generated in a highly controlled environment using CAD models.
- CMU Oakland Dataset [15]: Contains 44 labels as shown in Fig. 2. This is generated using a SICK LMS laser scanners and used in push-broom, collected around the CMU campus in the neighborhood of Oakland, Pittsburgh, and contains X, Y, and Z point coordinates, corresponding labels, and confidence as properties.

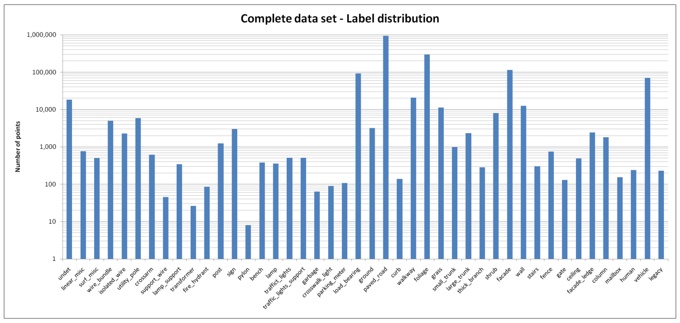


Fig. 2. A Log-scale Plot of the Distribution of the 44 Different Labels in the CMU Oakland dataset. The Four Labels with the Lowest Frequencies were Dropped out to Maintain Consistency with the 40 Outputs in the ModelNet40 Dataset.

A. Preparing Input Datasets

We faced some challenges in deciding how to generalize the input datasets to the model. The biggest is the representation of the data points in point clouds. Each prepared dataset uses a different representation of 3D point clouds, which is specific to a single model's architecture. Furthermore, the collection methods and reference frame of the collecting device also vary across datasets. Some datasets contain far more information (such as semantic annotations or segmentations) that cannot be found in other datasets. Each dataset also has a different density and distribution of points for every object in the set. The CMU Oakland dataset has labels for objects that occur so infrequently in the set that the data is unusable on a larger scale (mainly power cables or wires in the background). The PointGCN model constructs its initial graph data from two arrays found in an HDF5 file. The first, the 'data' array, is an N-by-M array with N point cloud examples and M points per point cloud. Each point in the point cloud is a 3-by-1 array consisting of the X, Y, and Z coordinates of the point. The second array is the 'label' array, an N-element array, with each index corresponding to a point cloud in the 'data' array. This meant there must be a generalized method of preparing or feeding the respective data into the model. As such, for object classification, we decided to use the bare minimum. A preprocessed dataset to the model should contain the X, Y, and Z coordinates per point, along with the object label to which that point belongs.

B. Dataset Creation

To overcome the issues we faced with point sparsity and the parsing of datasets, we wrote a script to process and 'mass-create' datasets that we could test on. This script was tested and used on the CMU Oakland dataset [15] which is shown in Fig. 3. It requires the point cloud x, y, and z coordinates, along with the corresponding label for the point, in four columns in a formatted text file as the input. This required some manual processing, including dropping points with the four least frequent labels (essentially unusable data). Dropping four labels from the 44-label Oakland dataset also allowed us

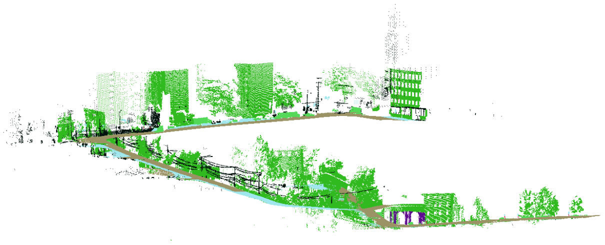


Fig. 3. A Visualization of the Complete, Preprocessed CMU Oakland Point Cloud Dataset, with the Lowest Frequency Labels Dropped from the Set.

to keep the model consistent with the total data labels in the ModelNet40 set. This would allow us to cross-test graphs made with one dataset on another dataset. A maximum number of point clouds per data file is set as well. These data points were later split into test and train datasets using a random distribution.

Once the data is read in and sorted, the creation process can begin. A label is randomly selected on a uniform distribution, and the required number of points from the corresponding data is picked to form a point cloud. In order to overcome the problem where a labeled object has too few points to create a substantially-sized point cloud for training, we add noise, or jitter, to the ground truth points, in order to create new points for the object.

The jitter fulfilled two purposes at once. First, since these points are only jittered with a value between 0.25 and 0.5 in Cartesian space, the created points would “fill in the blanks” in a sparsely-sampled object while remaining close enough in physical space to the object in question, as illustrated in Fig. 4. Secondly, the addition of random noise, and the randomized nature of label selection, allows every created dataset to be different. This would allow us to test the robustness of the model to small perturbations as well. Finally, the datasets were split into test and training datasets, of which five datasets were used for training, and two for testing.

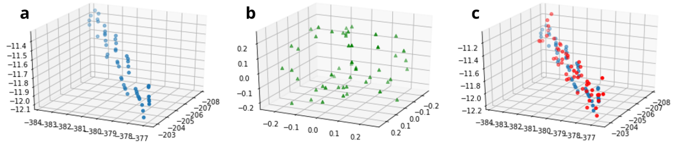


Fig. 4. A Visualization of the Jittering Process: (a) the Set of Points from the Point Cloud for the Label ‘curb’; (b) Randomly Generated Jitter Values Corresponding to Each of the Points; (c) the Original Points (blue) Along with the Jittered Original Points (red) are Added to the Point Cloud

VI. Results

The modified model was run using Tensorflow with GPU acceleration. The learning rate was set to between $12and24 \times 10^4$, and was halved every 20 epochs or so.Our first experiment simply involved training the PointGCN model on the Oakland dataset and the ModelNet40 dataset separately, and comparing the results. With the ModelNet40, after 260 epochs of testing and training, we received an average test accuracy of 86% among the 40 input classes. For the Oakland set, four of the least frequent data labels were dropped, to ensure consistency between the total number of PointGCN outputs. After 260 epochs of testing and training, we received an average test accuracy of 91%.

For the smaller point cloud tests, sets of 64 and 128 points per point cloud were used, with only 1000 training and test examples per set (5000 training examples and 2000 test examples in total). The maximum number of epochs were set at a 100. The graph batches used by the network were first built using the Oakland dataset, and then cross-tested with the ModelNet40 dataset, in order to test the flexibility of the nodes created by the model. The final results after a 100 epochs can be seen in Table I and Table II below.

TABLE I. Final Training and Test Accuracy of CMU Oakland and Princeton ModelNet40 Datasets with a Reduced Number of Points Per Cloud. With the Base Graph for the GCN Model Built using the Oakland Dataset

	64 Points per Cloud		128 Points per Cloud	
Accuracy (%)	Oakland	ModelNet40	Oakland	ModelNet40
Training Data	95.82	90.31	96.06	90.37
Test Data	91.30	87.67	92.00	88.80
Overall	**93.56**	**88.99**	**94.03**	**89.59**

TABLE II. Final Training and Test Accuracy of CMU Oakland and Princeton ModelNet40 Datasets with a Reduced Number of Points Per Cloud, with the Base Graph for the GCN Model Built using the ModelNet40 Dataset

	64 Points per Cloud		128 Points per Cloud	
Accuracy (%)	Oakland	ModelNet40	Oakland	ModelNet40
Training Data	94.65	87.72	94.23	93.73
Test Data	83.66	84.59	85.18	87.45
Overall	**89.15**	**86.15**	**89.71**	**90.59**

The results were quite surprising. Even with a significantly reduced number of points per point cloud, both the test and training accuracies of the model on both input datasets were extremely high. When a dataset is used to construct the graph batch used by the GCN model, it demonstrates a significantly higher training and test set accuracy. Nonetheless, given that both ModelNet40 and Oakland achieve 90% accuracy and 89% accuracy respectively, on graphs not even constructed from their data, shows the flexibility and robust nature of the model. Furthermore, the runtime is also drastically decreased, especially on a GPU setup. For 7000 input point clouds with 64 points per cloud, the graph build runtime was a mere 8 minutes, and the runtime for a 100 epochs was approximately 9 minutes and 40 seconds. For the same number of input clouds at 128 points per cloud, the graph build runtime was around 13 minutes and 38 seconds, with the runtime for a 100 epochs at 12 minutes and 34 seconds. This is quite remarkable given that the runtime with over 14,000 input clouds at 2048 points per cloud has a total runtime of over a day for similar levels of accuracy, even when GPU-accelerated.

VII. Discussion

From the results of the GCN’s prediction on the ModelNet40 and Oakland datasets, it is clear that the model has been extended to run with more than a single, specialized dataset. An approach using a GCN offers an adaptability that a CNN would not offer if trained on a single dataset. This would reduce the need to have to train a model on a different dataset for a different use case every time. From cross-testing one dataset on a graph built from another dataset, it is clear that the GCN has the ability to learn and refine graphs such that it can reflect the overall structure of an input point cloud. It is also a model that has the potential to be robust to changes in the input dataset. However, on comparison with a variety of other models on the ModelNet leaderboard at [16], this model falls some ways short of the state-of-the-art methods. Nonetheless, it is still better than a lot of other models on the list and so is an approach that shows promise.

TABLE III. COMPARISON WITH STATE-OF-ART METHODS ON MODELNET40 & OAKLAND DATA SET

Method	ModelNet40	Oakland
MV3D [17]	62.94	57.31
AVOD [18]	77.90	62.03
PointPillars [19]	78.39	67.39
PointRCNN [20]	85.94	73.210
RGCN [13]	83.42	74.05
PointRGCN [13]	85.97	76.73
Extended PointRGCN(Ours)	90.59	89.71

We implemented the MV3D, AVOD, PointPillars, PointRCNN, RGCN, PointRGCN, and compared with the proposed approach. All the results reported above are the average results of 10 trials for each dataset for our proposed method taking 128 points per cloud and the results are presented in Table III. The high accuracy achieved by the model on the reduced-size point clouds was also a very curious result. As such, we looked closer at how the nature of the input data to the GCN could affect its performance. The GCN model we adapted could use one of two methods for sampling from the larger point cloud set - immediate sampling or uniform sampling. Immediate sampling just takes the first 64 or 128 points in the array from the entire point cloud. Uniform sampling attempts to take 64 or 128 samples from a uniform random probability distribution across the point cloud. If the points are unsorted, both sampling methods are equivalent. However, if the points are sorted before they are passed in, using immediate sampling shows very different effects. We attempted to sort the Oakland dataset point clouds before passing them into the GCN. We found that the accuracy on the Oakland set itself dropped to a meager 33% on the training set and 24% on the test set, while the ModelNet data achieved a 41% on the training set and 37% on the test set.

This poor performance could be attributed to the GCN receiving a poor understanding of the general structure of the point cloud from the input data. In other words, with the data sorted and sampled from the front, the GCN can essentially “visualize” only a very small fraction of the entire cloud. This does not allow it to truly infer any larger-scale features, and any features it were to infer would be grossly incorrect, given that it only has access to a small physical portion of the cloud. When the data was left unsorted, it allows the GCN to sample parts of the point cloud stochastically, thus giving it enough points to infer larger features in the cloud and encode it in the graph. We can conclude from this smaller investigation that allowing uniform, randomized sampling of the entire point cloud, even with a small number of points, can offer the same levels of accuracy while drastically decreasing graph building and prediction runtime.

VIII. CONCLUSION AND FUTURE WORK

In this paper a novel approach for extending and generalizing a GCN model to work on different point clouds such as objects in urban environments or the interiors and exterior of buildings, without suffering a significant performance loss. Further, we worked on Optimizing and generalizing current GCN model such that it can run on small computing processors such as arm cortex A7. Through our experimentation on a modified version of PointGCN, we conclude that it is possible to generalize a graph convolutional network across datasets. The proposed model seems to be able to learn and preserve underlying patterns that appear in point clouds regardless of the object represented by the point cloud, and thus make for powerful processors of the data type. Additionally, our model do not appear to lose accuracy significantly when the number of points in the point cloud input is reduced, as long as the points remain representative of the larger structure of the object at hand.

In future, we would like to extend our work by looking at how exactly the graph generated by the GCN seems to preserve the global statistics of features, and how it can effectively exploit those same statistics across datasets. We can improve the performance of our current model by adding deeper layers, where they can compute the interrelation between points to generate a feature map across the datasets. As part of future work, we also plan to see if we could combine LiDAR information with RGB images for better recognition.

REFERENCES

[1] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[2] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” *arXiv preprint arXiv:1312.6203*, 2013.

[3] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, “Convolutional networks on graphs for learning molecular fingerprints,” *arXiv preprint arXiv:1509.09292*, 2015.

[4] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 5308–5317.

[5] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, 2016, pp. 3844–3852.

[6] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[7] B. Ravi Kiran, L. Roldao, B. Irastorza, R. Verastegui, S. Suss, S. Yogamani, V. Talpaert, A. Lepoutre, and G. Trehard, “Real-time dynamic object detection for autonomous driving using prior 3d-maps,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[8] J. Zarzar, S. Giancola, and B. Ghanem, “Pointrgcn: Graph convolutional networks for 3d vehicles detection refinement,” *arXiv preprint arXiv:1911.12236*, 2019.

[9] Y. Zhang and M. Rabbat, “A graph-cnn for 3d point cloud classification,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6279–6283.

[10] Y. Guo, “Deep learning for 3d point clouds: A survey,” *https://arxiv.org/abs/1912.12033*.

[11] P. V. D. I. Shuman and P. Frossard, "Chebyshev polynomial approximation for distributed signal processing," *n Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems 2011*, 2011.

[12] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," *arXiv preprint arXiv:1703.06103*, 2017.

[13] Y. Lingfan, L. Mufei, and Z. Zheng, "Relational graph convolutional network," *DGL*, 2018.

[14] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[15] N. V. D. Munoz, J. A. Bagnell and M. Hebert, "Contextual classification with functional max-margin markov networks," *In IEEE Conference on Computer Vision and Pattern Recognition 2009*, 2009.

[16] Princeton, "Princeton modelnet." [Online]. Available: https://modelnet.cs.princeton.edu/

[17] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, in *Multi-view 3d object detection network for autonomous driving*. IEEE, 2017, pp. 1907–1915.

[18] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, in *Point 3d proposal generation and object detection from view aggregation*. IEEE, 2018, pp. 1–8.

[19] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[20] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

Techniques for Solving Shortest Vector Problem

V. Dinesh Reddy[1], P. Ravi[2], Ashu Abdul[3], Mahesh Kumar Morampudi[4], Sriramulu Bojjagani[5]
School of Engineering and Sciences, Department of CSE
SRM University, Amaravati, AP, India[1,3,4,5]
Assistant Professor, CSE,
Anurag University, Hyderabad, Telangana, India[2]

***Abstract*—Lattice-based crypto systems are regarded as secure and believed to be secure even against quantum computers. lattice-based cryptography relies upon problems like the Shortest Vector Problem. Shortest Vector Problem is an instance of lattice problems that are used as a basis for secure cryptographic schemes. For more than 30 years now, the Shortest Vector Problem has been at the heart of a thriving research field and finding a new efficient algorithm turned out to be out of reach. This problem has a great many applications such as optimization, communication theory, cryptography, etc. This paper introduces the Shortest Vector Problem and other related problems such as the Closest Vector Problem. We present the average case and worst case hardness results for the Shortest Vector Problem. Further this work explore efficient algorithms solving the Shortest Vector Problem and present their efficiency. More precisely, this paper presents four algorithms: the Lenstra-Lenstra-Lovasz (LLL) algorithm, the Block Korkine-Zolotarev (BKZ) algorithm, a Metropolis algorithm, and a convex relaxation of SVP. The experimental results on various lattices show that the Metropolis algorithm works better than other algorithms with varying sizes of lattices.**

Keywords*—*Lattice; SVP; CVP; post quantum cryptography

I. Introduction

A lattice an abstract structure defined as the set of all integer linear combinations of some independent vectors in R_n. Over the last two centuries, mathematicians have explored the fascinating combinatorial structure of lattices, and it has also been studied from an asymptotic algorithmic viewpoint for at least three decades. Most fundamental problems are not considered to be effectively solvable on lattices. In addition, hardness results suggest that such problems can not be solved by polynomial-time algorithms unless the hierarchy of polynomial-time collapses. Cryptographic constructions based on lattice hold a clear promise for cryptography with strong security proof, as was demonstrated by Ajtai [1], who came up with a construction of cryptographic primitives based on worst-case hardness of certain lattice problems.

Two main important and very closely related hard computational problems used by cryptographers are the Shortest Vector Problem (SVP) and the Closest Vector Problem (CVP). In the former, for a lattice specified by some basis we are supposed to find nontrivial and small nonzero vector (length of vector) in the lattice. The problem CVP is an inhomogeneous variant of SVP, in which given a lattice (specified by some basis) and a vector v, one has to find the vector in L closest to v. The hardness of these problems is partially due to the fact that multiple bases can generate the same lattice. This work presents the best hardness result known for SVP, compares different algorithms solving SVP with respect to their efficiency and optimal solution.

II. Preliminaries

A. Definitions and Related Problems

Let $(b_1,\ldots,b_n)$ be a basis in $\mathbb{R}^n$ and N a norm on $\mathbb{R}^n$. Let L denote $\bigoplus_{i=1}^{n}\mathbb{Z}b_i, L$ is called a lattice. Set

$$\lambda(L)=\min_{v\in L\backslash\{0\}} N(v)$$

We are now able to define SVP. For L a lattice, the exact form Shortest Vector Problem, denoted by $SVP(L)$ is as follows.

$$SVP(L):\ \text{Find } v\in L \text{ such that } N(v)=\lambda(L)$$

As it is often the case, we will actually be interested in algorithms solving an approximation of SVP. For $\gamma\geqslant 1$ the approximated form of $SVP(L)$, denoted by $SVP_\gamma(L)$ is

$$SVP_\gamma(L):\ \text{Find } v\in L\backslash\{0\} \text{ such that } N(v)\leqslant\gamma\lambda(L)$$

Now, a closely related problem is the Closest Vector Problem $CVP(L,x)$, defined for a lattice $L\subset\mathbb{R}^n$ and $x\in\mathbb{R}^n$ as

$$CVP(L):\ Find\ v\ \in L \text{ such that } N(v-x)=\lambda(L)$$

In its exact form. As in the above we can define an approximated problem CVP_γ

$$CVP_\gamma(L):\ Find\ v\in\ L\ such\ that\ N(v-x)\leqslant\gamma\lambda(L)$$

In fact, CVP_γ solves SVP_γ. Indeed, for $i=1,\ldots,n$ set $L^i=\mathbb{Z}b_1\oplus\ldots\oplus 2b_i\oplus\ldots\oplus b_n$ and x_i a solution to $CVP_\gamma\left(L^i,b_i\right)$, then one of the vectors in $\{x_i-b_i\}_{i=1\ldots n}$ is a solution to $SVP_\gamma(L)$. For we have the following claim:

Claim 1. Let x_i be a solution to $CVP\left(L^i,b_i\right)$ then $\{x_i-b_i\}_{i=1,\ldots,n}$ contains a solution to $SVP(L)$
Proof. Indeed, let $x=\sum n_ib_i$ be any solution to $SVP(L)$, as $N\left(\frac{x}{2}\right)<N(x)$ we get that $\frac{x}{2}\notin L$ and there is $i\in\{1,\ldots,n\}$ such that n_i is odd. Then, $x+b_i\in L^i$ Therefore, the set $\{x_i-b_i\}_{i=1,\ldots,n}$ contains a element of norm less than or equal to $N(x)$.

When N is the l^2 norm on $\mathbb{R}^n, CVP$ has another formulation that allows us to use convex optimization techniques. To this end, let B be the matrix given by the basis $(b_1,\ldots,b_n)$ and c a vector in $\mathbb{R}^n$. Then, $CVP(L,c)$ is equivalent to

$$\begin{aligned}&\text{minimize} && x^TB^TBx-2c^Tx\\ &\text{subject to} && x\in\mathbb{Z}^n\end{aligned}$$

This is readily seen by expanding out the quantity $\|Bx-c\|_2$. Following (ref), we will apply a convex relaxation to this problem to get an efficient randomized algorithm solving CVP_γ.

Finally, a last related problem is the Hermite Shortest Vector problem denoted by $HSVP$. Again, let $B \in \mathbb{R}^{n\times n}$ be the matrix given by $(b_1,\dots,b_n)$, for $\gamma \geqslant 1$, $HSVP_\gamma(L)$ is the following problem.

$$HSVP_\gamma(L): \text{ Find } v \in L \text{ such that } N(v) \leqslant \gamma|\det(B)|$$

According to a theorem of Minkowski's, HSVP and SVP are in fact closely related. The LLL algorithm described in Section III-A actually solves $HSVP_\gamma$.

B. Hardness: Average and Worst Case

An interesting feature of lattice problem is there hardness. M.Ajtai in his seminal papers [1], [2] showed in particular that worst-case hardness is related to average-case hardness and that SVP is NP-hard for randomized reductions. This makes lattice problems particularly interesting as basis for crypto systems. Example of such systems are systems of Ajtai and Dwork [3], Goldreich, Goldwasser and Halevi [4], the NTRU cryptosystem [5], and the first fully homomorphic encryption scheme by Gentry [6]. In this section we state a number of hardness results without proofs. For a more precise treatment see also [7]. The main result of worst-case to average case NP-hardness is as follows.

Theorem 1 Suppose there is a randomized polynomial time algorithm $\mathcal{A}$ such that for all $n, \mathcal{A}$ returns a vector of $\Lambda(X)$ of length $\leqslant n$ with probability $\frac{1}{n^{O(1)}}$ for $\Lambda(X)$ chosen at random in $\Lambda_{n,m,q}$ where $m = \alpha n \log(n)$ and $q = n^\beta$ for appropriate α, β. Then, there exist a randomized polynomial time algorithm $\mathcal{B}$ and constants c_1, c_2, c_3 such that for all $n, (b_1,\dots,b_n)$ basis of $\mathbb{R}^n$ and $L := \oplus\mathbb{Z}b_i, \mathcal{B}$ performs the following with high probability [3]:
1) Finds a basis $\{\beta_1,\dots,\beta_n\}$ for L such that

$$\max\|\beta_i\| \leqslant n^{c_1} \min_{(c_i)\ basis\ of\ L} \max\|c_i\|\,;$$

2) finds an estimate $\bar{\lambda}$ of $\lambda(L)$ such that,

$$\frac{\lambda_1(L)}{n^{c_2}} \leqslant \tilde{\lambda} \leqslant \lambda(L)$$

3) If moreover, L has a n^{c_3} unique shortest vector, finds this vector.

We are now oing to explain the content of this theorem. First of all, we must explain a number of concepts and notation.

Here, $\|.\|$ refers to the l^2 -norm.

A randomized algorithm is an algorithm that uses a degree of randomness as part of its process. As such, it does not guarantee success, but the probability of success for a given input size can be computed. A randomized algorithm is said to perform an event with high probability if the probability of this event goes to 1 as the input size goes to $+\infty$

We turn now to the definition of $\Lambda(X)$. Let $n, m, q \in \mathbb{N}, (\mathbb{Z}/q\mathbb{Z})^{n\times m}$ the space of matrices with entries in $(Z/q\mathbb{Z})$, and $\Omega_{n,m,q}$ the uniform distribution on $(\mathbb{Z}/q\mathbb{Z})^{n\times m}$ Now, for any $X \in (\mathbb{Z}/q\mathbb{Z})^{n\times m}$, set $\Lambda(X)$ as the lattice $\{y \in \mathbb{Z}^m \mid Xy \equiv 0[q]\}$ and $\Lambda_{n,m,q}$ the space of such lattices. We see $\Lambda_{n,m,q}$ as a probability space by choosing X according to $\Omega_{n,m,q}$ and computing $\Lambda(X)$. For any $X, \Lambda(X)$ is indeed a lattice, as $q\left(\mathbb{Z}^m\right) \subset \Lambda(X)$.

Finally, for $c \geqslant 1$ a lattice L endowed with a norm N is said to have a c-unique shortest vector v if the following holds

$$\{v \in L \mid N(v) \leqslant c\lambda(L)\} = \{v, -v\}$$

This result is called a worst-case to average case hardness because it links the lack of a polynomial ranodmized algorithm over 'all' lattices to the lack of such algorithm for a small class of lattices, namely $\Lambda_{n,m,q}$. We also note that similar results exist for related problems such as CVP. In the following, we list further hardness results.

Theorem 2 SVP is NP -hard under the l_∞ -norm. Moreover, CVP is NP-hard under the l_p -norm for all $p \geqslant 1$ [8]..

Theorem 3 For all γ, CVP_γ is NP-hard under any l_p -norm [9].

Building on results by Ajtai [2], Micciancio proved the following.

Theorem 4 For all $\epsilon > 0, SVP_{\sqrt{2}-\epsilon}$ is NP -hard under randomized polynomial time reductions [10].

Finally:

Theorem 5 There is $c > 0$ such that CVP is NP-hard to approximate within a factor $n^{\frac{c}{log(log(n))}}$, where n is the dimension of the lattice [11].

Many other results related to hardness exist. We only decided to mention some of the most fundamental ones.

III. Solving SVP

This section presents solving SVP and related results. Even though many hardness results have been found, we can still solve approximate SVP within some reasonable constant depending on the input size. Usually, the 'reasonable constant' is not known theoretically to be 'reasonable' but only empirically. For instance, the first algorithm on this list, LLL, is only known to solve SVP within a factor exponential in the input size. However, results with LLL are in practice much better than the bound.

A. Lenstra-Lenstra-Lováz Algorithm

This section presents the Lenstra-Lenstra-Lovász reduction basis algorithm, by Lenstra, Lenstra and Lovász, see [12]. Originally, the purpose of this algorithm is to factor polynomials, we will not consider this problem here.

Moreover, it is important to note that the techniques used were developed by Hermite, Minkowksy and others to study Siegel sets and lattices in algebraic groups. This is why LLL is in fact an algorithm solving $HSVP$. $HSVP$ and SVP are related thanks to the following results due to Minkowski.

Theorem 6 (Minkowski). Let $n \in \mathbb{N}$ and L be any lattice in $\mathbb{R}^n$, then

$$\lambda(L) \leqslant \gamma_n(\det(L))^{\frac{1}{n}}$$

where γ_n is some constant depending only on n. More precisely, γ_n is known to satisfy

$$\sqrt{\frac{n}{2\pi e}} \leqslant \gamma_n \leqslant \sqrt{\frac{n}{\pi e}}$$

This theorem admits a short and elementary proof but can also be seen as a consequence of Minkowski's convex body theorem. For proofs and further details see [13] and [14]. In particular, we get the following important fact.

Claim 2. For $\gamma \geqslant 1, HSVP_{\gamma_n\gamma}(L) \Rightarrow SVP_{\gamma}(L)$

The Gram-Schmidt orthogonalization method is critical in the LLL algorithm, we briefly recall what this method entails. Let $(V, \langle\cdot,\cdot\rangle)$ be a vector space endowed with a scalar product and $\|\cdot\|$ the l_2 -norm on V given by this scalar product. For any basis $(b_1, \ldots, b_n)$ we define inductively its orthogonalization $(b_1^*, \ldots, b_n^*)$ as

$$b_i^* = b_i - \sum_{j<i} \left\langle b_i, b_j^* \right\rangle b_j^*$$

This new basis $(b_1^*, \ldots, b_n^*)$ is orthogonal i.e. $\left\langle b_i^*, b_j^* \right\rangle = 0$ whenever $i \neq j$. More-over, $\operatorname{span}(b_1, \ldots, b_r) = \operatorname{span}(b_1^*, \ldots, b_r^*)$ for all $1 \leqslant r \leqslant n$. In the following, we will pay particular attention to the Gram-schmidt coefficients $\mu_{i,j} := \frac{\langle b_i, b_j \rangle}{\|b_j^*\|^2}$ for $i > j$. Moreover, let B^* be the matrix with columns the vectors $(b_1^*, \ldots, b_n^*)$ and L the lattice generated by $(b_1, \ldots, b_n)$. Then, it is readily seen that vol $(L) = \|b_1^*\| \cdot \ldots \cdot \|b_n^*\|$. (For vol $(L) = |\det(B^*)|$ as for all i we have $\operatorname{span}(b_1, \ldots, b_r) = \operatorname{span}(b_1^*, \ldots, b_r^*)$, but $\operatorname{diag}\left(\|b_1^*\|^{-1}, \ldots, \|b_n^*\|^{-1}\right) B^* \in O_n(\mathrm{R})$ so $\det(B^*) = \|b_1^*\| \cdot \ldots \cdot \|b_n^*\|$.) Now, assume that $\frac{\|b_i\|}{\|b_{i-1}\|}$ is bounded below by some constant $c > 0$. As $b_1 = b_1^*$, we get that

$$\|b_1\|^n \leqslant c^{-n(n-1)/2} \|b_1^*\| \cdot \ldots \cdot \|b_n^*\| = c^{-n(n-1)/2} \operatorname{vol}(L)$$

So b_1 is a solution to $HSVP_{c-(n-1)/2}(L)$.
In order to make use of this observation, we define the following condition.

Definition 1 (Lovász' condition). A basis $(b_1, \ldots, b_n)$ satisfies Lovasz' condition if there is $\delta \in \left(\frac{1}{4}, 1\right]$ such that for all $2 \leqslant i \leqslant n$

$$\frac{\|b_i\|^2}{\|b_{i-1}\|^2} \leqslant \delta - \mu(i, i-1)^2$$

Where $(b_1^*, \ldots, b_n^*)$ is the Gram-Schmidt orthogonalization of $(b_1, \ldots, b_n)$.

```
initialization;
Require: a basis $(b_1, \ldots, b_n)$ of $L$ and $\delta \in \left(\frac{1}{4}; 1\right]$
Ensure: the output basis $(b_1, \ldots, b_n)$ of $L$ satisfies
  Lovász' condition
$i \leftarrow 2$
while $i \leqslant n$ do
    $b_i \leftarrow b_i - \sum_{j<i} \mid \mu_{i,j} \rceil b_j^*$
    if $\|b_i^*\|^2 \geqslant \left(\delta - \mu_{i,i-1}^2\right) \|b_{i-1}\|^2$ then
        $i \leftarrow i + 1$
    else
        $swap\ b_i,\ b_{i-1};$
        $i \leftarrow \max(2, i-1);$
    end
end
```

Algorithm 1: LLL basis reduction algorithm

For the sake of simplicity, we removed the necessary parts that consist in updating the Gram-Schmidt orthogonalization and Gram-Schmidt coefficients. Therefore, the previous algorithm is only a sketch of the Lenstra-Lenstra-Lovász reduction algorithm. Let us now state the main result about this algorithm, namely its complexity and the properties of the output basis.

Theorem 7 Let $(b_1, ..., b_n)$ be a basis generating a lattice L and $\delta \in \left(\frac{1}{4}, 1\right]$. Given these as input, the LLL algorithm terminates in poly $\left(d, (1-\delta)^{-1}, \log(\max\|b_i\|)\right)$ and the output is a basis $(\beta_1, \ldots, \beta_n)$ generating L such that:

$$\frac{\|\beta_1\|}{\operatorname{vol}(L)^{1/n}} \leqslant \left(\frac{1}{\delta - \frac{1}{4}}\right)^{(d-1)/4}$$

In particular, for $\delta = 1 - \epsilon$ with e small, the LLL algorithm terminates in $poly\left(d, \frac{1}{\epsilon}, \log(\max\|b_i\|)\right)$ time and the output satisfies:

$$\frac{\|\beta_1\|}{\operatorname{vol}(L)^{1/n}} \leqslant \left(\frac{4}{3} + O(\epsilon)\right)^{(d-1)/4} \simeq 1.07^{d-1}$$

Empirically, it has been observed that for large n, that $\frac{\|\beta_1\|}{\operatorname{vol}(L)^{1/6}}$ is around 1.02^n which is a considerable improvement [15].

This result is about how well LLL solves the Hermite Shortest Vector Problem [12]. According to a remark made earlier, we obtain the following result concerning our main interest, the Shortest Vector Problem.

Corollary 1. With the same input as above, the output $(\beta_1, \ldots, \beta_n)$ satisfies:

$$\frac{\|\beta_1\|}{\operatorname{vol}(L)^{1/n}} \leqslant \left(\frac{1}{\delta - \frac{1}{4}}\right)^{(d-1)/2}$$

Or for $\epsilon > 0$ small and $\delta = 1 - \epsilon$ we get that:

$$\frac{\|\beta_1\|}{\operatorname{vol}(L)^{1/n}} \leqslant \left(\frac{4}{3} + O(\epsilon)\right)^{(d-1)/2} \approx 1.15^{d-1}$$

More precisely, we can get the following complexity.

Theorem 8 (Lenstra-Lenstra-Lovaisz, 1982,[L.LL 82]). The LLL algorithm has complexity $O\left(n^6 \log\max\|b_i\|\right)$.

Now, we briefly mention an improvement of LLL due to Schnorr and Euchner, see [16]. In the previous algorithm, when the swap function is applied, instead of swapping b_i and b_{i-1} we apply what Schnorr and Euchner call deep insertions. This entails to inserting b_i at the index j where j is the smallest index satisfying:

$$\delta \left\|b_j^*\right\| \geqslant \|b_i\|$$

The result of this change is shorter outputs but longer running time.

B. Block Korkine-Zolotarev Algorithm

The Block korkine-Zootarev (BKZ) algorithm is another algorithm solving SVP_γ. Even though it also runs in polynomial time, its purpose is to achieve better accuracy than LLL, therefore has longer running time. After breaking down how LLL algorithm works, we could see that it uses an SVP-oracle for lattices in dimension 2 called Gauss's reduction. Given more efficient oracles or higher dimensional SVP -oracles we could hope for a more efficient algorithm. This algorithm was proposed by Schnorr and Euchner in their paper [17].

Along this line of thought, given an SVP-oracle for all dimensions up to some integer k the BKZ algorithm finds a shorter vector than LLL does. Unsurprisingly, running BKZ takes more time than running LLL. Here, there is a clear tradeoff between the shortness of the output and the complexity of the algorithm. Moreover, as of today it is not known theoretically if BKZ terminates in polynomial time, only empirical results are known.

Here, an SVP-oracle is a mean by which we are able to solve the exact SVP. Many techniques were developed to solve SVP, but as one could guess from NP-hardness results, all these methods have at least exponential time complexity. Therefore, it is not deemed feasible to run these algorithms in high dimension. Among such oracles are enumeration algorihtms which are some kind of greedy algorithms that dates back to Pohst [18], sieving techniques which are applications of Monte-carlo methods (just as the next section is) due to Micciancio and Voulgaris in [19], and a method based on Voronoi cells due to the same authors in [20]. Even though these methods have exponential complexity, they can still be used for lattices in reasonably large dimension.

Let us now turn to the exposition of BKZ algorithm. To this end, we first need to define the notion of Korkine-Zolotarev reduced basis and a KZ-reduction algorithm. Suppose we have a SVP -oracle $\mathcal{O}$ up to dimension k and, for any basis $(b_1, \ldots, b_k)$ generating a lattice L define $\pi_i : \mathbb{R}^k \to \operatorname{span}(b_1, \ldots, v_i)^\perp$ as the orthogonal projection and $L_i = \pi_{i-1}(L)$. A basis $(b_1, \ldots, b_n)$ is called KZ-reduced if

$$||b_i^*|| = \lambda(L_i)$$

where $(b_1^*, \ldots, b_n^*)$ is the Gram-Schmidt reduction of $(b_1, \ldots, b_n)$. (Note that $\pi_i(b_i) = b_i^*$.) As for all i there is $\alpha_i \in [-1/2; 1/2]$ such that $\alpha_i b_i^\infty + b_{i+1}^* \in L_i$ we get by an induction argument the following theorem.

Theorem 9. Let $(b_1, \ldots, b_n)$ be a KZ -reduced basis. Then,

$$\|b_1\| = \lambda(L) \text{ and } \frac{\|b_1\|}{\|b_k\|} \leqslant k^{(1+\log(k))/2}$$

Moreover, we have the following KZ-reduction algorithm.

```
Result: Write here the result
initialization;
Require: a basis $B = (b_1, \ldots, b_n)$ for $L$ and a $SVP$
  -oracle $\mathcal{O}$ for up to $k$ dimensions.
Ensure: The output basis is KZ-reduced.
$m \leftarrow c(R)$
for $i = 1$ to $k$ do
    call $\mathcal{O}$ to find $b_i^* \in \pi_i(L)$ of length $\lambda(\pi_i(L))$
    lift $b_i^*$ into $b_i \in L$ such that $\|b_i^* - b_i\|$ is minimal;
    change $(b_{i+1}, \ldots, b_k)$ such that $(b_1, \ldots, b_k)$
      generates $L$
end
```

Algorithm 2: KZ-reduction algorithm

Note that step 4 may look as hard as solving CVP but is actually a lot easier thanks to the properties of orthogonal projections. Now, BKZ is as follows.

Algorithm 3 Schnorr and Euchner's BKZ-reduction algorithm
Require: a basis $B = (b_1, \ldots, b_n)$ for $L, \text{ak} \in \mathbb{N}, \delta \in \left(\frac{1}{4}, 1\right)$ and a SVP -oracle $\mathcal{O}$ for up to k dimensions.
Ensure: The output basis satisfies Lovisz' conditon with factor δ and $\|b_i^*\| = \lambda\left(\pi_{i-1}\left(\bigoplus_{j=1}^k b_{i+j-1}\right)\right)$ for i from 1 to $d-k+1$

```
initialization;
Require: a basis $B = (b_1, \ldots, b_n)$ for
  $L, \text{ak} \in \mathbb{N}, \delta \in \left(\frac{1}{4}, 1\right)$ and a $SVP$ -oracle $\mathcal{O}$ for up
  to $k$ dimensions.
Ensure: The output basis satisfies Lovisz' conditon
  with factor $\delta$ and $\|b_i^*\| = \lambda\left(\pi_{i-1}\left(\bigoplus_{j=1}^k b_{i+j-1}\right)\right)$
  for $i$ from 1 to $d-k+1$
repeat
    for $i = 1$ to $d-k+1$ do
        KZ-reduce the basis $\pi_{i-1}(b_i, \ldots, b_{i+k-1})$;
        Lift up to the closest vectors in $L$ and
          complete the basis.
    end
until no changes occur;
```

Algorithm 3: Schnorr and Euchner's BKZ-reduction algorithm

The theoretical complexity computation of this algorithm is not computed, but the following can be proved.

Theorem 10 Given a basis $(b_1, \ldots, b_n)$ generating L and an SVP oracle $\mathcal{O}$ for up to k dimensions, the BKZ algorithm outputs a basis $(\beta_1, \ldots, \beta_n)$ generating L that satisfies:

$$\frac{\|b_1\|}{\lambda(L)} \leqslant \left(k^{\frac{1+\log(k)}{2k-2}}\right)^{n-1} \text{ and } \frac{\|b_1\|}{\operatorname{vol}(L)^{1/d}} \leqslant \sqrt{\gamma_k}\left(k^{\frac{1+\log(\omega)}{2k-2}}\right)^{n-1}$$

C. Metropolis-Hasting's Algorithm

The next algorithm is a randomized algorithm applying the well-known Metropolis algorithm, a kind of Monte-Carlo method, to lattice problems. For a precise account on this method see for instance [21]. The following Metropolis algorithm is due to Ajitha, Biswas, and Kurur in [22]. This algorithm returns an approximate solution with respect to the euclidean norm, denoted by $\|\cdot\|$.

The search space is defined as follows.

Once again let $(b_1,\ldots,b_n)$ be a matrix of R^n, L the lattice it generates and B the matrix given by $(b_1,\ldots,b_n)$.

The main idea behind this Metropolis-Hastings is to work on a bounded set of vectors with integer entries and pair each of these vectors v to a cost $c(v) := \|Bv\|$. If $w \in L$ has norm less than or equal to k then $w = Bv$ where v has integer entries bounded above by $M = (\alpha n)^n$, where $\alpha = \max_{i,j}\left(|b_{ij}|, k\right)$. Thus, it is enough to look at vectors bounded above by M. As we want to modify the vector step by step, it is interesting to work with more than one vector at a time. Let us now define the Markov chain we will be working on.

Fix parameters $k \in \mathbb{R}, m \in \mathbb{N}$ and set M as above. The search space S consists of matrices $A' := [A \mid I]$ with integer entries bounded above by M where A is a $n \times m$ matrix and I is the $n \times n$ identity matrix. It remains only to describe the transition probability of the Markov chain. We first describe the neighbourhood of a given the state, the actual transition matrix will be described later on in the pseudo-code of the Metropolis-Hastings algorithm.

A matrix $S' = [S \mid I]$ is in the neighbourhood $N(R')$ of $R' = [R \mid I]$ if

(i) S is equal to R up to swapping two columns;
(ii) S is equal to R up to multiplying a column by -1
(iii) we can get from R' to S' by $r_i \leftarrow r_i \pm 2^l r_j$ for any $1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n+m, i \neq j$

Two interesting features of S are:

- For any $R' \in S, \#N(R') = O\left(m^2 \log(M)\right)$
- For any $R', S' \in S$, there is a path between R' and S' of length $O(mn\log(M))$

Let us now turn to the pseudo-code and the actual description of the transition probability.

Here, (p_i) is a probability distribution on $\mathbb{N}$ such that there is $n_0 \in \mathbb{N}$ satisfying $p_i = 0$ for all $i \geqslant n_0$. This algorithm seems to give better results than LLL for lattices in low dimensional vector spaces. However, in higher dimension such methods will have to deal with the following geometric issue, the so-called *curse of dimensionality*. Let $R > \epsilon > 0$ then

$$\frac{\mathrm{vol}\left(B_{\mathbb{R}^n}(R)\right)}{\mathrm{vol}\left(B_{\mathbb{R}^n}(R-\epsilon)\right)} \rightarrow_{n\rightarrow+\infty} 0$$

Therefore, as n goes to $+\infty$ 'most' of the states of the markov chain built in the Metropolis algorithm will lie in the annulus $B_{\mathbb{R}^n}(R) \backslash B_{\mathbb{R}^n}(R-\epsilon)$ and the markov chain might take more and more time before hitting any point inside $B_{\mathbb{R}^n}(R-\epsilon)$. For more on related heuristics see [23].

```
initialization;
Require: a basis $B = (b_1,\ldots,b_n)$ for $L$ and $K \in \mathbb{Q}$
Ensure: Matrix $R$ with integer entries such that $BR$
  has a column $c$ with $\|c\| \leqslant K$.
$m \leftarrow c(R)$
while  $m > K$ do
    Select $S$ a neighbour of $R$ by performing one of
      the following operations.
        • Swap two columns with probability $\frac{C_2^m}{\#N(R)}$;
        • Multiply a column by -1 with probability $\frac{m}{d}$
        • Add $2^i$ times a column of $R$ to another column of $R$
          with probability $\frac{d-C_3^m-m}{d} \cdot p_i$
    if *if* $m > c(R)$ then
        $m \leftarrow c(R)$;
    end
    if $S \in S$ then
        $R \leftarrow S$ with probability $\min\left(\frac{e^{-c(S)/T}}{e^{-c(R)/T}}, 1\right)$;
    end
end
```

Algorithm 4: Metropolis-Hasting's algorithm for SVP

Finally, note that a similar method is widely used to solve exact instances of SVP. The version of this algorithm used to solve SVP is called a sieve method and is often used as an SVP -oracle in the BKZ algorithm, see for instance the following paper by Micciancio and Voulgaris [19].

D. Convex Relaxation

This section presents the convex relaxation (see [24] for definitions and results on convex relaxation) to obtain another randomized algorithm to solve CVP_γ (thus, SVP_γ as well). As mentioned in the first part of this report, CVP_γ admits another formulation, more precisely $\mathrm{CVP}(L,c)$ is equivalent to

(1)
$$\begin{gathered} \text{minimize} \quad x^T B^T B x - 2c^T x \\ subject\ to\ x \in \mathbb{Z}^n \end{gathered}$$

.
Where B be the matrix given by the basis $(b_1,\ldots,b_n)$ and c a vector in $\mathbb{R}^n$. A convex relaxation consists in relaxing the hard condition $x \in \mathbb{Z}^n$ to a looser condition so that the problem considered is now a convex optimization problem. Here, the method is due to Park and Boyd in [25]. The convex relaxation is in this particular case an instance of semidefinite programming, for more on this topic see [26]. We first relax Problem 1 into a non-convex problem

(2)
$$\begin{gathered} \text{minimize} \quad x^T B^T B x - 2c^T x \\ \text{subject to } x_i(x_i - 1) \geqslant 0, \forall i \end{gathered}$$

Now, set $P := B^T B$ we can reformulate Problem 2

$$minimize Tr(PX) - 2c^T x$$

(3) subject to $\operatorname{diag}(X) \geqslant x$

$$X = xx^T$$

Finally, relax condition $X = xx^T$ of Problem 3 to $X \geq xx^T$ which means that $X - xx^T$ is semidefinite positive. But $X \geq xx^T$ is equivalent to $\begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \geq 0$. So we get the following relaxation.

(4)

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(PX) - 2c^T x \\ \text{subject to} & \operatorname{diag}(X) \geqslant x \end{array}$$

$$\begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \geq 0$$

Now, this is a semi definite relaxation that can be solved in polynomial time.

```
initialization;
Require: $P = B^T B, c \in \mathbb{R}^n, K \in \mathbb{N}$
Solve (4) to get $X^*$ and $x^*$
$\Sigma \leftarrow X^* - x^* x^{4T}$
Find Cholesky factorisation $LL^T = \Sigma$;
$x^{\text{best}} \leftarrow 0$;
$f^{\text{best}} \leftarrow 0$
for k=1,2, ..., K do
    $z^{(k)} \leftarrow x^* + Lw$ where $w \sim \mathcal{N}(0, I)$
    $x^{(k)} \leftarrow \text{round}\left(z^{(k)}\right)$
    if $f^{best} > f\left(x^{(k)}\right)$ then
        $x^{best} \leftarrow x^{(k)}$;
            $f^{best} \leftarrow f\left(x^{(k)}\right)$
    end
end
```

Algorithm 5: Randomized algorithm for suboptimal solution

Where the Cholesky factorisation is an algorithm computing the square root of a positive semidefinite matrix in $O\left(n^3\right)$ [27].

```
initialization;
Require: $x \in \mathbb{Z}^n, P = B^T B, c \in \mathbb{R}^n$.
$g \leftarrow 2(Px + c)$
repeat
    Find index $i$ and integer $a$ minimizing $a^2 P_{ii} + cg_i$;
    $x_i \leftarrow x_i + c_3$
    $g \leftarrow g + 2cP_i$
until $\operatorname{diag}(P) \geqslant |g|$;
```

Algorithm 6: Greedy descent algorithm

This last algorithm runs in polynomial time as well. Combining these two algorithms we get a approximate solution such that for all integer $a \in \mathbb{Z}$ and all index i, x is a better solution than $x + ae_i$

IV. Experimental Results

This section describes how the said approaches perform on benchmark instances. We see that the output of the LLL algorithm is near to the shortest vector which is $2^{\frac{n-1}{2}}$ of the shortest vector for the given lattice, it can be used for polynomial factoring etc. LLL algorithm complexity is $O\left(n^6 \cdot \log^3 \beta\right)$, where β is $\max_{1 \leq i \leq n} ||b_i||$. We ran tests to compare LLL algorithm and BKZ algorithm on data from [22], [21] and [24].

For experimental analysis a class of SVP instances are generated using the techniques developed by Richard Lindner and Michael Schneider [28]. They have given sample bases for Modular, Random, ntru, SWIFT and Dual Modular lattices of dimension 10. We have tested our code for all these instances and found that our algorithm works faster and gives shorter lattice vector when compared to LLL. The tested results are given in the Table I.

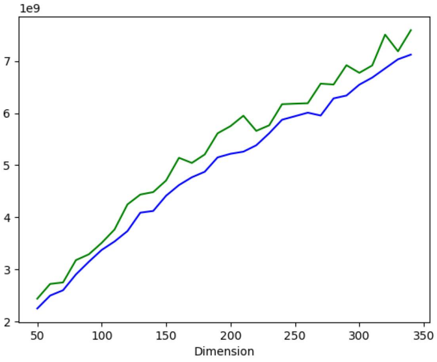


Fig. 1. Norm of the Output/ Dimension; Blue: BKZ / Green: LLL.

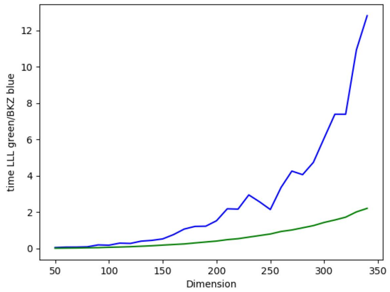


Fig. 2. Running time/ Dimension; Blue: BKZ / Green: LLL.

A. LLL/BKZ

Fig. 1 and Fig. 2 compare the basis reduction algorithms LLL and BKZ for a given family of Lattices parametrized by their dimension. Moreover, BKZ given an enumeration oracle for blocks of size 30. In particular, we see that with these parameters, the BKZ algorithm needs considerably more time to get a slightly shorter vector. To some extent, the LLL algorithm already seems to be efficient. A comparative results for running time of BKZ and LLL algorithms for varying dimension of lattice is given in Fig. 2.

TABLE I. RESULTS OF LLL AND MH FOR VARIOUS LATTICES

n	Input size	t (LLL: MH algo)	Norm (LLL: MH algo)
15	150	1234.58: 1147.2279	0.016: 201.651
20	8	3: 2.8284	0.2680: 0.052
25	8	1.73: 1.73	0.008: 0.004
30	8	4.123: 3.8729	0.008: 0.008
50	100	20.49: 8.66	0.108: 291.892

B. Metropolis-Hasting's Method

For a specific class of lattices [21], we have tested LLL and MH algorithms and the results are shown in Table I. Moreover, these data seem to support the curse of dimensionality heuristic as we can see the times needed to solve cases with large input size is high compared to small input size.

C. Convex Relaxation

The algorithm from [24] returns an approximate solution to CVP shown in Table II. To test their method, they computed solutions for instances randomly generated : all entries ar $\sim \mathcal{N}(0,1)$ then normalized so that the solution to problem (1) without the integer constraint has value -1. In what follows, n will denote the dimension, opt the percentage of outputs that were optimal solutions and t the time it took to get this output. All values are averages.

TABLE II. RESULTS FOR CVP FOR RANDOM INSTANCES

n	t	opt
50	0.397	90%
60	0.336	94%
70	0.402	89%

Because of the NP -hardness of CVP it is hard to compute the quantity opt in higher dimensions. However, the break down of running times of the method for larger $n's$ presented in Table III.

TABLE III. RUNNING TIMES OF THE CONVEX RELAXATION

n	t_{total}	SDP	Random sampling	Greeedy 1-opt
50	0.397	0.296	0.065	0.036
60	0.336	0.201	0.084	0.051
70	0.402	0.249	0.094	0.058
100	0.690	0.380	0.193	0.117
500	20.99	12.24	4.709	4.045
1000	135.1	82.38	28.64	24.07

In spite of the theoretical running time being $O\left(n^3\right)$ we can note that the total running time seems to grow subcubically.

V. CONCLUSION

In this paper we have discussed Shortest Vector Problem, Closest Vector Problem and their average case and worst case hardness results. Further, this work presented solving SVP using the LLL algorithm, BKZ algorithm, a Metropolis algorithm and a convex relaxation. We have compared the performance of these algorithms on various lattices by varying input sizes and the results we have obtained are fairly encouraging. The experimental results on various lattices shows that Metropolis algorithm works better than other algorithms.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments.

REFERENCES

[1] M. Ajtai, "Generating hard instances of lattice problems," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 99–108.

[2] ——, "The shortest vector problem in l2 is np-hard for randomized reductions," in *Proceedings of the 30th annual ACM symposium on Theory of computing*, 1998, pp. 10–19.

[3] M. Ajtai and C. Dwork, "A public-key cryptosystem with worst-case/average-case equivalence," in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, 1997, pp. 284–293.

[4] O. Goldreich, S. Goldwasser, and S. Halevi, "Public-key cryptosystems from lattice reduction problems," in *Proceedings of the Annual International Cryptology Conference*. Springer, 1997, pp. 112–131.

[5] J. Hoffstein, J. Pipher, and J. H. Silverman, "Ntru: A ring-based public key cryptosystem," in *Proceedings of the International Algorithmic Number Theory Symposium*. Springer, 1998, pp. 267–288.

[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 169–178.

[7] J.-y. Cai, "The complexity of some lattice problems," in *Proceedings of the International Algorithmic Number Theory Symposium*. Springer, 2000, pp. 1–32.

[8] J. C. Lagarias, "The computational complexity of simultaneous diophantine approximation problems," *SIAM Journal on Computing*, vol. 14, no. 1, pp. 196–209, 1985.

[9] S. Arora, L. Babai, J. Stern, and Z. Sweedyk, "The hardness of approximate optima in lattices, codes, and systems of linear equations," *Journal of Computer and System Sciences*, vol. 54, no. 2, pp. 317–331, 1997.

[10] D. Micciancio, "The shortest vector in a lattice is hard to approximate to within some constant," *SIAM journal on Computing*, vol. 30, no. 6, pp. 2008–2035, 2001.

[11] I. Dinur, G. Kindler, R. Raz, and S. Safra, "Approximating cvp to within almost-polynomial factors is np-hard," *Combinatorica*, vol. 23, no. 2, pp. 205–243, 2003.

[12] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische annalen*, vol. 261, pp. 515–534, 1982.

[13] H. Minkowski, "U rather the positive square forms and "u over chain break ä similar algorithms," *Journal f "u r pure and applied mathematics (Crelles Journal)*, vol. 1891, no. 107, pp. 278–297, 1891.

[14] R. Kannan, "Minkowski's convex body theorem and integer programming," *Mathematics of operations research*, vol. 12, no. 3, pp. 415–440, 1987.

[15] N. Gama and P. Q. Nguyen, "Predicting lattice reduction," in *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2008, pp. 31–51.

[16] C.-P. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Mathematical programming*, vol. 66, no. 1, pp. 181–199, 1994.

[17] ——, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Mathematical programming*, vol. 66, no. 1-3, pp. 181–199, 1994.

[18] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications," *ACM Sigsam Bulletin*, vol. 15, no. 1, pp. 37–44, 1981.

[19] D. Micciancio and P. Voulgaris, "Faster exponential time algorithms for the shortest vector problem," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2010, pp. 1468–1480.

[20] ——, "A deterministic single exponential time algorithm for most lattice problems based on voronoi cell computations," *SIAM Journal on Computing*, vol. 42, no. 3, pp. 1364–1391, 2013.

[21] C. P. Robert and G. Casella, "The metropolis—hastings algorithm," in *Monte Carlo statistical methods*. Springer, 2004, pp. 267–320.

[22] S. K. Ajitha, S. Biswas, and P. P. Kurur, "Metropolis algorithm for solving shortest lattice vector problem (svp)," in *Proceedings of 11th International Conference on Hybrid Intelligent Systems (HIS)*. IEEE, 2011, pp. 442–447.

[23] T. Carson, *Empirical and analytic approaches to understanding local search heuristics*. University of California, San Diego, 2001.

[24] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[25] J. Park and S. Boyd, "A semidefinite programming method for integer convex quadratic minimization," *Optimization Letters*, vol. 12, no. 3, pp. 499–518, 2018.

[26] L. Lovász, "Semidefinite programs and combinatorial optimization," in *Recent advances in algorithms and combinatorics*. Springer, 2003, pp. 137–194.

[27] L. N. Trefethen and D. Bau III, *Numerical linear algebra*. Siam, 1997, vol. 50.

[28] "The sage development team, s. reference v4.7," *Cryptography, www.sagemath.org/doc/reference/ sage/crypto/lattice.html*, 2009.

Black-box Fuzzing Approaches to Secure Web Applications: Survey

Aseel Alsaedi[1], Abeer Alhuzali[2], Omaimah Bamasag[3]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Web applications are increasingly important tools in our modern daily lives, such as in education, business transactions, and social media. Because of their prevalence, they are becoming more susceptible to different types of attacks that exploit security vulnerabilities. Exploiting these vulnerabilities may cause damage to the web applications as well as the end-users. Thus, web apps' developers should identify vulnerabilities and fix them before an attacker exploits them. Using black-box fuzzing techniques for vulnerability identification is very popular during the web apps' development life cycle. These techniques pledge to find vulnerabilities in web applications by constructing attacks without accessing their source codes. This survey explores the research that has been done in the black-box vulnerability finding and exploits construction in web applications and proposes future directions.**

***Keywords*—Black-box fuzzing; web application security; vulnerability scanning; automatic web app testing; vulnerability detection; survey**

I. Introduction

Web applications are significant components in various fields: commercial, banking, entertainment, education, healthcare, and social networking. To reach international markets, organizations have utilized web applications to promote their products/goals and provide end-user services. In recent years, web applications have evolved rapidly. This rapid development has led to the use of different technologies and libraries, which results in more complex and feature-full applications. Many application developers do not have the necessary security skills that prevent them from writing buggy code. That, in turn, allows attackers to exploit those vulnerabilities and may cause damage. Thus, the security of web applications is of paramount concern.

Verizon's 2019 report [1] stated that most data breaches occurred due to attacks on web applications. Two out of three of all the examined data breaches were in web applications. Another report on web application vulnerability [2] analyzed nearly 5,000 different web apps from March 2019 until February 2020 and found that 26% of these applications have critical vulnerabilities. The remaining apps had medium-level vulnerabilities. All these studies indicate that vulnerabilities are prevalent in web applications.

Typically, security staffs use manual analysis to identify vulnerabilities to fix. This process is time-consuming and error-prone, as it depends on the expertise level of the analyst, and it is challenging because of the increasing complexities of web applications. Additionally, manual analysis is difficult to perform in each new update on applications that can lead an attacker to exploit vulnerabilities.

Many automated techniques for vulnerability analysis and exploit constructions have been proposed. These approaches aim to reduce the cost, time, effort, and precision during testing. Due to these benefits, these types of approaches have become a hot topic in recent years, especially when applications tend to be complex. Broadly, these approaches can be categorized into white-box, and black-box fuzzing approaches. The white-box testing is based on examining the source code and the behavior of a web application to find security vulnerabilities. Several studies utilized this technique to identify critical vulnerabilities in web applications such as [3], [4], [5], [6], [7], [8], and [9]. Unfortunately, this type of approach cannot apply to a wide range of web apps; many times, it is specialized for a particular programming language. Additionally, the source code of the web app is not available at each time. On the other hand, the black-box fuzzing is a vulnerability-analysis and exploits construction approach that automatically discovers vulnerabilities and generates exploits without accessing the source code. Compared with white-box, the major benefit of using black-box fuzzing is that it is fast and efficient, and it can find security bugs in any web application, regardless of its implementation details. Thus, this technique applies to a wide range of web applications. Additionally, the black-box fuzzing approach is helpful for developers who have little or no experience in writing secure source code.

As a result of the pressing need to protect web applications without accessing the source code, a significant research effort has been geared towards developing many techniques for detecting web applications using the black-box fuzzing approach. Much of this research addresses a specific class of vulnerabilities or delivers fewer false positives, such as [10] and [11].

As such, the primary goal for this survey is to analyze the last ten years of existing bug-finding techniques in web applications , focusing on the black-box fuzzing approach. Further, this survey contributes towards identifying the challenges of the black-box fuzzing approaches, which aids the research community in determining where further research can be performed and provides valuable insights for improving the crawling module. Because of our goal, this survey intends to answer the following questions:

1) What are the techniques utilized by the approach?
2) Is the approach applicable to be used in modern web applications?
3) How does the approach construct benign inputs needed by web applications to explore further and test the application?

This paper is organized as follows. Section 2 provides an overview of the web application and its common vulnerabilities. Section 2 describes a black-box fuzzing approach, and Section 4 analyzes and compares existing approaches. Section 5 discusses the results by identifying current techniques' main weaknesses and then identifying potential research areas. Finally, Section 6 concludes our review.

II. Web Applications' Architecture and Common Vulnerabilities

A. Web Application Architecture and Characteristics

The traditional web applications have a three-tier architecture: client-side, server-side, and back-end databases that provide and store the web application data. The client-side part is executed on the user's web browser, allowing the user to interact and communicate with the web application via user inputs, links, etc. The client-side code is written using different technologies such as HTML, JavaScript, and CSS. On the other hand, the server-side is executed on the web server, responding to the client messages and managing the business logic. Server-side code is commonly written in PHP, Java, and so on. The client-side and server-side communicate with each other through messages, which are HTTP requests and responses. To demonstrate how a web application works, Fig. 1 shows a high-level view of how a typical web application works when communicating with the server-side to register a new user. It can be summarized as follows:

1) The user opens the web page, which its browser can render, and fills out all the necessary inputs to complete the registration. Filling out inputs on web forms provides interactions between users and browsers that enable users to deal with different input types, such as numeric and text. Once the user has completed the fill-in and submitted the form (usually by clicking the submit button), the inputs are encapsulated into an HTTP request and sent to the server.
2) The server-side receives the client request and then processes it.
3) The server-side sends the web form data to the database as a query to add the user to the web application list.
4) The server-side replies to the user as an HTTP response to the present result.

Today, web applications are getting more complex and dynamic because of the adaptation of different technologies such as AJAX (**A**synchronous **J**avaScript **A**nd **X**ML). The web form in our example, as shown in Fig. 2 can dynamically update part of the page to prevent unwanted requests when the data involves errors. It can also change the page's display contents dynamically without waiting for the server-side to deliver a new HTML page.

B. Web Applications' Vulnerabilities

Web applications play critical roles in our lives and are used in various activities and services. Typically, web applications deal with private or sensitive data, which becomes a valuable target for attackers. As shown in Fig. 2, a web application accepts potentiality dangerous inputs since the entered user inputs may include malicious code that harms the application. There are many types of vulnerabilities; in the following, we will focus solely on the most common vulnerabilities.

1) Cross-Site Scripting: XSS is a vulnerability executed on the client-side of the web application. The XSS is a code injection that enables the attacker to execute a malicious script (e.g., JavaScript code) in the victim's browser. The exploitation of XSS vulnerability is very dangerous, according to OWASP [12], because the XSS enables the attacker to modify web pages and steal sensitive information such as cookies, session tokens, and users' credentials. There are two classes of XSS attacks: reflected and stored.

- **Reflected XSS** occurs when the attacker successfully injects a malicious script into an HTTP request. The victim browser receives an HTTP response, including a malicious script, and executes it.
- **Stored XSS** occurs when an attacker injects a malicious script in content such as post comments and store permanently, often on the database. Suppose that malicious content is retrieved from the database without filtering. In that case, the malicious code will be executed on the victim's browser at all times when any user visits the infected page.

2) SQL Injection: SQLI is ranked as the most common injection vulnerability on web applications according to OWASP [12]. It occurs when an attacker manipulates the original logic, semantic, or syntax of an SQL query by using specially designed inputs such as SQL keywords or operators into original queries, which aims to control the back-end databases of the web application. There are two types of SQLI attacks: first-order SQLI and second-order SQLI.

- **First-order SQLI** The malicious queries are loaded and executed directly on the database.
- **Second-order SQLI** The crafted input is inserted into the database without sensitization. After that, that malicious content is retrieved without sanitization, which will allow the execution of malicious content.

III. Black-Box Fuzzing Approach as Defensive Technique for Protecting Web Applications

A. Typical Scenario

The black-box scanner consists of three primary modules: the crawling module, attack module, and analysis module. Crawling is a fundamental component in web application scanners which explores the applications that determine the scanner's capability to identify vulnerabilities. If a vulnerability scanner can discover subtle vulnerabilities on the application's deep locations, this indicates that the scanner has an effective crawling component. To understand how black-box fuzzing works, we will use the following example.

Envision a simple web application with a home page, registration page, and course-view page. As shown in Fig. 3, the user can access the course-view page only after completing the registration process. The crawler starts with a seed URL and extracts all reachable pages. The crawler identifies entry points and analyzes web forms to assign input values

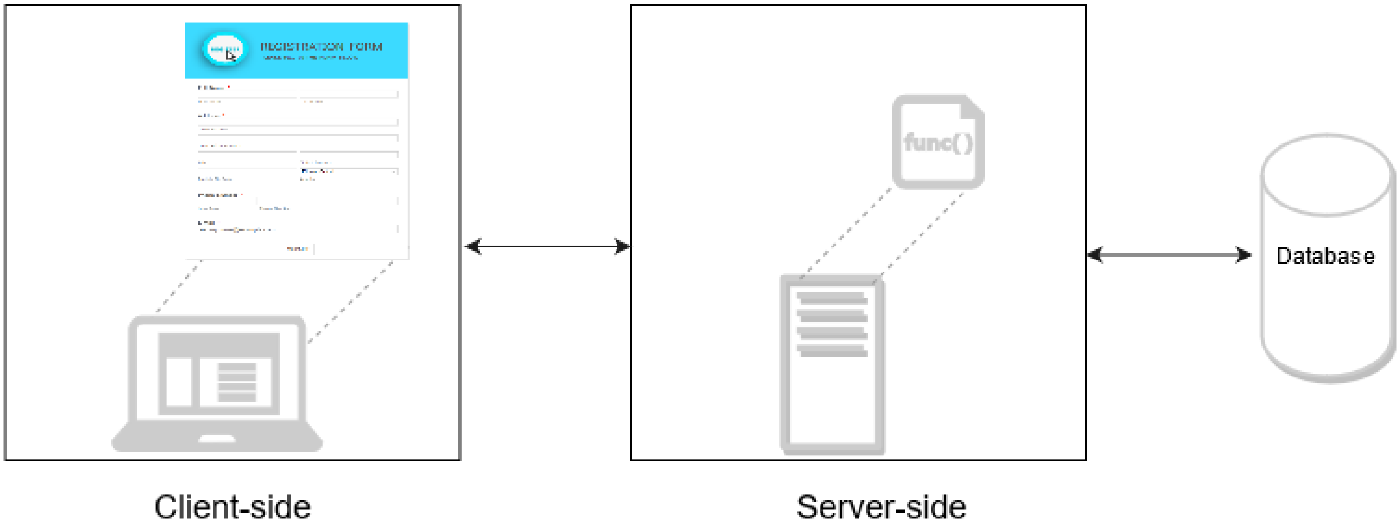
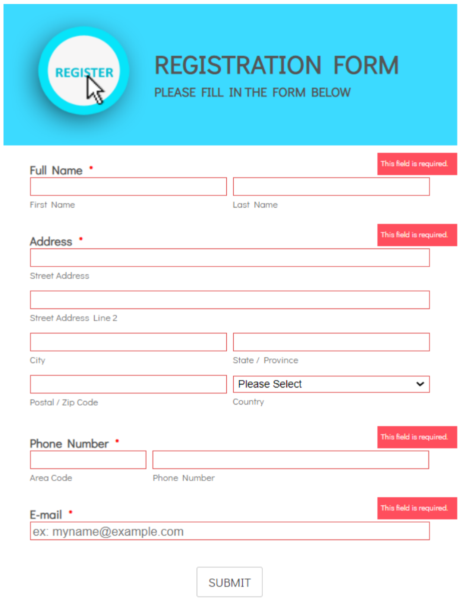


Fig. 1. Web Applications Architecture.



REGISTER

REGISTRATION FORM

PLEASE FILL IN THE FORM BELOW

Full Name * — This field is required.
First Name | Last Name

Address * — This field is required.
Street Address
Street Address Line 2
City | State / Province
Postal / Zip Code | Please Select — Country

Phone Number * — This field is required.
Area Code | Phone Number

E-mail * — This field is required.
ex: myname@example.com

SUBMIT

Fig. 2. Web form Input Validation.

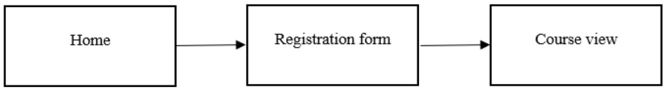


Fig. 3. Describes the Navigation Graph of the Example.

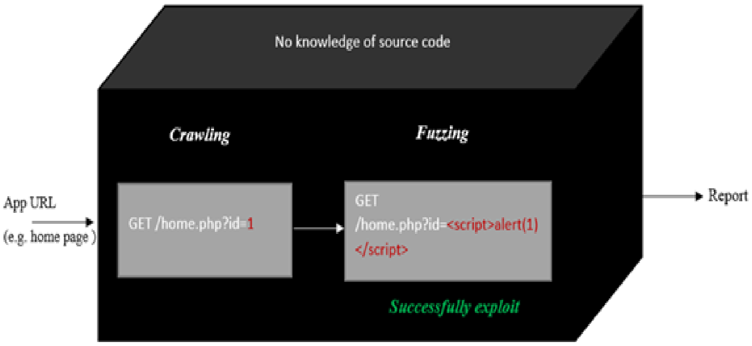


Fig. 4. Black-box Fuzzing Approach.

to discover several links on the application [13]. Then, the attack module fuzzes the application using information learned from the crawler. In other words, the attack module produces attack strings for each input or entry points to fuzz the web application. Finally, the analysis module analyzes the response page after launching the attack to determine where the attack string is reflected in the web application and reports the vulnerabilities [13]. In Fig. 4, the analysis module checks if the script code of the XSS attack is executed.

B. Automated Tools

Many black-box vulnerability scanners, commercial and open-source, are available, all of which have unique characteristics. Here, we review some black-box scanners.

Acunetix [14] is a commercial tool for automatic scanning web applications. It crawls web applications, even the AJAX-heavy ones. It provides different technologies, such as AcuSensor, to increase coverage and obtain higher accuracy for some vulnerabilities, such as accessing the application's source code. It discovers a wide range of vulnerabilities, such as XSS, SQLI, and so on.

AppSpider [15] is a commercial tool to find security bugs automatically. It complies with sophisticated technologies of modern web applications such as AJAX. It has discovered more than 95 attacks, including the OWASP Top 10 vulnerabilities [12]. It can construct values for web forms based on test cases and modify them to belong to different languages and more values.

Burp Suite Professional [16] is a commercial tool that web application scanners provide as part of their features. The scanner covers the most vulnerabilities on OWASP [12], such as XSS, SQLI, XPath, and so on. It has an advanced crawler that discovers vulnerabilities behind advanced JavaScript and dynamic features. It also constructs inputs for forms randomly and uses an intelligent manner to determine what fields are encountered on the page. It provides the ability to tune the configuration of the security test.

Nessus Professional [17] is a commercial vulnerabilities scanner. It can identify vulnerabilities accurately because of its high crawling coverage adapting different technologies to support modern web applications, including Ajax. Additionally, it scans at high speed and has very low false positives.

Netsparker [18] is a commercial web-vulnerability scanner that aims to identify different types of vulnerabilities such as XSS, SQLI, and so on. The crawler can handle different technologies, regardless of their complexity, platform, or architecture. Hence, it can be used for scanning modern web applications. The power of the tool is developed to identify vulnerabilities and gain nearly zero false positives precisely.

Grabber [19] is a free and open-source web-application-vulnerability scanner. It discovers security bugs involving XSS, SQLI, file inclusion, backup files, simple Ajax, and JS source-code analyzer. It scans the web application automatically by identifying the application's entry points into which the data can be injected. It is suitable to use for scanning small web applications.

Vega [20] is a free and open-source web-application scanner. It discovers many vulnerabilities, including SQLI, XSS, remote file inclusion, shell injection, and so on. The Vega crawler can extract URLs and forms and automatically constructs inputs for web applications. Further, it gives the user an additional feature to work as a proxy and be semi-automated to maximize the crawling coverage of web applications.

W3AF (**W**eb **A**pplication **A**ttack and **A**udit **F**ramework) [21] is a free and open-source vulnerability scanner. It discovers different types of vulnerabilities, including XSS and SQLI. Further, the crawler module extracts URLs and forms, and it uses an intelligent manner to give a fake value based on developers' knowledge of the tool.

Wapiti [22] is a free and open-source web-vulnerability scanner. It identifies various vulnerabilities, such as XSS, XXE, XPath, SQLI, and so on. The tool works in this way: the crawler extracts the URLs and forms from HTML, Flash, and basic JavaScript. After that, it launches a payload of attacks on the scripts and forms gained from the crawler to determine vulnerable locations. Finally, it generates reports to show the vulnerabilities and their locations. The crawler support fills form through some placeholder values that are consistent with most policies for web applications.

WFuzz (the **W**eb **Fuzz**er) [23] is a free and open-source web-application-vulnerability scanner. It is designed to perform brute-force attacks against web applications. It can find the essential web pages that are difficult to reach through normal browsing. It also discovers XSS, SQLI, and XXE attacks on GET and POST parameters. Table I shows an overview of the existing web application scanners in the black-box fashion.

TABLE I. SUMMARY OF THE CHARACTERISTICS OF BLACK-BOX WEB APPLICATION SCANNERS

Name	License	Price starts at	Dynamic features
Acunetix [14]	commercial	$4,500	✓
AppSpider [15]	commercial	$2000	✓
Burp Suite Pro. [16]	commercial	$399	✓
Nessus Pro. [17]	commercial	$3,438.50	✓
Netsparker [18]	commercial	●	✓
Grabber [19]	open-source	N/A	⋆
Vega [20]	open-source	N/A	×
w3af [21]	open-source	N/A	×
Wapiti [22]	open-source	N/A	×
WFuzz [23]	open-source	N/A	×

● Not Provided by the Vendor
⋆ Partially Support

IV. BLACK-BOX FUZZING RESEARCH APPROACHES

To develop an effective black-box fuzzing approach, the increase in crawling coverage has become significant. Thus, this section analyzes the existing primary research in this domain and answers the previously mentioned questions.

Question 1: What are the techniques utilized by the approach?

Many approaches have been used to improve the coverage of the crawler to detect web applications in a black-box fashion. By answering this question, it will be possible to identify the approach and determine if there are any contributions or limitations of these approaches, as the following:

Bisht et al. [24] proposed an approach to detect server-side parameter tampering vulnerabilities in web applications. It is based on extracting constraints from the web forms (i.e., client-side) and uses constraint solving technology to generate test cases that expose the parameter tampering opportunities.

Doupé *et al.* [25], these authors proposed an approach that aimed to enhance the crawling of black-box scanners to discover a wide range of vulnerabilities. Their approach is based on capturing the changes in the application states and using the changes discovered to enhance the crawling coverage. However, their approach does not handle dynamic contents implemented by AJAX. Additionally, it does not support enhanced form submission.

Djuric [26] developed a tool called SQLIVDT to generate SQLI against web applications. His approach uses two types of crawling, which are automated and manual, to identify all the gateways to web applications and can be used to execute SQLI. However, this approach handles dynamic features manually via proxy.

Li and Xue [27] capture web applications' behavior as a finite state machine (FSM) and discover logic-flow vulnerabilities from the difference between the FSM of the expected behavior of an application without bugs and the actual FSM. However, their approach cannot handle dynamic contents.

Pellegrino *et al.* [28] proposed a semi-automated scanner that aims to expand the code coverage of web applications. It uses dynamic analysis on the client-side code to handle JavaScript-based web applications. As a result of this work, the author improved the crawling coverage by more than 86% compared to other tools. However, this research's primary goal is to analyze the client-side of dynamic web applications without adept form submission. Moreover, this approach is limited to finding XSS vulnerabilities.

Muñoz *et al.* [29] proposed a novel approach to maximize the crawler's code coverage. Their approach is based on analyzing web forms to extract fields that fill with appropriate values from external sources. In general, this approach cannot guarantee that the application's server-side will accept the constructed input values.

Deepa *et al.* [30] proposed *DetLogic* to improve the crawler by discovering the logic vulnerabilities in web applications. *DetLogic* acts as a proxy between the client-side and server-side of the applications that utilizes the information coming from the server to model web application behavior as a finite state machine (FSM). Logical constraints are then constructed from the FSMs to launch attacks. However, their approach is more applicable to static web applications rather than modern web applications. Additionally, this approach does not handle modern web forms that include certain restrictions on the form inputs.

Deepa *et al.* [31] proposed an approach to detect XQuery and parameter-tampering vulnerabilities in XML-database-based applications. Their approach aims to reduce false alarms and expand the coverage of crawlers. To achieve more web exploration, it analyzes the client-side code to handle the constraints of the web forms. However, their approach is limited to find few types of vulnerabilities.

Koswara *et al.* [32] developed *W3AF+* that extends their basic functionality from *W3AF*[21]. The authors developed a traditional crawler to adapt their method to handle dynamic web applications (e.g., Ajax applications). Their method depends on recording changes of the inner states on the applications. They handle the dynamic application by extending their crawler via capturing the changes caused by the event generator and saving the new state of the resulting DOM in the original state machine when it differs from the current state. However, their method is restricted to *on-click* event.

Liu *et al.* [33] developed a tool to increase the crawler coverage of the scanner via filling the required information, such as login forms, through giving correct values by the user when the crawler gives the user an instruction to complete. However, this approach cannot guarantee that the application's server-side code will accept the constructed input values.

Aliero *et al.* [34] proposed an SQLIVS tool that aims to maximize the coverage of crawlers to find subtle vulnerabilities and minimizes false alarms. Their approach is based on analyzing different HTTP responses to determine the presence of SQLI vulnerabilities.

Eriksson *et al.* [35] proposed a tool, *Black Widow*, to maximize the code coverage of black-box-vulnerability scanners in web applications. Their approach is based on capturing the application's inner state to identify the sinks and sources and enhancing the navigation model to navigate more dynamic workflows. However, their approach is limited to finding XSS vulnerability and construct inputs on forms randomly.

Question 2: Is the approach applicable to be used in modern web applications?

Generally, modern web applications have many dynamic features that appear when rendering applications at runtime. These features can generate a change in the pages' DOM, which includes generating contents dynamically, such as forms, links, and so on. Consequently, these DOM changes can affect the navigation graph of the web application, which may impact finding the vulnerable path and, therefore, lead to miss vulnerabilities. Table II shows a comparison of the approaches regarding the adaptation of dynamic features on their crawling.

TABLE II. Comparison of the Approaches Regarding Adaptation Dynamic Features

Approach	Dynamic features
[24]	×
[25]	×
[26]	✓
[27]	×
[28]	✓
[29]	✓
[30]	×
[31]	×
[32]	✓
[33]	✓
[34]	×
[35]	✓

Question 3: How does the approach construct benign inputs needed by web applications to explore further and test the application?

Traditionally, constructing inputs for web forms that mimic user interaction with web applications has been done mostly manually or randomly. Filling forms enable users to deal with different types of inputs, such as numeric, text, etc. This variety of input restrictions makes the automatic construction of the correct inputs challenging. Thus, web application analyses infer the required inputs to be generated to explore the application's workflow and find more vulnerabilities. Table III shows a comparison of the approaches regarding constructing inputs on their crawling.

TABLE III. Comparison of the Approaches Regarding Construct Inputs on their Crawling

Approach	Input generation
[24]	constraint solver
[25]	database (external sources)
[26]	user
[27]	N/A
[28]	N/A
[29]	external sources
[30]	constraint solver
[31]	constraint solver
[32]	N/A
[33]	user
[34]	N/A
[35]	random

V. Discussion

Many researchers focus their efforts on discovering security bugs on the web app before attackers exploit them. In Section IV, we discussed several research efforts tackling the challenge of code coverage on the black-box fuzzing approach. We found that the crawling module of the black-box fuzzing approach is a significant challenge in modern web applications. The first limitation is that many existing research systems still suffer from shallow coverage of the black-box fuzzing approach due to difficulties in handling the dynamic features of modern web apps. Unless the crawling module of the black-box fuzzing approach can support dynamic features such as JavaScript, it is difficult for that fuzzer to analyze most modern web applications that may include critical vulnerabilities behind elements generated on the fly. As shown in Table II, the results indicate that there are no significant differences between studies on supporting dynamic features. In fact, most of the previous studies that handle contents generated dynamically are limited to specific types of events. The second limitation that has hampered crawling involves the design of web forms. Input validation with its restrictions makes it difficult for an automatic generation to fill inputs correctly and submit them to the server-side code. Due to this limitation, the scanners fail to find vulnerabilities in deep locations in web application structures. As shown in Table III, several techniques used to construct input on forms, which do not guarantee that the values will be accepted by the server code, for all previous techniques except [24], [30] and [31]. To precisely infer valid inputs, the study [24], [30], [31], [8], and [9] use constraint solver to construct inputs on the forms by deriving constraints of HTML form inputs.

As a result, the black-box fuzzing approach can achieve better performance when considering the inner state of applications, dynamic features, and input generation. An intriguing research direction is exploring how to combine existing techniques to overcome these limitations and produce a fully automatic scanner in a black-box fashion.

VI. Conclusion

This paper reviews recent and existing techniques regarding black-box fuzzing to discover vulnerabilities. Our survey shows that new technologies and programming capabilities have accelerated the evolution and the complexity of web applications. As a result, automatically analyzing web apps for security purposes becomes more challenging. Additionally, we found that the crawling module of black-box fuzzing approaches still suffers from shallow coverage, which affects identify vulnerabilities on dynamic web applications. Improved crawling module of the black-box fuzzing approach to include dynamic features is necessary to assist the crawler in discovering more links and, therefore, find more vulnerabilities. Further, it is necessary to consider the practical approaches for input generation that can analyze web forms and infer their restrictions to generate correct input values automatically.

As far as these issues, the central problem of the black-box fuzzing approaches is how to deal with these challenges to enhance performance, which leads us to propose another area where black-box scanners should be improved.

References

[1] Verizon, "White paper: 2019 data breach investigations report," Verizon Business, Tech. Rep., 2019. [Online]. Available: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf

[2] Acunetix, "White paper: Acunetix web application vulnerability report 2020," Acunetix, Tech. Rep., 2020. [Online]. Available: https://www.acunetix.com/acunetix-web-application-vulnerability-report/

[3] M. C. Martin and M. S. Lam, "Automatic generation of xss and sql injection attacks with goal-directed model checking." in *17th {USENIX} Security Symposium*, 2008, pp. 31–44.

[4] G. Wassermann, D. Yu, A. Chander, D. Dhurjati, H. Inamura, and Z. Su, "Dynamic test input generation for web applications," in *Proceedings of the 2008 international symposium on Software testing and analysis*, 2008, pp. 249–260.

[5] A. Kieyzun, P. J. Guo, K. Jayaraman, and M. D. Ernst, "Automatic creation of sql injection and cross-site scripting attacks," in *2009 IEEE 31st international conference on software engineering*. IEEE, 2009, pp. 199–209.

[6] P. Bisht, T. Hinrichs, N. Skrupsky, and V. N. Venkatakrishnan, "Waptec: Whitebox analysis of web applications for parameter tampering exploit construction," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, ser. CCS '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 575–586. [Online]. Available: https://doi.org/10.1145/2046707.2046774

[7] S.-K. Huang, H.-L. Lu, W.-M. Leong, and H. Liu, "Craxweb: Automatic web application testing and attack generation," in *2013 IEEE 7th International Conference on Software Security and Reliability*. IEEE, 2013, pp. 208–217.

[8] A. Alhuzali, B. Eshete, R. Gjomemo, and V. Venkatakrishnan, "Chainsaw: Chained automated workflow-based exploit generation," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 641–652.

[9] A. Alhuzali, R. Gjomemo, B. Eshete, and V. Venkatakrishnan, "{NAVEX}: Precise and scalable exploit generation for dynamic web applications," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 377–392.

[10] M. N. khalid, M. Iqbal, M. T. Alam, V. Jain, H. Mirza, and K. Rasheed, "Web unique method (wum): An open source blackbox scanner for detecting web vulnerabilities," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.081254

[11] F. Duchene, S. Rawat, J.-L. Richier, and R. Groz, "Kameleonfuzz: evolutionary fuzzing for black-box xss detection," in *Proceedings of the 4th ACM conference on Data and application security and privacy*, 2014, pp. 37–48.

[12] OWASP TOP 10, "Owasp top ten web application security risks — owasp," 2021. [Online]. Available: https://owasp.org/www-project-top-ten/

[13] A. Doupé, M. Cova, and G. Vigna, "Why johnny can't pentest: An analysis of black-box web vulnerability scanners," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2010, pp. 111–131.

[14] acunetix. Acunetix web vulnerability scanner. [Online]. Available: https://www.acunetix.com/

[15] Rapid7. Appspider - application scanner. [Online]. Available: https://www.rapid7.com/products/appspider/

[16] PortSwigger. Burp suite - application security testing software. [Online]. Available: https://portswigger.net/burp

[17] Tenable. Nessus professional vulnerability assessment. [Online]. Available: https://www.tenable.com/products/nessus

[18] Netsparker. Netsparker - web vulnerability scanner. [Online]. Available: https://www.netsparker.com/

[19] R. Gaucher. Grabber. [Online]. Available: http://rgaucher.info/beta/grabber/

[20] Subgraph. Vega vulnerability scanner. [Online]. Available: https://subgraph.com/vega/index.en.html

[21] w3af. w3af - web application attack and audit framework. [Online]. Available: http://w3af.org/

[22] N. Surribas. Wapiti web application scanner. [Online]. Available: https://wapiti.sourceforge.io/

[23] C. Martorella, C. del ojo, and X. Mendez. Wfuzz - the web fuzzer. [Online]. Available: https://wfuzz.readthedocs.io/en/latest/index.html

[24] P. Bisht, T. Hinrichs, N. Skrupsky, R. Bobrowicz, and V. Venkatakrishnan, "Notamper: automatic blackbox detection of parameter tampering opportunities in web applications," in *Proceedings of the 17th ACM conference on Computer and communications security*, 2010, pp. 607–618.

[25] A. Doupé, L. Cavedon, C. Kruegel, and G. Vigna, "Enemy of the state: A state-aware black-box web vulnerability scanner," in *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, 2012, pp. 523–538.

[26] Z. Djuric, "A black-box testing tool for detecting sql injection vulnerabilities," in *2013 Second International Conference on Informatics & Applications (ICIA)*. IEEE, 2013, pp. 216–221.

[27] X. Li and Y. Xue, "Logicscope: automatic discovery of logic vulnerabilities within web applications," in *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, 2013, pp. 481–486.

[28] G. Pellegrino, C. Tschürtz, E. Bodden, and C. Rossow, "jäk: Using dynamic analysis to crawl and test modern web applications," in *International Symposium on Recent Advances in Intrusion Detection*. Springer, 2015, pp. 295–316.

[29] F. R. Muñoz and L. J. G. Villalba, "Web from preprocessor for crawling," *Multimedia Tools and Applications*, vol. 74, no. 19, pp. 8559–8570, 2015.

[30] G. Deepa, P. S. Thilagam, A. Praseed, and A. R. Pais, "Detlogic: A black-box approach for detecting logic vulnerabilities in web applications," *Journal of Network and Computer Applications*, vol. 109, pp. 89–109, 2018.

[31] G. Deepa, P. S. Thilagam, F. A. Khan, A. Praseed, A. R. Pais, and N. Palsetia, "Black-box detection of xquery injection and parameter tampering vulnerabilities in web applications," *International Journal of Information Security*, vol. 17, no. 1, pp. 105–120, 2018.

[32] K. J. Koswara and Y. D. W. Asnar, "Improving vulnerability scanner performance in detecting ajax application vulnerabilities," in *2019 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2019, pp. 1–5.

[33] C.-H. Liu, W.-K. Chen, and C.-C. Sun, "Guide: an interactive and incremental approach for crawling web applications," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1562–1584, 2020.

[34] M. S. Aliero, I. Ghani, K. N. Qureshi, and M. F. Rohani, "An algorithm for detecting sql injection vulnerability using black-box testing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 249–266, 2020.

[35] B. Eriksson, G. Pellegrino, and A. Sabelfeld, "Black widow: Blackbox data-driven web scanning," *proceedings of IEEE SSP*, 2021.

Improved Trust Model to Enhance Availability in Private Cloud

Vijay Kumar Damera[1]
Research Scholar, Department of CSE
JNTU Hyderabad, Telangana, India

A Nagesh[2]
Professor, Department of CSE
MGIT Hyderabad, Telangana, India

M Nagaratna[3]
Professor, Department of CSE
JNTUCEH Hyderabad, Telangana, India

***Abstract*—In the process of cloud service selection, it is difficult for users to choose trusted, available, and reliable cloud services. A trust model is a perfect solution for this service selection problem. In cloud computing, data availability and reliability have always been major concerns. According to research, around $285 million is lost per year due to cloud service failures, with a 99.91 percent availability rate. Replication has long been used to improve the data availability of large-scale cloud storage systems where errors are anticipated. As compared to a small-scale environment, where each data node can have different capabilities and can only accept a limited number of requests, replica placement in cloud storage systems becomes more complicated. As a result, deciding where to keep replicas in the system to meet the availability criteria is an issue. To address above issue this paper proposes a trust model which helps in selecting appropriate node for replica placement. This trust model generates comprehensive trust value of the data center node based on dynamic trust value combined with QoS parameters. Simulation experiments show that the model can reflect the dynamic change of data center node subject trust, enhance the predictability of node selection, and effectively decreases the failure rate of node.**

Keywords*—Trust; trust model; cloud; availability; reliability

I. Introduction

Cloud computing provides usable, convenient, and on-demand cloud services in the form of shared computing resource sharing pools. It includes three levels of services: Infrastructure as a service (IaaS), Platform as a service (PaaS), and Software as a service (SaaS), where IaaS provides consumers with comprehensive computer infrastructure services, such as hardware server rental, PaaS is to provide software development platform as a service to consumers, SaaS is to provide software to consumers through the network mode. It is a general practice to choose cloud service providers based on a comprehensive evaluation of QoS metrics [1]. The same is considered here in this work to evaluate the trust value of data center nodes. High availability and high-performance are essential features user expects from cloud service providers[3]. Replication plays an important role for any system to improve availability, throughput and response time for user [24].

In fact, replication is an essential corner stone in data storage not only for cloud computing but also for traditional storage systems [12], [13], [14], [15], because it can relatively impact the performance of cloud storage in terms of storage cost, network usage, response time, etc. [22], [23], [24], [25]. Therefore, maintaining static number of replicas in cloud storage for every data file would be inefficient for storage cost and data availability [16], [17], [18]. As a consequence, the determination of the optimum number of replicas and the suitable nodes for replicas has become a key issue in the cloud computing [19], [20], [21].

To address the above issue this paper proposes a node selection model based on the dynamic trust. The trust value is calculated by direct and recommended trust. In order to reflect trust value more comprehensively the concept of Dynamic Trust Value (DTV) is introduced in this model. Further the concept of information entropy is introduced to solve the problem of weighting of trusted parameters, so that the comprehensive trust value and QoS value are weighted to obtain the optimal node selection for replica placement.

II. Related Works

For the study of cloud service selection, many scholars at home and abroad have done a lot of work. For example works such as multi-objective genetic algorithm, particle swarm optimization PSO [1], artificial neural network algorithm [2].The application of these methods are generally aimed at a relatively static scenarios. The generation and disappearance of network services in a cloud computing environment are often dynamic, so it is necessary to solve the authenticity judgment of services in a cloud computing environment. Further it is important to solve the problem of degree of understanding of service quality when selecting a service algorithm [3], the trust of cloud services and other problems.

In response to the above problems, at the same time, in order to ensure the success of the service selection algorithm, the research based on the trust degree of the service subject is particularly important. Ma You et al. [4] proposed a new ESOW algorithm for QoS measurement. The ESOW algorithm is based on the user's subjective trust. The two parts of preference weight and objective weight are synthesized. The calculation of user's subjective weight is based on the adaptive SWDM algorithm, and the objective weight is calculated according to the OWDM algorithm. Sarbjeet [5] proposed a method based on the past experience and third-party service recommendation trust evaluation mechanism.

DASA et al. [6] proposed a dynamic trust calculation model that can effectively evaluate the behavior of malicious agent strategies. It mainly analyzes and evaluates all relevant elements to make correct decision. Zhouao et al. [7] proposed a dynamic virtual resource lease method from the perspective of service provider's benefit maximization, from the perspective of price allocation and request urgency. Cao Jie et al. [8] based on the interpersonal relationship in sociology relationship combined with user satisfaction evaluation, recommendation evaluation and third-party supervision feedback, a new credible measurement model was proposed. Zhang Lin et al. [9] combined relevant ranking factors, attribute factors, and intervals according to the behavior and dynamics of information services factor, penalty factor, and other four factors, a new dynamic trust monitoring model is proposed. Abawajy [10] proposes a distributed trust management framework based on reputation, which can pass the past experience, trust level and first level of honesty to determine the trust value of cloud computing entities.

III. Proposed Methodology

A. Basic Definitions

The service feedback results of the system described in this paper can be expressed by the two values namely positives and negatives. Therefore, the trust value of the data center node can be defined as the probability P of providing a good service, that is, using the evaluation information to reflect the probability P of providing a good service as accurately as possible. The process of cloud computing includes three main entities:

Service Provider: Represented as pro_j, representing the $j-th$ node, this node can provide users with the resources required by the user's cloud service request.

Service Consumer: Represented as $user_i$, i represents the $i-th$ service request user, the node can send service request information to the service intermediary, and can provide service history and service results to the service intermediary. As an evaluation of pro_j, user $user_i$ vs. pro_j is defined as a binary.

Service Broker: Represented as a broker, responsible for processing request response and management, and responsible for processing $user_i$ feedback information, providing $user_i$ with pro_j node evaluation information.

B. Direct Trust and Recommended Trust

The direct trust value is calculated using the past historical transactions and feedback information of two trading entities[16]. In the process of calculating the trust value, the Bayesian theory [11] is introduced to calculate the trust value of the node.

Suppose that the probability of good service provided by pro_j is P_j. After several transactions between the nodes and $user_i$. The evaluated binary "$Positives_{ij}$, $Negatives_{ij}$" is obtained, and the probability density function of P_j is obtained as:

$$\begin{gathered} f(p|\alpha,\beta)=\tfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1} \\ 0\leqslant p\leqslant 1,\alpha\geqslant 0,\beta\geqslant 0 \end{gathered} \tag{1}$$

In the formula:$\alpha=$ positives $_{ij}+1,\beta=$ negatives $_{ij}+1$

The Gama function is represented by:

$$\Gamma(z)=\int_0^{+\infty}\mathrm{e}^{-t}t^{z-1}\mathrm{d}t \tag{2}$$

Then there is a probability density function that can obtain the Bayesian estimate of P_j as:

$$P_{ij}=\int_{P_j=0}f\left(P_j|\alpha_{ij},\beta_{ij}\right)P_j\mathrm{d}P_j=\tfrac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}= \\ \tfrac{\text{positives }_{ij}+1}{\text{negatives }_{ij}+\text{ positives }_{ij}+2} \tag{3}$$

Then get the direct trust value : $T^{\mathrm{dt}}=P_{ij}$

The recommended trust value refers to the user $user_i$'s trust value for the service provider pro_j is obtained through the recommendation between other related entities, and the related entities are synthesized based on empirical evaluation. During the evaluation, the set of recommended entities is assumed to be R. The associated user-recommended user R in R recommends the two-tuple to $user_k$:

$$\left\{\begin{array}{l} \mathrm{Re}_{-}^{\text{postives}}\ \sum_{k\in R}\ \text{postives }_{kj} \\ \mathrm{Re}_{=}^{\text{negatives}}\ \sum_{k\in R}\ \text{negatives }_{kj} \end{array}\right. \tag{4}$$

Similarly, through the Bayes principle, the recommended trust value can be obtained:

$$T^{rt}=\frac{\mathrm{Re}_{\text{possives}}^{ik}\ +1}{\mathrm{Re}_{\text{negatives}}^{k}\ +\mathrm{Re}_{\text{postives}}^{ik}\ +2} \tag{5}$$

C. Time Decay Function

Not all user feedback can truly reflect the trust status between entities. Because over time, old user feedback may not accurately reflect the current trust value. For example, it may be evaluated that the service behavior of the entity has been modified or improved. So at this time set the weight according to the time of feedback to accurately reflect the user's feedback. This can be achieved by setting the time decay function mechanism. Assume that at time τ, after the unit time t, the user feedback trust formula for the attenuation of the value over time is shown in equation (6).

$$f_{ij}(\tau+t)=\left\{\begin{array}{ll} f_{ij}(\tau), & f_{ij}(\tau)\geqslant\theta_1,t\leqslant\theta_2 \\ f_{ij}(\tau)\mathrm{e}^{-\lambda(t-\theta_2)}, & f_{ij}(\tau)\geqslant\theta_1,t>\theta_2 \\ b, & \quad\text{other} \end{array}\right. \tag{6}$$

In the formula: λ is the decay constant, which is used to control the decay rate of the trust value. The value of λ can be set according to different service types. At the same time, the user's true intention of evaluating the entity and reducing trust are considered for the update frequency. A start attenuation threshold θ_2 and a stop attenuation threshold θ_1 are set here. When the elapsed time t is greater than θ_2, the trust value decreases gradually according to the attenuation constant, and when the trust value is less than or equal to the stop attenuation threshold at θ_1, the trust value is set to a fixed value b, and the trust value will not change with time.

D. Calculation of Comprehensive Trust

The calculation of comprehensive trust includes two aspects: the direct trust value T^{dt} and the recommended trust value T^{rt}, and the weight between the two can be set according to different service types. Set them to α, β and $\alpha + \beta = 1$, in this paper, they are set to 0.5, 0.5 respectively, then the calculation formula of the comprehensive trust $\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)$ of the data center node r_k is:

$$\begin{gathered}\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)=\alpha * \frac{\text{positives}_{ij}+1}{\text{negatives}_{ij}+\text{positives}_{ij}+2}+ \\ \beta * \frac{\mathrm{Re}_{\text{positives}}^{ik}+1}{\mathrm{Re}_{\text{negatives}}^{k}+\mathrm{Re}_{\text{positives}}^{k}+2}\end{gathered} \tag{7}$$

E. Dynamic Trust Value

Dynamic Trust Value (DTV), which represents the trend of trust value with respect to change with time. It can reflect the historical change of trust value, and has a pre-judgment indicator for the next node selection, thus improve the efficiency of node selection.

In order to be able to quantify the value of DTV, the least squares data fitting method is introduced. The least squares fitting method is a method to approximate or compare the functional relationship between the coordinates represented by discrete point groups on the plane with a straight line. Assume that the trust degree of pro_j changes with time as $\left\{\left(t_k, \text{trust}_k^j\right): k \in [1, n]\right\}$, where the node $\left(t_k, \mathrm{tr}\, ust_k^j\right)$ represents the trust value P_k of pro_j at time t_k. According to the least squares method, the fitted straight line equation is assumed to be:

$$y_j = \mathrm{DTV}_j * t_k + b \tag{8}$$

The slope of the straight line DTV_j is the Dynamic Trust Value of the defined data center node, b represents the intercept. In order to determine the value of DTV_j, b, according to the principle of the least square method, all data nodes $(t_k, y_k)\,(k = 1, 2 \ldots n)$. The square sum of the deviation values of all data nodes is minimized, that is:

$$\begin{gathered}M=\sum_{k=1}^{n}\left(\text{ trust }_k-y_k\right)^2= \\ \sum_{k=1}^{n}\left(\text{ trust }_k-\mathrm{DTV}_j * t_k-b\right)^2 subjectto \min(M)\end{gathered} \tag{9}$$

The condition for obtaining the minimum value is that the corresponding binary function takes the extreme value of 0, that is:

$$\frac{\partial M}{\partial \mathrm{DTV}_j}=\frac{\partial M}{\partial b}=0 \tag{10}$$

After finishing, the normal equations are obtained:

$$\left\{\begin{array}{l}\sum_{k=1}^{n} \text{trust}_k - \mathrm{DTV}_j \sum_{k=1}^{n} t_k - nb = 0 \\ \sum_{k=1}^{n} \text{trust}_k * t_i - \mathrm{DTV}_j \sum_{k=1}^{n} t_k^2 - b \sum_{k=1}^{n} t_k = 0\end{array}\right. \tag{11}$$

The linear parameter values DTV_j and b can be obtained by solving the normal equations, namely:

$$\begin{gathered}\mathrm{DTV}_j=\left(n \sum_{k=1}^{n} t_k * \operatorname{trus} t_k-\sum_{k=1}^{n} \text{trust}_k \sum_{k=1}^{n} t_k\right) / \\ \left(n \sum_{k=1}^{n} t_k^2-\left(\sum_{k=1}^{n} t_k\right)^2\right)\end{gathered} \tag{12}$$

$$\begin{gathered}b=\left(\sum_{k=1}^{n} t_k^2 \sum_{k=1}^{n} \operatorname{trus} t_k-\sum_{k=1}^{n} t_k \sum_{k=1}^{n} t_k * \operatorname{trus} t_k\right) \\ \left(n \sum_{k=1}^{n} t_k^2-\left(\sum_{k=1}^{n} t_k\right)^2\right)\end{gathered} \tag{13}$$

Then you can get the Dynamic Trust Value DTV_j of pro_j's trust change.

For the research needs of this paper, after the calculation of the Dynamic Trust Value DTV_j, it is normalized and converted into the following formula:

$$\mathrm{TCV}_i=\frac{\mathrm{TCV}_i-\min \left\{\mathrm{TCV}_k\right\}}{\max \left\{\mathrm{TCV}_k\right\}-\min \left\{\mathrm{TCV}_k\right\}} \tag{14}$$

The trust value can be converted to the range of [0, 1]. Since the Dynamic Trust Value of trust degree (DTV) reflects the change trend of trust degree, the level of DTV reflects the change of trust degree, so it can be based on the value of DTV. It is used to predict the value of future trust. The higher the DTV, the higher the trust value of the data center node. On the contrary, it indicates that the node provides false information, so that the trust is in a downward trend.

At the same time, when choosing the range of Dynamic Trust Values, the range of different trust change trend values will be normalized to the interval of [0, 1]. Choosing different intervals will not have much impact on the experimental results. However, in order to reflect the user's true trust feedback behavior, this paper selects the range of Dynamic Trust Values pertaining to e-commerce platforms such as eBay and Amazon [-0.875, 0.875].

F. Quality of Service (QoS)

QoS describes the ability of a product or service to meet consumer demand. To achieve better availability of cloud the following aspects, such as processing time, storage capacity, link capacity and type of operating system are considered as QoS parameters. They reflect service availability from different perspectives. This paper considers the QoS attributes of data center nodes from four aspects: Processing Capacity, Storage Capacity, Link Capacity and Operating System. The calculation of service performance can be calculated through user feedback or a third-party monitoring mechanism.

Assuming that there are a group of n services that meet the functional requirements, the vector of each group corresponding to the QoS attribute is set as: $Q_j = (q_p, q_s, q_l, q_o)$

$j = 1, 2, 3 \cdots, n, q_p, q_s, q_l, q_o$, respectively represent the processing capacity, storage capacity, link capacity and type of operating system of the j data center node. Because the value span between these attributes is relatively large, all QoS attributes need to be converted to normalization. q^{+}, q^{-}, respectively represents the value after the positive and negative QoS attributes are normalized, and their normal conversion methods are:

$$q^{-} = \begin{cases} \frac{q^{\max} - q}{q^{\max} - q^{\min}}, q^{\max} - q^{\min} \neq 0 \\ 1, q^{\max} - q^{\min} = 0 \end{cases} \tag{15}$$

$$\mathrm{q}^{+} = \begin{cases} \frac{q - q^{\min}}{q^{\max} - q^{\min}}, q^{\max} - q^{\min} \neq 0 \\ 1, q^{\max} - q^{\min} = 0 \end{cases} \tag{16}$$

Equation (15) shows that the attribute is negatively correlated with performance, that is, the larger the attribute value, the worse the performance, such as processing capacity and storage capacity; Equation (16) indicates that the attribute is positively correlated with performance, that is, the larger the attribute value, the better the performance and reliability. $q^{\max}$ and $q^{\min}$ respectively represent the maximum and minimum values in the attribute group. Then the QoS value can be obtained by linearly weighting each attribute:

$$Q = w_1 * q_p + w_2 * q_{\mathrm{s}} + w_3 * q_l + w_4 * q_o \tag{17}$$

G. Optimal Node Selection Strategy based on Information Entropy

The optimal node selection in the cloud computing environment is not only related to the QoS value, but also closely related to the trust value and the value of feedback by the user. It is a comprehensive reflection of the user's service quality. In previous studies, the analysis was only from the perspective of pure QoS. Some of them considered the trust value, but only considered the trust of the interaction process, and quantified the trust value as a single QoS value. This lacks in-depth study of service credibility. This paper believes that the optimal node selection should be considered in combination with both trust and QoS. Only considering any single aspect is unreasonable. At the same time, it is aimed at existing research where it only quantifies trust into a single value to evaluate the trust degree, which does not reflect the problem of trust degree changing with time. This paper integrates the Dynamic Trust Value (DTV) into the calculation process of node selection, and combines the concept of information entropy to determine the parameter weight between trust and QoS, so as to obtain the optimal node selection strategy.

Information entropy is a concept used to measure the amount of information. It is often used to give a rough measure of the uncertainty of information. Information entropy is a measure of the uncertainty of the result before the event. After the event, It is a measure of information obtained from the event. Therefore, the information entropy of an event is not only a measure of the amount of information of an event, but also can be included as relevant information in the event itself [12].

TABLE I. INFORMATION ENTROPY AND WEIGHT CALCULATION

Index	Service Quality	Trust Value
Value	0.6	0.8
Initial Value	0.4	0.2
Information Entropy	0.9708	0.7205
Weights	0.426	0.574

According to the calculation principle of information entropy, the information entropy determined by the quality of service QoS_j is:

$$\begin{gathered} H\left(\mathrm{QoS}_j\right) = -\mathrm{QoS}_j \log \mathrm{QoS}_j - (1 - \mathrm{QoS}_j) \log \\ \left(1 - \mathrm{QoS}_j\right) \end{gathered} \tag{18}$$

The information entropy determined by the trust value is:

$$\begin{aligned} & H\left(\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)\right) = -\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right) \log \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right) \\ & -\left(1 - \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)\right) \log\left(1 - \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)\right) \end{aligned} \tag{19}$$

Then the weight of the quality of service QoS_i can be determined:

$$\delta_1 = H\left(\mathrm{QoS}_j\right) / \left(H\left(\mathrm{QoS}_j\right) + H\left(\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)\right)\right) \tag{20}$$

And the weight of trust value $\delta_2 = 1 - \delta_1$

Among them, $\mathrm{CT}\left(\mathrm{pro}_{\mathrm{r}}^{\mathrm{j}}\right)$ means the degree of reliability of the subject j, and 1-$\mathrm{CT}\left(\mathrm{pro}_{\mathrm{r}}^{\mathrm{j}}\right)$ indicates the undeterminable component; QoS_j is the evaluation of the service quality of the subject j, and $1 - QoS_i$ is the uncertain component of its service quality.

For example, a service subject's (service quality, trust value) is (0.6, 0.8), then their information entropy is (0.970, 0.720), then the weight indicators of the two are: 0.4, 0.5, see Table I for details.

After the weights are calculated, the calculated trustworthiness value is trend-corrected according to the DTV_j value, so that the trustworthiness value can more accurately reflect the trustworthiness of the node. According to the dynamic trust value of trustworthiness, the correction formula is as follows (21).

$$\overline{\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)} = \begin{cases} \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right) + \left(1 - \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)\right) \times \\ \mathrm{DTV}_j \mathrm{DTV}_j > 0.5 \\ \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right) \mathrm{DTV}_j = 0.5 \\ \mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right) * e^{DTV_j - 0.5} \mathrm{DTV}_j < 0.5 \end{cases} \tag{21}$$

When the normalized Dynamic Trust value of trust degree DTV_j is equal to 5, we know that the change of trust value is in a relatively stable state at this time, so we do not change the trust value, and make a difference when the value of DTV_j is greater than or less than 5. The treatment is to punish some of the service providers' false information caused a decline in the trend of the change in trust, while the increase in the value of the trend of the change in the degree of trust changes the value of the trust appropriately. According to the revised trustworthiness value,QoS value and information entropy, the calculation weight is determined to obtain the comprehensive value of optimal node selection:

$$\Delta = \delta_1 * \mathrm{QoS}_j + \delta_2 * \overline{\mathrm{CT}\left(\mathrm{pro}_{r_k}^{j}\right)} \tag{22}$$

The following is a data center node selection strategy process based on the dynamic trust value:

1) The service broker receives the node selection request from the service user ($user_i$).
2) The service intermediary selects the data center node provider (pro_j) that meets the user's functional requirements by analyzing the service request.
3) According to formula (7), the service consumer obtains the comprehensive trust value through the direct trust value and the indirect trust value.
4) According to equation (8), the Dynamic Trust Value DTV of the node provider is obtained.
5) According to formula (17) to obtain the quantified QoS value of the overall service quality and formula (21) to modify the trust value, and through the information entropy calculation to obtain the corresponding weight.
6) According to equation (22), the overall evaluation value of the optimal node selection can be obtained.
7) Choose the node with the largest overall evaluation value as the best choice for replica placement.

IV. Results and Discussions

For this work simulation method is opted to verify the effectiveness of the node selection model proposed in this paper based on Dynamic Trust. The experiments were run on the cloud simulation software CloudSim3.0 [13]. CloudSim is a new universal and extensible simulation framework that supports seamless modeling and simulation, based on a specific environment and configuration, by extending its basic functions, can conduct experiments on cloud computing infrastructure and management services.

When CloudSim starts the simulation, you first need to create a data center (Datacenter), create resources such as CPU and memory in the data center, you can map user requests to the appropriate service provider (DatacenterBroker) through CIS (CloudInformationService), according to service selection Strategy for resource allocation. The operating environment is the Eclipse integrated development platform based on java development, and the CloudSim simulation program runs on the Intel Pentium dual-core G630, 2.7GHz, 2GB memory, Ubuntu18.04 64 bit Operation System on the desktop.

In this experiment, according to the different trend values of the trust degree of the node provider SP, the SP is divided into three categories:

1) Class A: The Dynamic Trust Value of SP is monotonously increasing. For example, due to the improvement of technical quality, the value of trust gradually increases.
2) Type B: The Dynamic Trust Value of SP's trust degree is relatively stable, that is to say, this type of SP provides stable cloud service functions and has good trust.
3) Type C: The Dynamic Trust Value of the SP is monotonously decreasing. For example, the SP provides a cloud service product with a false function description, which leads to the decrease of the trust value.

TABLE II. Simulation Parameters

Parameter		Default Value	Description
Operating Parameters	CloudletNum	100	Total number of tasks
	VmNum	500	Total number of SP
Algorithm Parameters	Num_luser	100	Total number of SC
Weights	α,β	0.5	Trust Weight
	w_i	0.25	QoS weight parameters

In order to conduct metric comparison, two test indicators are set: success rate and predicted success rate.

Within a certain time interval T, the CIS (CloudInformationService) in the CloudSim simulator provides the number N of service providers to all SCs according to the selection strategy, where the number of SPs that successfully interact with the SC is S (here, there is no fraud Behavior), the success rate is:

$$\theta = \frac{S}{N} \times 100\% \tag{23}$$

Where: θ is the average degree of cooperation between network nodes.

Within a certain time interval T, the CIS in the CloudSim emulator provides the number of service providers N to all SCs according to the selection strategy, encountered a situation where the node happened to be in the blacklist, or encountered during the interaction deception, the total number of times in both cases is M, then the model's predicted success rate is:

$$\varphi = \frac{N - M}{N} \times 100\% \tag{24}$$

In the formula: φ can reflect the model's ability to predict the next step.

In the verification test, it is assumed that the number of SP subjects is 500 and the number of SC subjects is 100 in a cloud environment. The trust value of each virtual machine starts to be randomly generated. The node selection policy input parameters, namely, CloudletNum, vmNum, num_user, α, β, w_i are shown in Table II.

In the table, CloudletNum represents the total number of tasks requested by users in the CloudSim simulation environment, and the tasks from different users are relatively independent; vmNum is the number of virtual machines; num_user represents the number of users; α and β are the direct trust weights and recommendation Trust weights, respectively. The data results of each group are averaged after 10 times.

Case Study 1: In a cloud environment, the nature of data center node is highly dynamic. It is assumed that a service node rj is malicious, but the resource may be used to cover its important service transactions. Good service quality establishes good trust. After a period of time, the service provider lowers the standard of service quality in order to reduce costs, but due to the establishment of early trust, the poor post-service has a higher trust value. So, it is necessary to introduce the trust time decay mechanism. When there is no transaction in the middle of a period of time, the trust value will decrease with time. Also set a decay constant, you can set the credibility decay speed according to different service types. The service occurs in a malicious time period, and a larger attenuation constant can be set. Fig. 1 shows the attenuation of the trust value under different attenuation constants, namely, type1(λ=y1=exp(-0.025x)), type2(λ=y2=exp(-0.05x)) and type3(λ=y3=exp(-0.1x)). Simulation results show that the larger the attenuation parameter, the faster the attenuation rate.

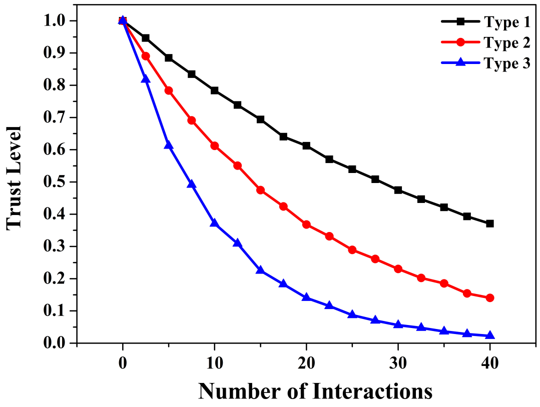


Fig. 1. Change of Trust under Different λ Values.

Case Study 2: Fig. 2 shows the change of the trust value of the three types of service providers during the interaction of the service subjects. From the figure, it can be seen that the class A SP is in the increasing trend as the number of interactions are increasing and its credibility has been maintained with a growing trend. The Trust value of the B-type SP increases with the number of interactions, but when it reaches a certain level, it will be in a relatively stable state. The Trust value of C-type SP begins to increase with the number of interactions. The trust value also increases. Although the C-type SP maintains a high trust value in the early stage, due to the provision of false services, the trust level in the later stage decreases, which ultimately leads to a lower level of trust value.

Case Study 3: Fig. 3 shows the comparison of the three different strategies namely the node selection method based on the Dynamic Trust Value, the node selection method based on trust degree, and the node selection method based on non-trust degree. It is an indicator of success, so its success rate has a greater advantage than the latter. But at the same time, the node selection method based on the dynamic trust has always maintained a better growth trend, because its DTV which effectively reflected the change trend of trust. This method has the ability to predict, to a certain extent can filter out false information, thereby improving the success rate of node selection in a better way.

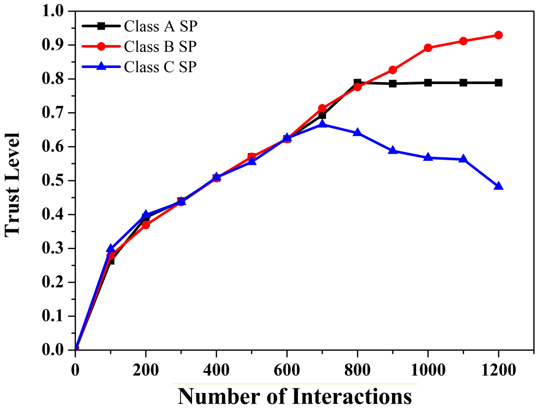


Fig. 2. Comparison of Different SP's under Dynamic Trust.

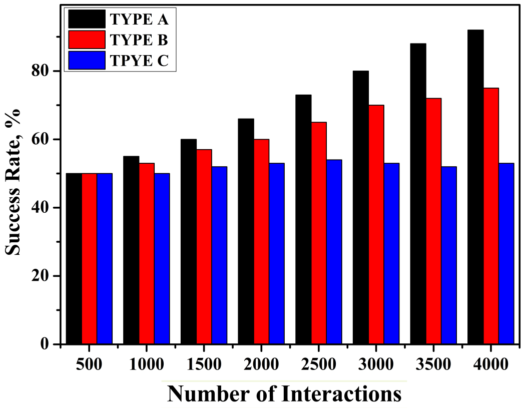


Fig. 3. Comparison of Node Selection Rate of Different SP's.

Case Study 4: Fig. 4 shows the comparison of the prediction success rate among the three methods namely the node selection method based on the Dynamic Trust Value, the node selection method based on trust degree, and the node selection method based on non-trust degree. The proposed method DTV has a certain predictive ability for node selection, so the prediction success rate has always shown a relatively stable growth state, while the node selection method based on trust does not predict the success rate after reaching a certain level. Again, this is because the trust-based node selection method does not contain the Dynamic trust evaluation and lacks continuous predictive ability. The trust-based node selection method has no consideration of trust factors, so it has the great blindness that led to the development of the prediction success rate in molar shape. The experimental results show that the model based on the DTV proposed in this paper effectively improves the prediction ability of node selection for replication.

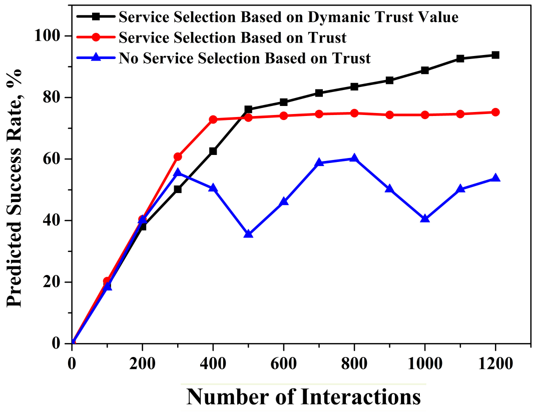


Fig. 4. Comparison of Predicted Success Rate of Proposed Model with Others.

V. Conclusion

In the cloud computing environment, this paper aims at the problem of replica placement in data center nodes for enhanced availability. The node selection for replica placement is a difficult process. Based on the proposed trust model it improves the ability to predict service quality of data center node and increases the accuracy of data center node selection for replication. The concept of Information Entropy is introduced to avoid the shortcomings that only perform simple weighted analysis of trusted parameters, and effectively improves the success rate of node selection. Experimental analysis shows that this method can better meet users' service quality and trust in node selection. Further the data center node which gets selected for replication using this model exhibits less failure rate, there by enhances availability in cloud.

References

[1] Wand et al., "Particle swarm optimization with skyline operator for fast cloud-based web service composition ", Mobile Network s and Applications, 2013 , 18(1) :116-121.

[2] Zhang et al., "Preference-aware QoS evaluation for cloud web service composition based on artificial neural networks", Web Information Systems and Mining , 2010 , 18(1) :116-121.

[3] H U Chunhua et al., "Services selection based on trust evolution and union for cloud computing", Journal on Communications, 2011 , 32 (7) :71-79.

[4] MA You et al., "Web service quality metric algorithm employing objective and subjective weight", Journal of Software. 2014 , 25 (11) : 2473 -2485.

[5] Sarabjeet S et al., "Trust evaluation in cloud based on friends and third party's recommendations", RAECS Conference. Panjab University Chandigarh: IEEE, 2014 :1-6.

[6] Das A et al., "Secured trust : a dynamic trust computation model for secured communication in multi agent systems", IEEE Transactions on Dependable and Secure Computing , 2012 , 9(2) :261-274.

[7] Zhou et al., "Dynamic virtual resource renting method for maximizing the profits of a cloud service provider in a dynamic pricing model" , International Conference on Parallel and Distributed Systems. Seoul :IEEE, 2013: 118-125.

[8] Cao , Jiang Huowen et al., "Trust-aware dynamic level scheduling algorithm in cloud environment", Journal on Communications , 2014 , 35(11) :40-49 .

[9] Zhang Lin and Wang Hai-yan, "Dynamic trust monitoring model supporting behavior in information services", Journal of Nanjing University of Posts and Telecommunications, 2013 , 33(1) :68-73.

[10] Abawajy J., "Determining service trustworthiness in intercloud computing environments", Proceedings of the 10th International Symposium in Pervasive Systems, Algorithms, and Networks. National Cheng Kung University: IEEE, 2009 :784 -788 .

[11] Josang A and Ismail R., "The beta reputation system", Bled Electronic Commerce Conference. Bled Slovenia: IEEE, 2002: 324-337.

[12] Huang Ying-jie et al., "Hybrid particle swarm optimization based on entropy for flexible job shop scheduling problems", Journal of Hunan Universit y: Natural Sciences, 2012 , 39(3) :48-52.

[13] Rahul M., "Study and comparison of CloudSim simulators in the cloud computing", The SU Transactions on Computer Science Engineering &. its Applications, 2013 , 1(4) :111-115.

[14] Tjang C et al., "Research on evaluation of SaaS SP service quality based on SLA", In Journal of Computer Engineering, 2013, Page:31-36.

[15] Dantas J et al, "Eucalyptus-based private clouds: availability modeling and comparison to the cost of a public cloud", 2017, Page:1130–1140

[16] Fan W., Perros, H., "A novel trust management framework for multi-cloud environments based on trust service providers", Knowl. Based Syst.70, 2014, 392–406.

[17] Rajendran, V.V., Swamynathan, S., "Hybrid model for dynamic evaluation of trust in cloud services", Wirel. Netw., 2015, 1–12.

[18] Jabbar, S., Naseer, K., Gohar, M., Rho, S., Chang, H., "Trust model at service layer of cloud computing for educational institutes", J. Supercomput., 2015,1–26.

[19] Chiregi, M., Navimipour, N.J., "A new method for trust and reputation evaluation in the cloud environments using the recommendations of opinion leaders' entities and removing the effect of troll entities", Comput. Human Behav. 60, 2016. 280–292.

[20] Selvaraj, A., Sundararajan, S., "Evidence-based trust evaluation system for cloud services using fuzzy logic", Int. J. Fuzzy Syst., 2017, 1–9.

[21] Lynn, T., van der Werff, L., Hunt, G., Healy, P., "Development of a cloud trust label: a Delphi approach", J. Comput. Inf. Syst. 56, 2016, 185–193.

[22] Tang, M., Dai, X., Liu, J., Chen, J., "Towards a trust evaluation middleware for cloud service selection", Future Gener. Comput. Syst. 74, 2017, 302–312.

[23] P. T. Endo, M. Rodrigues, G. E. Gonçalves, J. Kelner, D. H. Sadok, and C. Curescu, "High availability in clouds: systematic review and research challenges", J. Cloud Comput., vol. 5, no. 1, Oct. 2016, p. 16.

[24] Vijay Kumar, A Nagesh and M Nagaratna, "SLA-Based Trust Model to Enhance Availability in Private Cloud", International Journal of Advanced Science and Technology, 29(05), 2020, pp.13941 - 13954.

[25] J. Riley, J. Noss, W. Dillingham, J. Cuff, and I. M. Llorente, "A High-Availability Cloud for Research Computing," Computer, vol. 50, no. 6, 2017, pp. 92–95.

Supervised Learning-based Cancer Detection

Juel Sikder[1], Utpol Kanti Das[2], Rana Jyoti Chakma[3]
Dept. of Computer Science and Engineering
Rangamati Science and Technology University
Rangamati, Bangladesh

***Abstract*—The segmentation, detection and extraction of the infected tumor from Magnetic Resonance Imaging (MRI) images are the key concerns for radiologists or clinical experts. But it is tedious and time consuming and its accuracy depends on their experience only. This paper suggest a new methodology segmentation, recognition, classification and detection of different types of cancer cells from both MRI and RGB (Red, Green, Blue) images are performed using supervised learning, Convolutional Neural Network (CNN) and morphological operations. In this methodology, CNN is used to classify cancer types and semantic segmentation to segment cancer cells. The system trained using the pixel labeled the ground truth where every image labeled as cancerous and non-cancerous. The system trained with 70% images and validated and tested with the rest 30%. Finally, the segmented cancer region is extracted and its percentage area is calculated. The research examined on the MATLAB platform on MRI and RGB images of the infected cell of BreCaHAD dataset for breast cancer, SN-AM Dataset for leukemia, Lung and Colon Cancer Histopathological Images dataset for lung cancer and Brain MRI Images for Brain Tumor Detection dataset for brain cancer.**

***Keywords*—Semantic segmentation; CNN; brain; breast; leukemia; lung**

I. Introduction

Muscles, bones, the liver and the lungs are forming rapidly from trillions of cells produced by the human body. Lung cancer starts its journey from the lung. Secondary type cancer spreads to the lung from another part of the body. Cancer which cannot compete against infection spreads in the bone marrow and also a cancer that happens in white blood cells is known as Leukemia. The brain tumor is a collection or a combination of unnatural tissue in the brain. MRI, computerized tomography (CT) scan and the Biopsy technique used to detect the brain tumor. MRI images are safer than CT scan images, because MRI can produce higher contrast compared with CT scan and MRI images also do not harm the body. As the MRI images produce a pure resolution of brain tissues, they are used in detection and classification techniques. The lymph vessels and blood vessels help breast cancer to propagate outside the breast.

Nowadays, it is common to see that deep learning-based segmentation is a more useful segmentation technique in the present world which introduces deep learning in every sphere of the digital world. Semantic segmentation is one of the best deep learning-based segmentation techniques that introduces the machine's knowledge to human beings. A good example is supervised learning that includes labeled the ground truth dataset of learning object categories. This outstanding supervised methodology used SegNet architecture for the segmentation purpose of cancer cells that are of different types. Classification of objects is more useful and necessary for the recognition of testing objects like cancer cells in this study. This research used Convolutional Neural Network(CNN) for classification purposes. It takes the help of the resnet50 architecture that uses fc1000 layers for this deep classification of the testing image [1]. The authors implement an idea that turns the machine into a human brain. To test this study, different types of cancer datasets for segment and detect cancer cells of heterogeneous have been used. The system split all dataset images of the different format into different classes, categorized them into Breast, Leukemia, Lung and Brain where every category of the dataset is labeled using the pixel label named Cancerous and Non_Cancerous class. During performance as a human being, test image is classified using CNN classifier and based on the classified predict label, classified category along with its labeling are passed into the SegNet architecture to build a SegNet semantic segmentation procedure using VGG16 layers. Creating semantic segmentation procedure, a test image is segmented and segmented cancer cell is detected using morphological operations and detected cancer area is calculated. Many researchers have only worked with specific types of cancer cell detection with different methods, but in this research we focused on many types of cancer cell detection using the proposed system. In brief, we develop a system to segment, detect and classify the cancer cells more especially Brain, Breast, Leukemia, and Lung using semantic segmentation and classification using the CNN classifier which is more convenient than traditional existing method.

The structure of this research has been designed as follows. The next Section II contains a review of some recent related research works. The proposed methodology is described in brief in Section III. Results and discussion sections are described in Section IV and conclusions and future research plans are discussed in Section V.

II. Literature Review

There have been many works done to detect and analyze cancer cells. Some of the existing works are described below.

Fahad Lateef and Yassine Ruichek [2] have completed a survey. According to the shared conceptions underlying their structure, they categorized different methods among ten different classes in their survey. They also provided a brief description of the publicly available datasets. Besides, to measure their accuracy, they used an evaluation matrix. Furthermore, they paid attention to different methodologies and examined to achieve their desired performances. Finally, they ended the research by describing discrete amounts of open questions and their probable solutions.

Alqazzaz, Salma, et al. [3] have proposed a methodology where they applied a CNN e.g. SegNet to 3D datasets for automation of the brain tumor and sub tumor portions segmentation, as well as enhancing tumor, edema and necrosis. They applied a process to further advance tumor segmentation. In this process, to produce four maximum feature maps, they integrated the four distinctly trained SegNet architecture. To encrypt fascinating info into a feature presentation, they combined both the intensity of the pixels of the primary MRI modalities and the large feature maps. To classify the MRI voxels, the combined features are taken as input and applied to a tree that can make decisions. The authors used the BraTS 2017 challenge dataset to evaluate their methodology.

Ahmed Ghoneim, Ghulam Muhammad, M. Shamim Hossain [4] have proposed a method to classify and detect Cervical cancer. They used CNN and extreme learning machines to classify. Images of the cells are passed into a CNNs frame so that deep-learned features can be extracted. Then, extracted features have been used to classify the test images. They claimed that the classifier gives 91.2% accuracy.

Konstantinos Kamnitsas, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, Ben Glocker [5] have applied effective multi-scale 3D CNN with complete associated CRF for segmentation of the brain lesion. They projected a double route i.e. three-dimensional CNN and 11-layers deep to segment brain lesions. The authors used BRATS 2015 and ISLES 2015 dataset to examine their architecture.

Almajalid, Rania, Juan Shan, Yaodong Du, and Ming Zhang [6] have modified the u-net model for breast ultrasound imaging (BUS) segmentation. They applied a post processing after the segmentation stage. The authors claimed that the method achieved 82.52% accuracy in the term of average DICE.

Al-jaboriy, Saif S., Nilam Nur Amir Sjarif, Suriayati Chuprat, and Wafaa Mustafa Abduallah [7] have introduced a machine learning-based autonomic leukocyte cell segmentation procedure. Using artificial neural networks (ANNs) and 4-moment statistical features, the features were extracted from blast cells.

Kamal, Uday, Abdul Muntakim Rafi, Rakibul Hoque, Jonathan Wu, and Md Kamrul Hasan [8] have proposed a deep learning model for the segmentation of lung tumors from CT scans called Recurrent 3D-DenseUNet. To differentiate between tumorous and non-tumorous image-slices, they applied morphological operations and selective thresholding. They claimed that the system gained an average dice score of 0.7228.

III. Methodology

The proposed system divided into the following three steps:

- Classification stage.
- Semantic segmentation and detection stage.
- Calculate cancer area.

”Fig. 1” illustrates the proposed methodology.

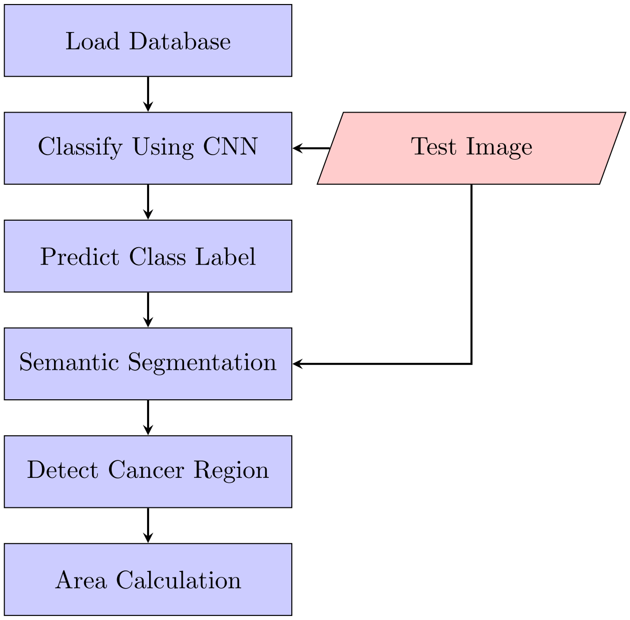


Fig. 1. Block Diagram of the Proposed System.

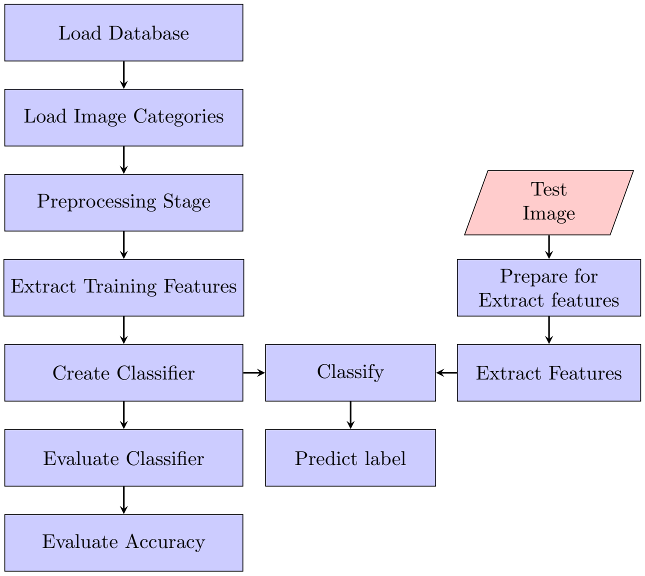


Fig. 2. Block Diagram of the CNN Classifier.

A. Classification Stage

To classify a dataset into user-defined classes, a supervised learning classification method is used [9]. It is the procedure that predicts the class of given data. The task of resembling a mapping procedure from input to output is known as classification predictive modeling. ”Fig. 2” illustrates the block diagram of the CNN classifier.

For training a classifier, a pre-trained CNN is used as a feature extractor in the proposed system. CNNs are trained to gather a large set of varied images. From these vast collections of data, CNNs can extract important features viz. SURF, LBP and HOG [11].

1) Load Data Base: To test the classification approach of cancer cells, the system used Brain, Breast, Leukemia and Lung category here. The data sets loaded to the system for creating the CNN classifier.

2) Load image category: The system constructed an image data store (imageDatastore) function based on the following categories ‘Brain’, ‘Breast’, ‘Leukemia’ and ‘Lung’. The system used the image data store (image Datastore) function to help manage the data. As an image data store (imageDatastore) function operates on image file locations, all the images don’t load into memory.

3) Preprocessing Stage:

a) Apply CNN using ResNet-50: ‘ResNet-50’ net can be loaded from neural network toolbox^{TM}. The system needs to initialize ‘ResNet-50’ applying resnet50. The first layer shows the input dimensions. CNN required 224-by-224-by-3 image size. A package of convolutional layers interspersed with rectified linear units (ReLU) and max-pooling layers exist in this architecture.

b) Prepare Training and Testing Image Sets: The system fragments the dataset into training and validation data. The system picked 60% of cancer cells for training and the remaining, 40%, for the testing. It randomized the dividing process to remove any biasing. The CNN architecture processed training and testing datasets. A supervised procedure and the training data trained the model [10]. The parameters of the model are accustomed depending on the judgement results. Both the parameter estimation and variable selection can be taken by the model fitting. Using some performance metrics the accuracy of the model is evaluated using the test set. Besides the training and testing set, a validation dataset is used to validate the trained model [11]. It is a very general way to assign more than 50% images for the training set, 25% for the test set, and the remainder for the validation set.

c) Pre-process Images For CNN: Resnet-50 uses 224-by-224-3 dimension images to perform. So, the system converts grayscale images into RGB and resizes them according to the requirement of resnet50. During network training, additional data augmentation processes use that resized image. The system resized both the training and testing images.

4) Extract Training Features: To extract features, the CNN model has few layers. The features are extracted from edges and blobs at the beginning of the networks. The second convolutional layer weights are required to visualize first convolutional layer weights [12]. Then the system resized and scaled the weights for visualization. In the first layer, there are about 96 discrete groups of weights. The first convolutional layer weight is given in “Fig. 3”. The activation method is used to extract the deep learning features. ’fc1000’ layers are used to extract features using activation functions. Either a graphics processing unit (GPU) or a central processing unit (CPU) is used by the activation function automatically.

5) Create Classifier using extracted features: A multiclass support vector machine(SVM) classifier uses extracted features to train a classifier. The fitcecoc function is used primarily by a fast-stochastic Gradient Descent solver. To create the classifier firstly, the system collects training labels from the training dataset. After that multiclass SVM classifiers are trained by the system using the fastest-linear solver [13][14].

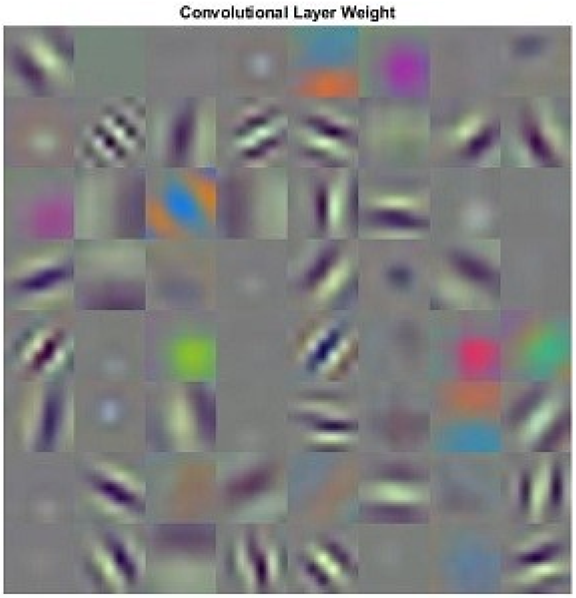


Fig. 3. First Convolutional Layer Weight.

6) Evaluate Classifier: In this step, the system tests extracted features using the trained CNN classifier. To evaluate the classifier, extracted features fed to the classifier, then the system identified class labels and tabulated the outcomes via a confusion matrix. Finally, the confusion matrix is converted into percentage form.

7) Classification of the test cancer cell: To classify the test cancer cell, the system needs to pre-process the test image or test the cancer cell. For this purpose, the system has to apply pre-processing stages before applying classification. When image features are extracted using an activation function, then extracted image features passed to the classifier and the previously created classifier classifies the test image [1].

8) Predict class label: When the classifier classifies the test image, the classifier predicts a label that represents which class the test image belongs to. Based on this predicted class label, the next segmentation stage starts to process. This predicted class label helps the next process to minimize its dataset and its processing time or execution time also.

B. Semantic Segmentation and Detection Stage

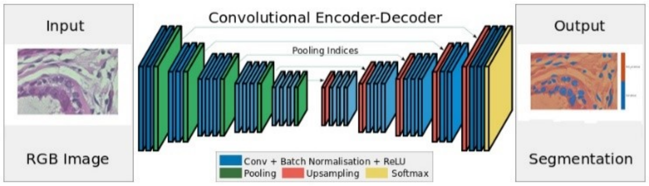


Fig. 4. Architecture of Semantic Segmentation.

“Fig. 4” illustrates the architecture of semantic segmentation. The encoders (i.e., a sequence of non-linear processing layers) are contained by the SegNet architecture. A pixel-wise classifier also follows the decoders (i.e., an equivalent group of decoders). Characteristically, one or more convolutional layers are contained by each encoder. Applying max-pooling directories in the encoding order, the decoder up-samples the light encoding because of the max-pooling [2].

To execute up-sampling of the low-resolution feature model in the decoders, max-pooling indices are used as one major component of the SegNet. This procedure bears some significant merits of holding the highest frequency niceties in segmented images. In the decoders, this procedure also helps to diminish trainable constraints. Using the stochastic gradient descent, the complete system can be accomplished end-to-end. The block diagram of semantic segmentation is given in “Fig. 5”.

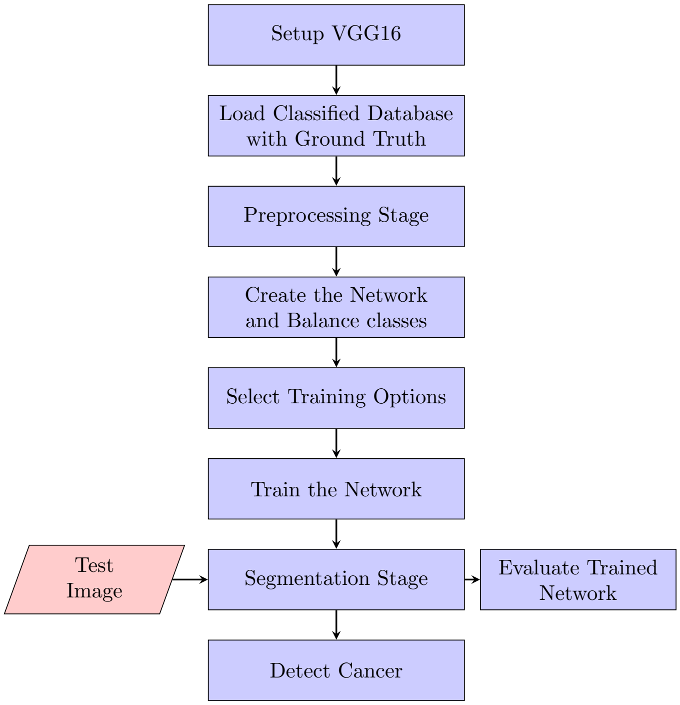


Fig. 5. Proposed block Diagram of Semantic Segmentation.

1) Setup VGG16: VGG-16 net is a 16 layers deep CNN net. The system loads a pre-trained network trained.To classify pictures into 1000 object classes (such as Brain, Breast, Leukemia, Lung and many other cancer types),the pretrained network is used. The system loaded a pretrained VGG-16 CNN net and examined the layers and categories. The pretrained VGG-16 network loaded using the vgg16 function. A sequence of the network object is used to produce output. The network has 41 layers. There are 16 layers with learnable weights: 3 fully connected layers and 13 convolutional layers exist in the net and all those 16 layers are learnable [3].

2) Load Classified Database with Ground Truth: After setting up the neural network, the system loaded the database with its ground truth. For this reason, the system classified the class of the test image. Based on the classified class the system loaded the only classified class database with its ground truth. Some labeled image with its original image is given in “Fig. 6”.

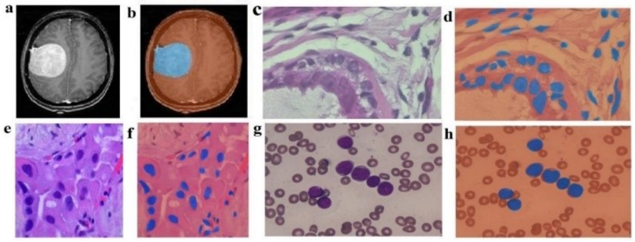


Fig. 6. Original images (a. Brain, c. Breast, e. Lung, g. Leukemia) and Labeling Images (b. Brain, d. Breast, f. Lung, h. Leukemia).

3) Preprocessing Stage:

a) Identify class and PixelLabelId: The system identified all class names and their corresponding pixel label id based on the classified loaded database. A large set of images and its counter set of labeled data are required to train a semantic segmentation network. To state the mapping between class names and label IDs, the pixel label data used. In the used dataset, the labels are “Cancerous” and “Non-cancerous” having 1, 2 label IDs respectively.

b) Analyze Dataset Statistics: In this study count pixels of Each Label used to observe the circulation of class labels in the database. The system can count the number of pixels in each label of a class and also can count the total pixels of each class in the used database. Using those pixel values, the percentage of cancer area and the non-cancer area are calculated. Those can be calculated using the following formula:

$$Frequency\ of\ each\ labels = \frac{Sum\ of\ label\ pixels}{Sum\ of\ image\ pixels} \quad (1)$$

c) Resize both Database Images and Labeled Images: In this research SegNet used the only process RGB images that are 300-by-300. To avoid re-saving all database images to this setup, the system used a converted and resized RGB image. The resized procedure has been applied not only for extra data augmentation, but also for training the network. Both the dataset image and labeled images are resized by the system.

d) Prepare both Training and Test Set: The system has described in the previous CNN section. The system needs to split the total dataset images and their pixel labels for training and testing sets. The system split the dataset 60% for the training set and 40% for the testing set.

4) Create the Network and Balance Classes: The system used the ‘segnetLayers’ procedure to make a neural network depending on VGG-16. The system specified the network image size. Then the number of categories is specified. Finally, the system created the SegNet layer network using image size, the number of categories and VGG-16. The classes are not balanced in an earlier dataset. To improve training, the system used class weight to equilibrium the classes. The pixel label computed earlier with a count each of label to compute the average frequency class weights, the pixel label measured past with the count of each label. The system identified class weights using a pixel classification layer. Then the system needs to update the pixel classification Layer of SegNet and delete the current pixel classification Layer and add the newly defined layer.

5) Select Training Options: The gradient method used Stochastic gradient descent momentum (SGDM). The system also needs to adjust the mini-batch size according to the memory size installed in the GPU to be used. The system applied training options to regulate the hyper-constraints used for SGDM. A factor of 0.3 every 10 epochs has been used to diminish the learning rate. Setting the 'ValidationData' parameter against the validation data every epoch, the system has tested the network. When the validation accuracy joins, stopping the criteria of the training has been set to 4 by the 'ValidationPatience'. To diminish memory procedure while training, a mini-batch size of 8 has been used. Depending on the quantity of GPU on the system, this mini-batch size can increase or decrease. In addition, At the end of every training epoch, the saving of network checkpoints is enabled by this name-value pair. The system can continue training from the kept checkpoint, if training is hindered because of a system deduction or power outage.

6) Train the Network: Data augmentation has been used during training to progress the accuracy of the network. Data augmentation has been completed using randomized X/Y translation and left/right reflection. The system used the image data augmenter to regulate these data augmentation constraints. Image data augmenter assists different other categories of data augmentation. During start training, the system combined data augmentation selections and the training data using pixel label image datastore. Batches of training data are read by the pixel label image datastore. To pass the resultant data of the augmentation process to the training process, data augmentation is applied. Pixel label image source is used to define the final data used for learning. Augmentation is also done here. After completing training, the system can segment the test image of the cancer cell using the created semantic segmentation network. For this reason, the system has to load the test image to the semantic segmentation network.

7) Semantic Segmentation: The field of computer vision has been empowered by deep learning. Semantic segmentation is one part of computer vision. Labeling each pixel is the purpose of semantic segmentation. Because the system is predicting for each pixel in the image, the dense prediction is referred to as the system. As the system created the SegNet network and trained it successfully, the system can now segment the loaded test image to the network. Using the trained net, the system can segment the test image and overlap the pixel label in it. It can be seen that the regions have been relatively neatly divided. Some test images are given in "Fig. 7".

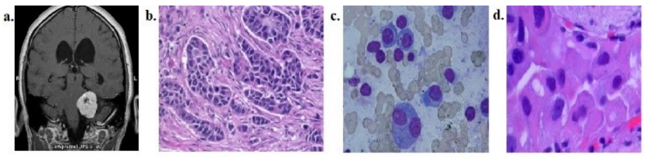


Fig. 7. Input Images (a. Brain, b. Breast, c. Leukeima, d. Lung).

The system can compare the segmentation result with the ground truth. As a result of overwriting, the green and magenta areas are different from the true values. For each class segmented by SegNet, the system can evaluate how much the ground truth region is included. This is an index called IoU and can be measured using the jaccard function. In addition to jaccard, the dice coefficient and BF coefficient are often used as the coefficient for expressing the similarity to the true value. The system uses the dice function and the bfscore function on MATLAB respectively. Segmented result of the test images given in "Fig. 8".

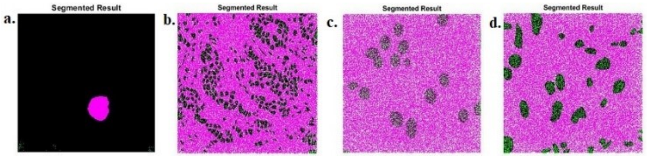


Fig. 8. Segmented Results (a. Brain, b. Breast, c. Leukemia, d. Lung).

8) Evaluate Trained Network: Semanticseg function has been run on the complete test set, to calculate the accuracy. Semanticseg returns the outcomes as a pixel label datastore object. The 'WriteLocation' parameter is used to write the location of the actual pixel label data. The system used the evaluate semantic segmentation function to calculate semantic segmentation metrics. For individual classes and for each input image, the semantic segmentation process returns various metrics for the complete dataset. The dataset metrics used to identify the network performance.

9) Detect Cancer: After segmentation, the system can detect any object from the segmentation output. Based on the cancer cell labeled, the system can separate the cancer cell from another labeling object. To get a meaningful segmentation output, the system applies different morphological operations and finally converts it into a binary form so that it can separate the cancer cell and others. Cancer cells represented in white e.g. 1 and non-cancerous are represented as black e.g. 0 [15]. The binary representation of segmented cancer cells given in "Fig. 9".

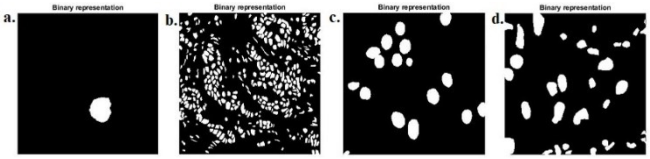


Fig. 9. Binary Representation (a. Brain, b. Breast, c. Leukemia, d. Lung).

After detecting cancer cells, the boundary of the cancer cells is identified by applying boundary detection algorithms and the boundary is represented by the red color. Detected cancer output is given in "Fig. 10".

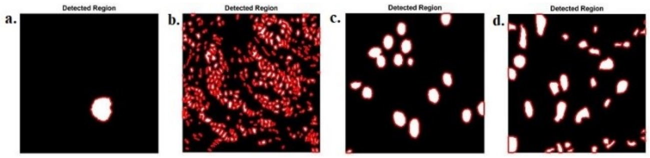


Fig. 10. Detected Region (a. Brain, b. Breast, c. Leukemia, d. Lung).

C. Area Calculation

Traced and detected cancerous outcomes are in the binary form. So, to calculate the percentage area of the cancerous region, the intensity (i.e., 1) of each pixel is summed and the total number of pixels of the test image is counted. The following equation is used to count both the area of the detected cancerous area and the total area of the test image [3].

$$Area = \sum_{i=1,j=1}^{n,m} P(i,j) \tag{2}$$

In the system, the cancerous area is identified with the pixel intensity having ones and the non-cancerous area identified with the pixel intensity zeros [1]. The percentage of the cancerous area is calculated with the ratio of the cancerous area and the total test image are as follows:

$$The\, percentage\; area = \frac{The\ area\ of\ Cancerous\ Cell}{The\ area\ of\ tested\ image} \times 100\% \tag{3}$$

IV. Result and Discussion

This paper developed a system to experiment on different databases of different types of cancerous cells. For brain cancer or brain tumor dataset purpose, the authors used Brain MRI Images for Brain Tumor Detection dataset. It consists of 154 brain images. For breast cancer histopathological annotation and diagnosis (BreCaHAD) dataset contains required files for Breast Cancer Histopathological Annotation and Diagnosis. It contains three folders named groundTruth_display (ground truth used on actual images), groundTruth (json files), images (original images). From those folders, the authors used only original images. This dataset contains 162 test images. For leukemia, the authors used SN-AM Dataset [16][17][18][19][20], which has three subsets. They used subset-2, and subset-3 where each of the subset contained 30 images. Lung and Colon Cancer Histopathological Images dataset [21], which consists of 25000 histopathological images and five categories, is used for Lung Cancer. Each image's size is 768 x 768 pixels, and each of them is in a jpeg file format. The authors created new polygonal ground truth for all datasets.

For evaluation of the proposed methodology Intersection over Union (IoU) is used. To evaluate IoU, detected cancerous areas and another area counted from ground truth are required. More formally, the hand-labeled ground truth that can form a polygon in the cancerous region and another polygon created with detection from the system methodology is required to estimate the proposed system accuracy using IoU. As the system used by the hand labeled ground-truth of different datasets and applied Intersection over Union (IoU) evaluation metric [22].

$$IoU = \frac{Region\ of\ the\ Overlap}{Region\ of\ the\ Union} \tag{4}$$

Table I. illustrates the IoU measuring raw values calculated from the total database and individual IoU. Calculating the average IoU from the above table, the system produces 93% accuracy.

TABLE I. Evaluation of the System

Cancer Type	Area of Overlap	Area of Union	IoU
Brain	93,195	94,529	0.9859
Breast	2,724,105	2,988,369	0.9116
Leukemia	597,690	642,560	0.9302
Lung	72,695,210	80,560,420	0.9024

Again, in the classification, the confusion matrix is applied to evaluate the exactness of CNN. It is used to identify the accuracy of the classifier. By observing how the classifier predicts a correct one as a truly positive and also observing a relationship of each category it can predict a class. The main job of a confusion matrix is to predict the results of a classifier and identify how correctly the classifier predicts a known class. Using true negative (x), true positive (w), false negative (z) and false positive (y) of the confusion matrix, the system can calculate the following performance.

$$Accuracy(Ac) = \frac{w+x}{w+x+y+z} \times 100 \tag{5}$$

$$Misclassification\ Rate = 1 - Accuracy \tag{6}$$

$$Specificity = \frac{x}{x+y} \times 100 \tag{7}$$

$$Precision(Ps) = \frac{w}{w+y} \times 100 \tag{8}$$

$$Recall(Re) = \frac{w}{w+z} \times 100 \tag{9}$$

$$F - measure(F-m) = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100 \tag{10}$$

TABLE II. Performance Calculation of CNN Classifier

Cancer Type	w	x	y	z	Ac(%)	Ps(%)	Re(%)	F-m(%)
Brain	18	54	0	0	100	100	100	100
Breast	18	54	0	0	100	100	100	100
Leukemia	18	54	0	0	100	100	100	100
Lung	18	54	0	0	100	100	100	100

Table II shows that the CNN classifier gave a 100% accuracy rate.

V. Conclusion

This research overcame the limitation of using individual methods for the cancer detection system for both MRI and Histopathology images. The results obtained in this research notify that the author's strategy is practical and can also pointedly support precise, can automatically detect cancerous cells and the proposed combination gives better accuracy after applying the machine learning algorithm.The authors reduced the training time and increased the segmentation accuracy using a CNN classifier and SegNet model with a morphological operation which is better than the common SegNet architecture. The experimental result achieves detection of cancer cells from various cancer databases that represent proper accuracy rates of an average of 93%.

Future research will be focused on using the more accurate, advanced deep learning algorithms to reduce the time complexity for large datasets and improve the accuracy rate. We will use more types of cancer cells and their categories. This research will open many opportunities for further research in image segmentation and cancer cell detection.

REFERENCES

[1] Sikder, Juel, Utpol Kanti Das, and AM Shahed Anwar. ”Cancer Cell Segmentation Based on Unsupervised Clustering and Deep Learning.” In International Conference on Intelligent Computing & Optimization, pp. 607-620. Springer, Cham, 2020.

[2] Lateef, Fahad, and Yassine Ruichek. ”Survey on semantic segmentation using deep learning techniques.” Neurocomputing 338 (2019): 321-348.

[3] Alqazzaz, Salma, Xianfang Sun, Xin Yang, and Len Nokes. ”Automated brain tumor segmentation on multi-modal MR image using SegNet.” Computational Visual Media 5, no. 2 (2019): 209-219.

[4] Ghoneim, Ahmed, Ghulam Muhammad, and M. Shamim Hossain. ”Cervical cancer classification using convolutional neural networks and extreme learning machines.” Future Generation Computer Systems 102 (2020): 643-649.

[5] Kamnitsas, Konstantinos, Christian Ledig, Virginia FJ Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. ”Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.” Medical image analysis 36 (2017): 61-78.

[6] Almajalid, Rania, Juan Shan, Yaodong Du, and Ming Zhang. ”Development of a deep-learning-based method for breast ultrasound image segmentation.” In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1103-1108. IEEE, 2018.

[7] Al-jaboriy, Saif S., Nilam Nur Amir Sjarif, Suriayati Chuprat, and Wafaa Mustafa Abduallah. ”Acute lymphoblastic leukemia segmentation using local pixel information.” Pattern Recognition Letters 125 (2019): 85-90.

[8] Kamal, Uday, Abdul Muntakim Rafi, Rakibul Hoque, Jonathan Wu, and Md Kamrul Hasan. ”Lung cancer tumor region segmentation using recurrent 3d-denseunet.” In International Workshop on Thoracic Image Analysis, pp. 36-47. Springer, Cham, 2020.

[9] Mahmud, Tanjim, Juel Sikder, Rana Jyoti Chakma, and Jannat Fardoush. ”Fabric Defect Detection System.” In International Conference on Intelligent Computing & Optimization, pp. 788-800. Springer, Cham, 2020.

[10] Tanjim Mahmud, Juel Sikder, Umme Salma, Sultana Rokeya Naher, Jannat Fardoush, Na-hed Sharmen, Sajib Tripura. “An Optimal Learning Model for Training Expert System to Detect Uterine Cancer” The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) March 23-26, 2021, Warsaw, Poland, Procedia Computer Science 184 (2021) 356–363. Published by Elsevier B.V, https://doi.org/10.1016/j.procs.2021.03.045.

[11] de Brebisson, Alexander, and Giovanni Montana. ”Deep neural networks for anatomical brain segmentation.” In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 20-28. 2015.

[12] Damodharan, Selvaraj, and Dhanasekaran Raghavan. ”Combining tissue segmentation and neural network for brain tumor detection.” International Arab Journal of Information Technology (IAJIT) 12, no. 1 (2015).

[13] Wang, Meng, Xiaobo Zhou, Fuhai Li, Jeremy Huckins, Randy W. King, and Stephen TC Wong. ”Novel cell segmentation and online learning algorithms for cell phase identification in automated time-lapse microscopy.” In 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 65-68. IEEE, 2007.

[14] Tanjim Mahmud, Sajib Tripura, Umme Salma, Jannat Fardoush, Sultana Rokeya Naher, Juel Sikder and Md Faisal Bin Abdul Aziz. “Face Detection and Recognition System”, 2nd In-ternational Conference on Technology Innovation and Data Sciences (ICTIDS)-2021, Kuala Lumpur, Malaysia, Published in Lecture Notes in Networks and Systems” Series (LNNS) - Springer Nature. DOI: 10.1007/978-981-16-3153-5.

[15] Alexandre de Brebisson, Giovanni Montana, “Deep neural networks for anatomical brain segmentation”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20-28, 2015.

[16] Gupta, Anubha, Rahul Duggal, Shiv Gehlot, Ritu Gupta, Anvit Mangal, Lalit Kumar, Nisarg Thakkar, and Devprakash Satpathy. ”GCTI-SN: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images.” Medical Image Analysis 65 (2020): 101788.

[17] Gupta, Ritu, Pramit Mallick, Rahul Duggal, Anubha Gupta, and Ojaswa Sharma. ”Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple myeloma.” Clinical Lymphoma, Myeloma and Leukemia 17, no. 1 (2017): e99.

[18] Duggal, Rahul, Anubha Gupta, Ritu Gupta, Manya Wadhwa, and Chirag Ahuja. ”Overlapping cell nuclei segmentation in microscopic images using deep belief networks.” In Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1-8. 2016.

[19] Duggal, Rahul, Anubha Gupta, and Ritu Gupta. ”Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks.” CME Series on Hemato-Oncopathology, All India Institute of Medical Sciences (AIIMS), New Delhi, India (2016).

[20] Duggal, Rahul, Anubha Gupta, Ritu Gupta, and Pramit Mallick. ”SD-layer: stain deconvolutional layer for CNNs in medical microscopic imaging.” In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 435-443. Springer, Cham, 2017.

[21] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019.

[22] Zheng, Zhaohui, et al. ”Distance-IoU loss: Faster and better learning for bounding box regression.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

Analysis of the Use of Videoconferencing in the Learning Process During the Pandemic at a University in Lima

Angie Del Rio-Chillcce[1], Luis Jara-Monge[2], Laberiano Andrade-Arenas[3]
Faculty of Science and Engineering
University of Sciences and Humanities
Lima, Perú

***Abstract*—Due to the health emergency situation, which forced universities to stop using their centers as a means of teaching, many of them opted for virtual education. Affecting the learning process of students, which has predisposed many of them to become familiar with this new learning process, making the use of virtual platforms more common. Many educational centers have come to rely on digital tools such as: Discord, Google Meet, Microsoft Team, Skype and Zoom. The objective of the research is to report on the impact of student learning through the use of the aforementioned videoconferencing tools. Surveys were conducted with teachers and students who stated that 66% were not affected in their educational development. Most of them became familiar with the platforms; however, less than 24% qualified that their academic performance has improved, some teachers still have difficulties at a psychological level due to this new teaching modality. In conclusion, teachers and students agree that these tools are a great help for virtual classes.**

Keywords—Digital tools; health emergency; universities; video conferencing; virtual education

I. Introduction

Globally, due to the new pandemic SARS-CoV-2 or known as COVID-19, the use of different softwares for student education is increasing. Less than a decade ago, videoconferencing was only used to avoid unnecessary expenses on business trips, thus reducing savings on lodging and time.

There was an abrupt change with the arrival of this new virus. This has led to a growing need for all students (from kindergarten to university level) to establish and become familiar with this new teaching and learning process, making the frequent use of these platforms more common.

Online class imbalance learning is a new problem that is happening in the real world [1]. In Indonesia, there are two problems that inhibit its education system, namely, transportation and distance. As a first solution, they applied the use of information and communication technologies (ICT). To reach a large mass of students, they created the Virtual Class Box (VCB) 5.0 device to support digital distance learning between teachers and students [2].

On the other hand, China rushed and experimented for a two-month trial period by forcing its universities to work remotely [3], obtaining as a result of their ICT, acceptable academic performance. However, he also stated that online platforms are not explicitly for student purposes.

Latin American countries such as Uruguay, created a digital program designed exclusively for the quality of teaching and learning in students due to the shortage of teachers [4].

One of the ways of interactivity is the so-called Discord application [5], normally used for video games. In their latest updates they added the trend of video calls, based on this, the student-teacher activities improved, consolidating the educational synergy.

In this day and age, conventional education is having gaps when it comes to teaching. Because, due to factors beyond their control, the immobilization of people decreed by many governments worldwide has left them alone at home, unable to go out, much less to places where there is a conglomeration of people. Education at any level has been affected in the delivery of classes [6].

Currently, the need for Peruvian students has increased by 80% the use of online classes. Because virtually the entire 100% of their houses of study will be disabled in person to prevent the spread of the virus from March 16, 2020 to date. By carrying out virtual meeting platforms such as Zoom or Google meet.

Today, they are the most innovative and adaptable means of telecommunication to our new national lifestyle. Since, it allows to establish a more real and visual connection between teachers and students. Helping also not to delay education. A study in 2019 [7], demonstrates that online learning helps university students improve their academic performance.

As it is in one of the Peruvian professional study houses, Universidad de Ciencias y Humanidades, in which the pandemic forced all its teachers to use videoconferencing without having previous knowledge of these, making a sudden change in the teaching methodology, since the virtual meeting platforms such as zoom and google meet had a time limit. Therefore, by having unique videoconferencing platforms, students will be able to receive instructions as if they were in person, and maintain the synergy of real-time intercommunication. However, these habits are not well received by some courses that require mandatory explanations by the physical presence of a teacher; discarding the disadvantages that some may have due to the quality of the network in their homes.

By carrying out this work we will make a research contribution and, above all, we will see that each person has a different learning pace, which leads us to question, with the excellent quality of the platforms, are videoconferences really effective for university teaching and learning?

The objective of this article is to know the impact of different videoconferencing tools that have been used in the teaching and learning process at a University in northern Lima in times of pandemic.

The structure of the article is as follows: Section II, Literature Review, the antecedents will be explained; in Section III, Methodology, the steps to be followed will be detailed; in Section IV Results are discussed; in Section V, Discussions and in Section VI Conclusions and Future Work are discussed.

II. Literature Review

Currently, in times of pandemic, teaching is at a distance where the use of different means of videoconferencing is relevant in education. Since, it has a very significant role in the learning experience of the students [8].

The author [9], indicates that ICT has contributed to the new educational reforms. Google meet was mostly used by students in work meetings as opposed to teachers who preferred to zoom in on class meetings.

The use of the Zoom Videoconferencing platform helps both teachers and students in their work, teaching and learning, which allows both parties to interact and learn about the benefits of the platform, in addition to creating a socially positive learning environment [10]. The author [9], [10] coincide in their methodologies, since both apply a quantitative approach, having as statistical results, where a good percentage of Zoom videoconferences are accepted by students and teachers.

On the other hand, the author [11] mentions in the results of his studies that students do not have problems related to virtual education using videoconferencing platforms, however, there is a latent concern about issues related to the laboratory since, in it, they needed instruments that only the University could provide and that the vast majority of students do not have the possibility of acquiring.

In similar instances, a university student group, in their article [12], emphasized that it is not only in the professional field that institutions use videoconferencing for teaching.

In a study of e-learning, the modern form of online learning, the following benefits were found [13], such as cost-effectiveness, learning flexibility and, above all, the independent part.

In an investigation [14], explains that students have learning effectiveness during their online classes using Discord, as it allows access to requested activities and availability. However, what is detrimental to its reinforcement is the low quality of its network speed and the self-conscious behavior on the part of the students when they do not contribute opinions by activating their microphones.

Students in Indonesia, however, refuted that answer [15], he mentioned that when his students use videoconferencing tools in their classes, the behavior of the students is different, assuring that many of them have managed to overcome the shyness of speech, in which, virtually, they have more courage to contribute their opinions.

III. Methodology

The following will describe the most influential points of Videoconferencing platforms that contribute to virtual education, such as: Discord, Google Meet, Microsoft Teams, Skype, Zoom.

A. Survey

In this work 2 surveys were applied to: 25 university students and 10 teachers of the University of Sciences and Humanities of the career of Systems and Computer Engineering, using the Google Form tool to carry out these surveys. These surveys will help us to collect data from both students and teachers.

-Link to the survey for teachers:

$docs.google.com/forms/d/1T6ZFX6Forz4$
$U_QhLKxHAILyA1ITYQ698i0uq8oAY0Vc$

Table I, lists the questions on the form for teachers.

TABLE I. Teacher Survey Table

N°	Teacher Survey
1	¿At what level do you consider you are psychologically and pedagogically qualified to work with virtual videoconferencing tools?
2	¿How many hours have you increased your workload to develop your online classes?
3	From your perspective, do you feel that stress has built up during the new teaching methodology?
4	¿Do you think you need to improve your digital skills to face the new teaching and learning processes in virtual environments?
5	Nowadays, the use of different videoconferencing platforms has diversified. On a scale of 1 to 5, indicate your knowledge of each of them, where 1 is none and 5 is total.
6	Rate the following tools, indicating the level of use in your online classes.
7	From your point of view, do you feel comfortable with the new teaching modality?
8	Based on the academic performance of your students, consider:

-Link to student survey:

$docs.google.com/forms/d/1o7cqxYVTB_zJd - EoRSS - 8W9LnyAIco5ll8IgiQQ2xJU$

In Table II, we will mention the questions in the survey focused on students.

B. Zoom

Zoom is a tool used for synchronous online teaching, which allows you to work efficiently. It includes several functions such as: annotation tools, polls, meeting rooms and video and screen sharing. These functions facilitate learning [16].Below in Table III, the most important tools of the Zoom platform will be detailed.

It offers innovative learning opportunities and tools, integrating video conferencing and teacher-student communication.

TABLE II. Student Survey Questions

N°	Student Survey
1	¿Do you consider that your level of learning has increased in the new virtual learning environment?
2	¿Does teaching teachers via videoconferencing help you learn?
3	According to the aforementioned. Give a brief explanation of your answer.
4	¿Do you think you need to strengthen your digital skills to adapt to new learning processes in virtual environments?
5	Currently, the use of the different Videoconferencing platforms has diversified. On a scale of 1 to 5, mention the knowledge of each of them, where 1 is nil and 5 is total.
6	Please rate the following tools, indicating the level of use in your online classes
7	¿From your point of view, do you feel comfortable with the new learning modality?
8	¿How often do you have trouble concentrating during your online classes?
9	Based on your academic performance, consider:

TABLE III. Main Zoom Tools

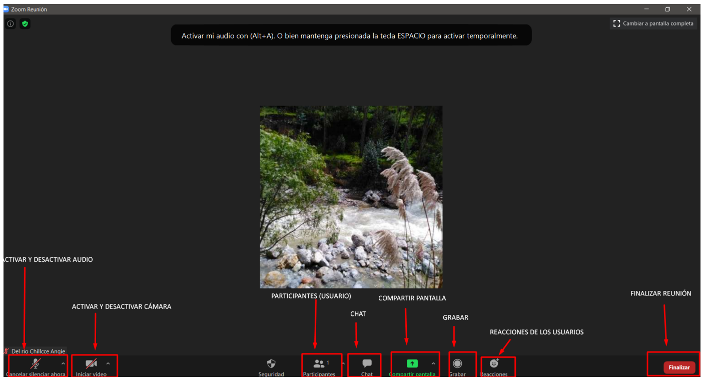
TOOLS	USAGE
Start a meeting	Create a videoconference.
Schedule a meeting	Allows you to schedule a specific day and time for the meeting.
Use of the calendar	Gives the option to use Google calendar, to receive notifications of meetings already scheduled.
Screen sharing	Allows all participants to have the option to choose what to share with other meeting participants
Screen Recorder	It is useful to replay the meeting as many times as you want, it is very helpful to take notes and remember some details.
Virtual whiteboard	Allows you to draw, write or carry out explanations in an easier way.
Chat	Participants have the option to interact both directly and privately.
Live Broadcasting	It is used to make live broadcasts using applications such as Facebook or Youtube.
User management	You have the option to enable and disable the audio and video of the participants, as well as manage which user enters the meeting.

In Table IV, some advantages and disadvantages of the platform are mentioned.

TABLE IV. Table Advantages and Disadvantages of Zoom

Advantages	Disadvantages
The pc version and mobile application has a relatively easy, comfortable and intuitive interface.	Despite being an easy to use tool, it can be confusing for some people who are not adapted to this new technology.
It has a free version	It has a 40 minutes limit for the free version.
Allows screen sharing in real time	As it is a synchronous application the use of internet or a stable connection is important.
No need to be registered to join a meeting, nor download the application	It collects data and emails from all connected devices exposing the user.
A permanent ID will be assigned to the user	It is no necessary to download and install the application to use it.

In Fig. 1, the main and basic Zoom tools are shown.



Activar mi audio con (Alt+A). O bien mantenga presionada la tecla ESPACIO para activar temporalmente.

Fig. 1. Zoom Interface.

C. Google Meet

Google Meet is a relatively recent application [17], was launched in April 2020 to all users, free of charge. It was previously known as Hangouts. It is mainly focused on companies and educational centers.

Below in Table V, the most important tools of the Meet platform will be detailed.

TABLE V. Main Meet tools

TOOL	USE
Create a video call	Allows you to create the video call, by logging in with your Google account.
Invite other users to join a meeting	Gives you a meeting link or code that you can send to other users.
Screen sharing	Allows you to show your screen or the window of an application.
Screen Recording	Allows you to record the sessions so that the student has all the information at hand.
Chat	Users can interact by sharing files and views.
Controls for hosts	The host can mute, set or delete a user.

In Table VI, some advantages and disadvantages of the Meet platform are mentioned.

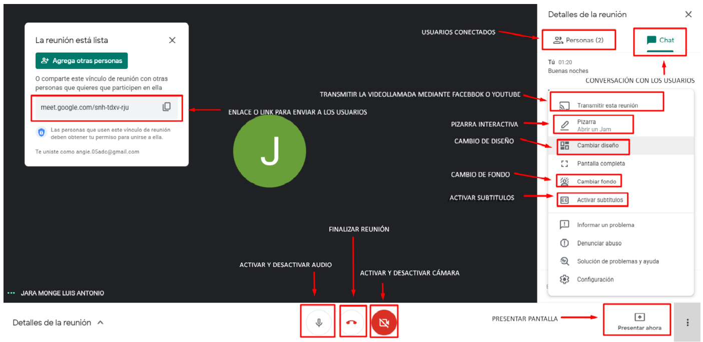
TABLE VI. Advantages and Disadvantages of Meet

Advantages	Disadvantages
Allows you to create meetings with more than 200 participants	60 minutes limit for the free version.
Facilitates real-time captioning during conversation	Has few mechanisms to control user audio.
It has a simple and deductible interface	Each participant must be registered or have a Google account.
The security of the videoconference is guaranteed due to the encryption of the transmissions	it is a synchronous application, this means that the internet connection is indispensable.

Google Meet is closely connected to Google Suite, this makes it possible to add meetings through Event Calender.

In Fig. 2, the main Meet tools are shown.

D. Skype



La reunión está lista

O comparte este vínculo de reunión con otras personas que quieres que participen en ella

meet.google.com/snh-tdxv-rju

Las personas que usen este vínculo de reunión deben obtener tu permiso para unirse a ella.

Te uniste como angie.05adc@gmail.com

Tú 01:20
Buenas noches

Detalles de la reunión

Fig. 2. Meet Interface.

Like the other virtual meeting platforms, this one offers its services at zero prices, i.e. free of charge, so that people can meet at a distance. Despite being an offshoot of Microsoft's own [18]. Table VII shows the features that Skype has.

TABLE VII. Skype Application Features

HERRAMIENTA	USO
Speed per calls	24kbps / 128kbps.
Courier writing	Unlimited for ongoing call.
Call recording	If in case the guests want to see the show again.
Shared screen	It has the ability to transmit live presentations from the phone or computer.
Multi platform	phone, pc, laptops, tablets, has accessibility to Skype.

We continue with Table VIII, the advantages and disadvantages of this virtual meeting platform will be detailed.

TABLE VIII. Advantages and Disadvantages of Skype

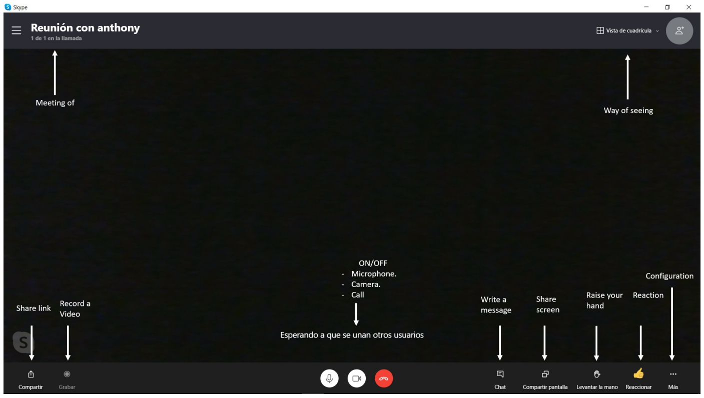
Advantages	Disadvantages
Skype has accessibility for the different Operating Systems such as: Windows, Mac, Linux, Android, iOS.	It is not used mainly for academic meetings, both university and school.
Advanced Encryption Standard Security.	Sometimes during video call, video and audio quality tends to drop, which is annoying for users.
It has the capacity of 50 people connected in real time.	If the time limit exceeds, the video transmission is cut off, however, the call continues as audio.

In the Fig. 3, the tools of said platform are indicated by arrows.

E. Microsoft Teams

This virtual meeting platform is also used by hundreds of students nationwide, launched since 2017 becoming better known by the education sector since the year 2019, as well as providing a good integrated teaching and learning space [19].

Next, it is detailed in Table IX, the characteristics that the platform carries within a videoconference on-line.



Esperando a que se unan otros usuarios

Fig. 3. Skype Platform.

TABLE IX. App Features Microsoft Teams

TOOL	USE
Messenger service	Ability to communicate personally with one of the members.
Time limit	The platform can be used between users for 24 hours at a time.
Time	Shows the time spent inside the room..
Setting	Shows other options available to the application, such as: audio distribution.
Leave	By pressing that option, the member can exit the live session.
Reactions	Interaction of members, such as: raising your hand to give your opinion, then the host will respond to your request.

To achieve a distinction between the aforementioned platforms, it will be detailed in Table X, the advantages and disadvantages of the Microsoft Teams video conferencing platform.

TABLE X. Advantages and Disadvantages of Microsoft Teams

Advantages	Disadvantages
PKI security Protect data used to encrypt transport layer connections.	It has a paid version for the full use of its services.
Has resources at the time of live meetings	It does not allow more than 300 users working simultaneously.
It has a free plan, however, it restricts some tools	A large percentage of users do not feel conformism with the tools, that is, difficult adaptability without prior training.
Your files are stored in the cloud. (Drive)	Requires the mandatory use of the internet.

In Fig. 4, each of the parts that Microsoft Teams has during a live meeting are noted.

F. Discord

The Discord platform is a means of creating a quality remote environment during emergency training. [20]. It is used

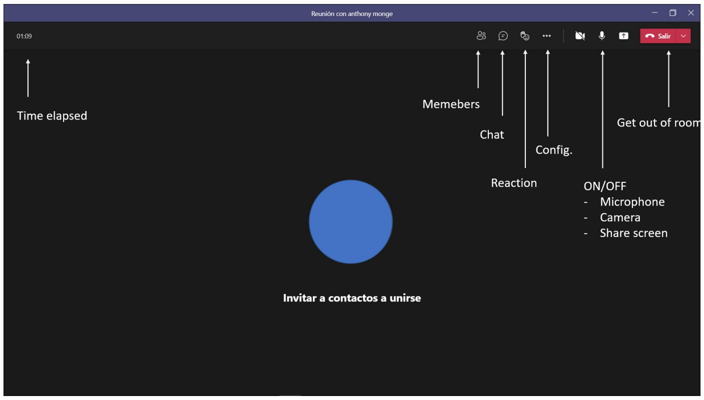
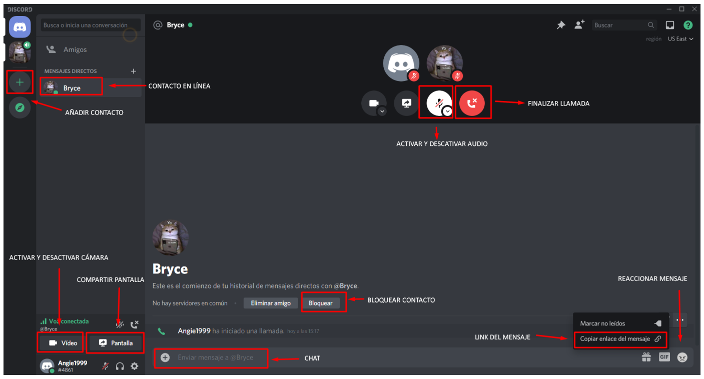


Fig. 4. Microsoft Teams Platform.



Este es el comienzo de tu historial de mensajes directos con @Bryce.

No hay servidores en común

Angie1999 ha iniciado una llamada. hoy a las 15:17

Fig. 5. Discord Conference.

to convert a traditional class to a virtual class, a monotonous listening class into an engaging listening class. In Table XI, we will proceed to list some of its tools.

TABLE XI. Discord App Features

TOOL	USE
Add contact	Be able to add a user to make the video call.
Create video call	Lets create the video call.
Share screen	Allows video call participants to choose what to share.
Chat	It serves to have a more direct communication.

The following is detailed in Table XII, some advantages and relevant disadvantages of the videoconferencing platform Discord.

TABLE XII. Advantages and Disadvantages of the Discord Application

Advantages	Disadvantages
Due to the optimized performance, you can communicate without losing call quality.	After the launch of the paid version (Discord Nitro), limited some features in version (Discord Classic), and much more in its free version.
The tool is cross-platform.	Requires previous training for its use.
The platform is free.	The video calling option is not optimized, so there is a chance it will fail.
It is customizable for a better user interface.	You do not have the option to record video calls.
Ability to provide class with up to a maximum of 50 people.	Discord is not available for the Linux Operating System.
Technical support available to your users.	He had complaints from scammers.

In Fig. 5, the main interface of the Discord tool is shown, in addition to having a good educational and group environment, with a knowledge of the application there will be no deficiencies in its use. Multiple options are displayed within the interface that can be performed during a call in progress, either by transmission or just aural.

Through the comparisons made, it can be clearly concluded that the platforms have similar characteristics as: Time in which the class takes place, interaction option, screen sharing, writing internal messages to a user as well as in public. Except for Discord, for not having the option of video recording.

Then, the videoconferences differ between them by one or another option of use, therefore, the student's learning and the teacher's teaching still remains to be analyzed, Therefore, the results of the survey carried out will be explained by means of graphics.

IV. Results

A. Regarding the teachers surveyed:

- The 66.6% considers that he is highly trained physically and psychologically to work with videoconferencing tools, as well as the 33.4% consider it to be at an intermediate level. As shown in Fig. 6.

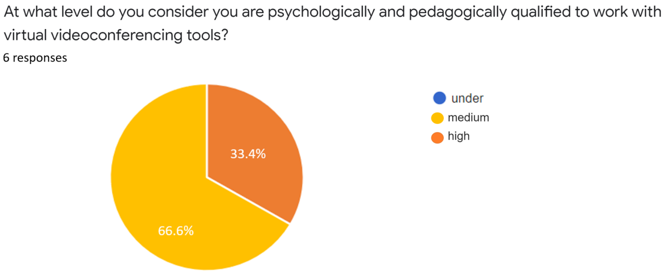


Fig. 6. Questions about Teacher Training.

- The 50% dedicates more than 3 hours (than normal) a day to develop his class, meanwhile the 33.3% dedicate 3 hours a day and 16.7% he claims he only spends 2 hours. As shown in Fig. 7.
- The 66.4% of teachers assess that they have accumulated a medium stress with the new teaching methodology, while the 16.8% considers that he accumulated high stress, in the same way the 16.8% accumulated low stress. As shown in Fig. 8.
- The 100% consider that you need to keep learning the new digital tools. As shown in Fig. 9.

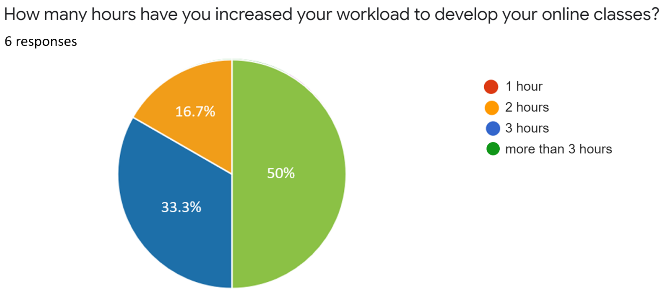


Fig. 7. Ask about the Hours Dedicated to Developing the Class.

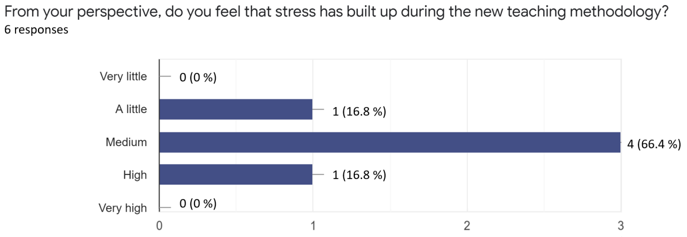


Fig. 8. Question about Accumulated Stress in Teachers.

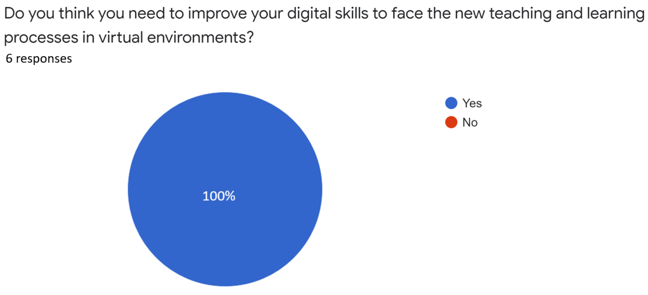


Fig. 9. Ask about the New Digital Tools.

- The vast majority of teachers have a very high knowledge of the Zoom and Meet platforms, however they have a medium-low knowledge of the Skype, Microsoft Teams and Discord tools. As shown in Fig. 10.
- Almost all teachers confirmed that they use the Zoom and Google Meet tools more frequently to teach their classes, as shown in Fig. 11.
- Only 16.7% considers that he feels very comfortable with this teaching modality; however, the 83.3% consider that you are comfortable with this mode of teaching. As shown in Fig. 12.

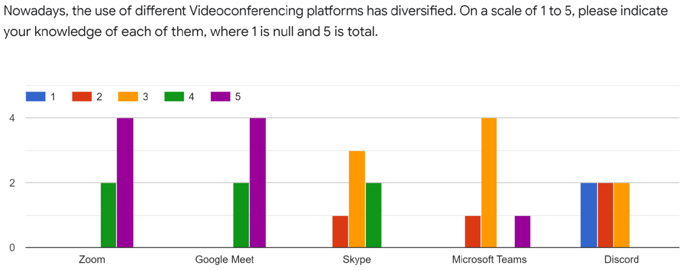


Fig. 10. Question about the Knowledge of Videoconferencing Tools.

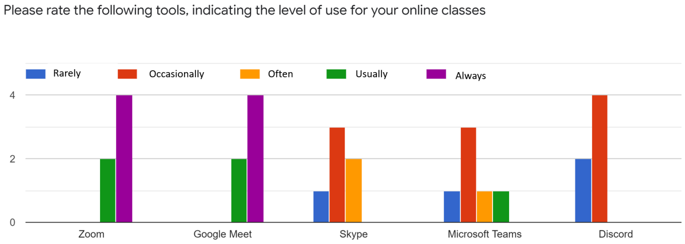


Fig. 11. Question about the use of Videoconferencing Tools.

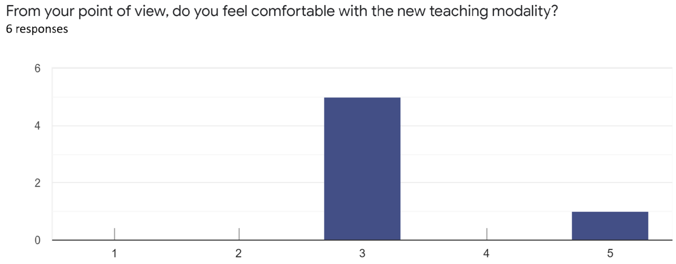


Fig. 12. Ask the Teacher about this New Teaching Method.

- An 83.4% of teachers consider that students' academic performance has remained on the sidelines and only 16.7% indicates that it has improved. As shown in Fig. 13.

B. Regarding the Surveyed Students:

- The 60% of students rate that their learning level has remained in a normal state, the 20% estimates that it has improved, and 8% qualifies that he has learned little; however, the 8% he considers that he has learned very little. As shown in Fig. 14.

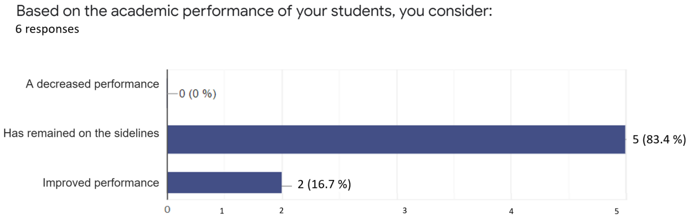


Fig. 13. Question about the Student's Academic Performance.

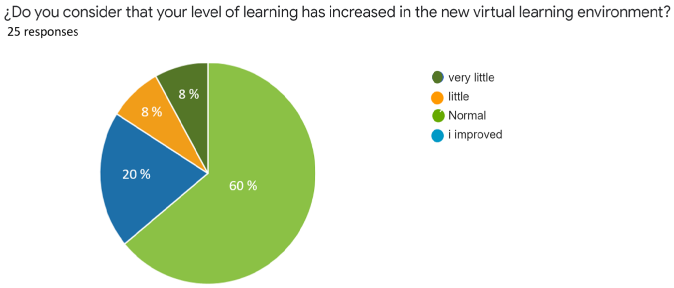


Fig. 14. Ask about the Student about their Level of Learning in New Virtual Learning Environment.

- The 68% of the surveyed students rate that they agree that video conferencing tools help them learn; however, the 32% considers that they do not agree. As shown in Fig. 15.

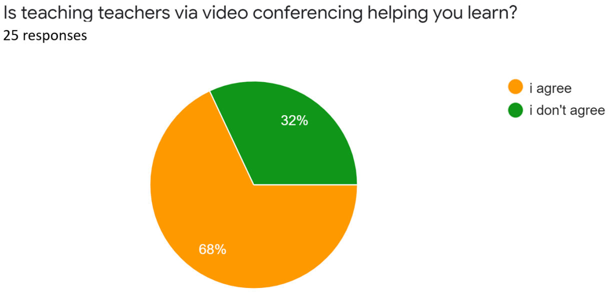


Fig. 15. Ask about the Student about their Level of Learning via Video Conferencing.

- The 72% of respondents believe that they need to continue strengthening their digital knowledge, although, 28% consider not.

- The vast majority of students have a medium-high knowledge of the Zoom and Meet platforms, however, they have low-zero knowledge of the Skype, Microsoft Teams and Discord tools. As shown in Fig. 16.

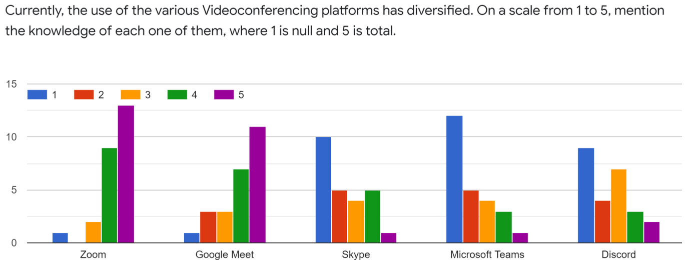


Fig. 16. Ask the Student about Knowledge of Videoconferencing Tools.

- Almost all students confirmed that they most frequently use the Zoom and Google Meet tools for their online classes. As shown in Fig. 17.

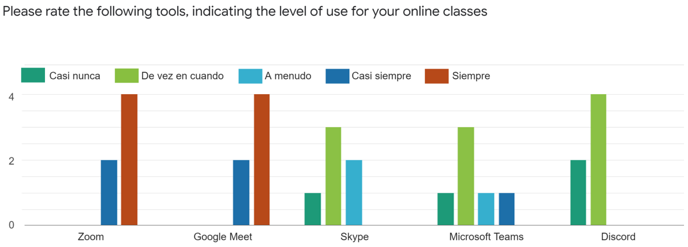


Fig. 17. Ask the Student about the use of Videoconferencing Tools.

- The 16% considers that he feels very comfortable regarding the new learning modality, a 20% rate that you feel comfortable, the 44% considers that it is normal, the 16% estimates that he feels uncomfortable, in the same way a 4% He rated that he is not comfortable. As shown in Fig. 18.

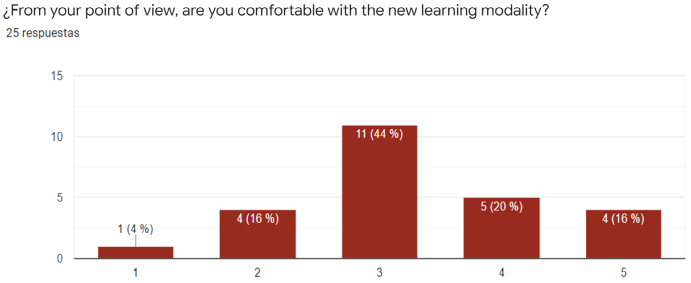


Fig. 18. Ask the Student about their Learning Comfort.

- The 52% indicated that he very seldom has concentration problems during classes, only on the 32% who has problems often, and the 16% indicates constantly has problems. As shown in Fig. 19.

- The 24% rated that their performance has improved considerably, the 60% considers that his performance has remained on the sidelines, and only the 16% indicates that its performance has decreased. As shown in Fig. 20.

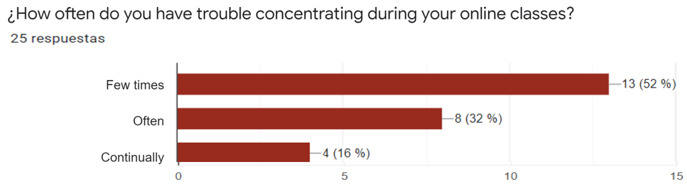


Fig. 19. Ask the Student about his Concentration During Class.

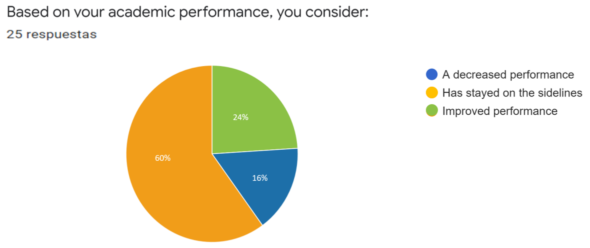


Fig. 20. Ask the Student about their Academic Performance.

V. Discussions

During the points mentioned and analyzed for the use of these different virtual platforms, the financing aspect is essential due to various factors. Both by the student entity (University or college), as well as by the students for the usability of the conference applications [21].

According to the author [22], mentions that Zoom features are optimal for conducting video conferencing in an intuitive way. Another Study Reveals About Improving Scores During Online Exams [23]. That is, in 2020 there were higher grades in its average. Which means that students' grades have risen significantly since e-learning was introduced at the beginning of the pandemic. In the results of [24], Through their survey, they detailed that the majority of students in a 56.3 % They assured that the most used and / or known platform is Zoom. Obtaining a result similar to our work as the most used videoconferencing application for this new mode of education. He also added, and we agree, that the solution of practical exercises such as: mathematics; require more than distance learning.

The results obtained in comparison with the author [10] they coincide, as, video conferencing platforms help teachers and students in teaching and learning work.

VI. Conclusions and Future Work

In conclusion, regarding the results obtained from the surveys, the vast majority of teachers and students perform at an advanced level in the platforms for the virtual teaching and learning environment. Likewise, teachers have constant and fluid communication with their students during class partitions. However, some teachers still have difficulties at a psychological level due to this new teaching modality. The teachers mentioned that they are in constant training to learn the new digital tools. Students feel comfortable and motivated during their virtual classes.

The contribution of this research work is aimed at the educational environment, making the results known through online learning surveys. The academic community of the University, made up of students, teachers and authorities, benefit from this work.

Further research is expected later, adding the other departments and comparing little-known video conferencing platforms.

References

[1] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1356–1368, 2015.

[2] J. Nainggolan, G. Christian, K. Adari, Y. Bandung, K. Mutijarsa, and L. B. Subekti, "Design and implementation of virtual class box 5.0 for distance learning in rural areas," in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2016, pp. 1–6.

[3] F. Lu, X. Chen, X. Ma, Z. Liu, and Y. Chen, "The exploration and practice of it solutions for online classes in higher education during covid-19 pandemic," in *2020 International Symposium on Educational Technology (ISET)*, 2020, pp. 298–302.

[4] C. Marconi, C. Brovetto, I. Mendez, and M. Perera, "Learning through videoconference. research on teaching quality," in *2018 XIII Latin American Conference on Learning Technologies (LACLO)*, 2018, pp. 37–40.

[5] M. Vladoiu and Z. Constantinescu, "Learning during covid-19 pandemic: Online education community, based on discord," in *2020 19th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 2020, pp. 1–6.

[6] C. Diaz-Nunez, G. Sanchez-Cochachin, Y. Ricra-Chauca, and L. Andrade-Arenas, "Impact of mobile applications for a lima university in pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120294

[7] A. García and E. Vidal, "Mobile-learning experience as support for improving the capabilities of the english area for engineering students," in *2019 International Conference on Virtual Reality and Visualization (ICVRV)*, 2019, pp. 202–204.

[8] A. F. Azmi, R. Nuravianty, T. I. Nastiti, and D. I. Sensuse, "Using social networking sites for learning experiences by indonesian university students," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 177–182.

[9] Y. Bandung, D. Tanjung, and L. B. Subekti, "Design of mleaming application with videoconference system for higher education," in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, 2017, pp. 1–6.

[10] J. Sutterlin, "Learning is social with zoom video conferencing in your classroom," *ELearn*, vol. 2018, no. 12, Dec. 2018. [Online]. Available: https://doi.org/10.1145/3302261.3236697

[11] E. E. A. Rahim, N. Daud, S. A. A. Kadir, and N. W. Jamil, "Students' perceptions of open and distance learn-

ing (odl) for theoretical and lab-related subjects," in *2020 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, 2020, pp. 29–32.

[12] J. S. César, R. G. Elar, T. E. Jhonathan, T. A. Anthony, V. R. Gary, and A. A. Laberiano, "Analysis of the use of technological tools in the e-learning process," in *2020 IEEE ANDESCON*, 2020, pp. 1–6.

[13] Rismayani and Y. J. W. Soetikno, "Using webqual 4.0 for measuring quality of e-learning services during covid-19 pandemic," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 2020, pp. 1–7.

[14] Y. Maher, S. M. Moussa, and M. E. Khalifa, "Learners on focus: Visualizing analytics through an integrated model for learning analytics in adaptive gamified e-learning," *IEEE Access*, vol. 8, pp. 197 597–197 616, 2020.

[15] A. Sufyan, D. Nuruddin Hidayat, A. Lubis, U. Kultsum, M. Defianty, and F. Suralaga, "Implementation of e-learning during a pandemic: Potentials and challenges," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 2020, pp. 1–5.

[16] L. Kohnke and B. L. Moorhouse, "Facilitating synchronous online language learning through zoom," *RELC Journal*, p. 0033688220937235, 2020. [Online]. Available: https://doi.org/10.1177/0033688220937235

[17] R. S. Al-Maroof, S. A. Salloum, A. E. Hassanien, and K. Shaalan, "Fear from covid-19 and technology adoption: the impact of google meet during coronavirus pandemic," *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–16, 2020. [Online]. Available: https://doi.org/10.1080/10494820.2020.1830121

[18] Y. Tabira and S. Goto, "Impact of international postures on willingness to communicate during international exchanges using skype," in *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, 2017, pp. 1–5.

[19] D. Pal and V. Vanijja, "Perceived usability evaluation of microsoft teams as an online learning platform during covid-19 using system usability scale and technology acceptance model in india," *Children and Youth Services Review*, vol. 119, p. 105535, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0190740920313992

[20] V. Kruglyk, D. Bukreiev, P. Chornyi, E. Kupchak, and A. Sender, "Discord platform as an online learning environment for emergencies," *Ukrainian Journal of Educational Studies and Information Technology*, vol. 8, no. 2, pp. 13–28, Jun. 2020. [Online]. Available: https://www.uesit.org.ua/index.php/itse/article/view/303

[21] N. Nasrat, A. Khamosh, and K. Lavangnananda, "Challenges and hurdles to e-learning implementation during covid-19 outbreak: A case of shaikh zayed university," in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2020, pp. 242–246.

[22] Y. Chaiko, N. Kunicina, A. Patlins, and A. Zhiravetska, "Advanced practices: web technologies in the educational process and science," in *2020 IEEE 61th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*, 2020, pp. 1–6.

[23] R. M. Coșniță, A. Maria Cimpean, and M. Raica, "Online versus on-site e-assessment in medical education: are we ready for the change?" in *2020 International Symposium on Electronics and Telecommunications (ISETC)*, 2020, pp. 1–4.

[24] S. M. Mambo and F. Makatia Omusilibwa, "Effects of coronavirus pandemic spread on science, technology, engineering and mathematics education in higher learning institutions," in *2020 IFEES World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, 2020, pp. 1–4.

Network Forensics: A Comprehensive Review of Tools and Techniques

Sirajuddin Qureshi[1], Saima Tunio[2], Faheem Akhtar[3], Ahsan Wajahat[4], Ahsan Nazir[5], Faheem Ullah[6]
Faculty of Information Technology,
Beijing University of Technology, Beijing 100124, China.[1,2,4,5,6]
Department of Computer Science, Sukkur IBA University, Pakistan[3]

***Abstract*—With the evolution and popularity of computer networks, a tremendous amount of devices are increasingly being added to the global internet connectivity. Additionally, more sophisticated tools, methodologies, and techniques are being used to enhance global internet connectivity. It is also worth mentioning that individuals, enterprises, and corporate organizations are quickly appreciating the need for computer networking. However, the popularity of computer and mobile networking brings various drawbacks mostly associated with security and data breaches. Each day, cyber-related criminals explore and devise complicated means of infiltrating and exploiting individual and corporate networks' security. This means cyber or network forensic investigators must be equipped with the necessary mechanisms of identifying the nature of security vulnerabilities and the ability to identify and apprehend the respective cyber-related offenders correctly. Therefore, this research's primary focus is to provide a comprehensive analysis of the concept of network forensic investigation and describing the methodologies and tools employed in network forensic investigations by emphasizing on the study and analysis of the OSCAR methodology. Finally, this research provides an evaluative analysis of the relevant literature review in a network forensics investigation.**

Keywords*—Network forensics; Tshark; Dumpcap; Wireshark; OSCAR; network security

I. Introduction

The evolution of computer networks and the internet has created many opportunities for the perpetration of cyber-related crimes. Numerous computing devices are connected to a complex mesh of computer networks all over the globe. Cyber attackers are continuously adapting complicated strategies to perpetuate cyber-related crimes. The nature and the type of crimes are costly to the affected victims [1]. In some instances, the committed cybercrimes not only cause significant financial losses but might also render the affected organization inoperable. Thus, it is essential to have a mechanism of performing necessary investigation and audit to establish the course and the perpetrators of the associated cybercrimes. In the context of cyber-criminal investigations, the mechanism is referred to as network forensics.

Network forensics is a digital forensic process that involves the investigation, Analysis, and monitoring of computer networks to discover essential information that helps in the apprehension of cybercriminals [2]. Network forensics also helps in gathering necessary and legal information, evidence, and traces of intrusion detection. In essence, network forensics helps a cyber-forensic investigator monitor network traffic and identify any malicious content within network traffic. Network forensics is data-centric, and thus it is not primarily restricted to the Analysis of network traffic. Instead, it is also associated with related concepts, notably mobile forensics, memory forensics, and host-based forensics [1].

Primarily recent Internet technology advances drive the evolution of network security and its associated forensic processes and related toolsets. When more facets of our everyday lives move to electronic networks and databases where they are vulnerable to illegal activity, there is a growing need for advanced analytical resources. Some widely mentioned explanations for the use of network forensics are based on

- Analysis of computer systems belonging to victims or authorities.
- Collection of facts for use in court; Recovery of lost data in the event of software and hardware failure.
- Analysis for a computer system after a break-in.
- Collection of information about how the computer systems function for debugging purposes, optimization of their computer systems The list only scratches the surface of what network forensics can do in the sense of risk management and data recovery;

The following example illustrates the critical role that this technology can play in an investigation process. The companies usually use different items when it comes to network security. Such devices typically approach protection from two main perspectives; detection and monitoring, in other words. Types of items for protection include firewalls and systems for access controls. Likewise, the intrusion detection systems and antivirus software are examples of detection products. Although the used products foil several attacks, novel attacks often bypass protection products without being detected. Investigating the attacks in these cases is a challenging job. Serious attackers are, in many cases, skilled at removing evidence. Consequently, firewall logs and intrusion detection warnings that miss such attacks entirely or may prove insufficient for a thorough investigation, mainly when the goal is to apprehend the attacker.

Network forensics has been suggested in information security literature to incorporate investigative capabilities in existing networks. This refers to a dedicated research infrastructure that enables network packets and events to be captured and evaluated for research purposes. Complementation of the above Network Security optimization is performed. The forensic network is of significant importance to companies worldwide. On the one hand, it helps learn the specifics of recent threats, ensuring that potential attacks are thwarted. Furthermore, network forensics is essential to investigate the abuses of insiders that constitute the second most costly model of corporate assault. Lastly, law enforcement refers to network forensics for cases in which a device or digital machine is either the object of a crime or used to carry a criminal offense.

Network forensics is a complex phenomenon that needs the utilization of a variety of tools and methodologies. It is thus essential to have a good understanding of how these tools and techniques can aid in the process of network forensics and the discovery of malicious activity and intrusion attempts. This paper aims to provide a comprehensive description of network forensics' concept to understand the tools and methodologies used. Emphasis is based on giving a vivid portrait of the OSCAR methodology as used in network forensics. An analysis and review of critical related works that illustrate the practical implementation of the network forensics concept are extensively discussed.

II. Related Work

The field of network forensics attracts diverse interests that ultimately have led to the publication of various research works aimed at bridging the knowledge gap within the topic domain. In particular, much of the related works in the field of network forensics is related to security. It is essential to note that any network provided that is connected to the internet is prone to a variety of cyber-attacks. The attacks are generally designed in such a way that they exploit ay vulnerabilities within the network. A forensic investigator is thus tasked with the responsibility of coming up with essential strategies to perform a comprehensive network forensic process to identify potential cases of network intrusion [3]. In addition to the fact that the legislature has borne some of the cost of crime prevention, company secrets are compelled to utilise the most dynamic security measures available to secure their essential information [4].The advent of information and communication technologies has ushered in a new era of human existence known as the information society. As the most well-known product of this community, cyberspace has provided people with enormous opportunity to search for and store large volumes of data. This has not only improved the visibility of information, particularly scientific and economic conclusions, but it has also resulted in an increase in targeted cyber-attacks aiming at gaining unauthorised access to such sensitive data. Meanwhile, the concept of safeguarding trade secrets has taken on new significance as information with independent economic or competitive worth [5]. One of the numerous issues that trade secrets have produced as valuable and sensitive knowledge as a result of the expanding space of information and communication interchange is the widespread response of governments to the use of coercive instruments with powerful deterrent effects, such as Terry's case [6]. This research comprehensively discusses it as discussed in the related domain [7], [8], [9], [10], [11].

A. Network Security and Network Forensic

Apart from assisting in identifying and apprehending cyberterrorists and attackers, network forensics also plays a significant role in extending the security model within a network. As noted by Almulhem, network forensics helps network administrators to enhance the prevention and detection of network and cyber-related attacks. In essence, network forensics makes it possible to perform a comprehensive vulnerability analysis process to identify potential threats facing a network [12]. Almulhem adds that network forensics is more associated with a security model than a product or service aimed at enforcing security or network prevention. Instead, network or digital forensics emphasizes the design and implementation of methodologies, tools, and concepts that aim to enhance the process of forensic investigation [12].

Kilpatrick et al. suggest the implementation of SCADA (supervisory control and data acquisition systems that form a vital infrastructure for network forensics [13]. SCADA networks are essential for forensic investigations in that the underlying architecture makes it possible to analyze, monitor, and monitor network behavior [13]. In particular, the SCADA network forensics makes it possible to design and build robust SCADA networks. This is because traffic analysis is an essential constituent of the architecture of a SCADA network.

Network forensics also plays a significant role in the implementation of security mechanisms in the machine to machine networks (M2M) [14]. M2M networks utilize artificial intelligence and machine learning to improve the communication process. Network forensics is used to identify security issues in M2M networks by implementing two distinct modules; forensic and attack detection module. Further, a forensics strategy that uses anti-distributed honeypot is used to aid in detecting and preventing DDoS attacks [14].

To illustrate and reiterate the importance of network forensics investigations, it is paramount to review several case studies whereby the concept has been adequately implemented. Particularly, Kurniawan and Riadi [15] managed to explore and device a unique framework upon which it was made possible to utilize the concept of network forensics to analyze and identify the behavior of the notorious Cerber Ransomware. The approach is aimed explicitly at establishing an attempt to reconstruct the timestamp of an attack [15]. Focus is placed on the need to exact malware deemed to have infected a particular network host. The eventual results indicate that analysis of network forensics behavior can identify patterns of infections, exploits channels, and the ultimate payload associated with the Cerber Ransomware.

1) Network Security Forensic Mechanisms: A firewall within a network environment provides a network forensic investigator with a perfect opportunity to conduct a comprehensive analysis of all the previous network intrusion attempts. As noted by Messier, the majority of firewall systems are equipped with the ability to either implement the software capability in UNIX or Windows [16]. Consequently, a forensic network investigator can either analyze Syslog or Event Logs files to identify and analyze the nature of intrusion activities within and targeted towards a network. An analysis of firewall logs is also essential. It greatly assists in identifying the existing security vulnerabilities and eventually enables the security administrator to develop essential security enhancements.

Bensefia and Ghoualmi reiterate the importance of having a unique branch of network forensics primarily dedicated to analyzing firewall logs [17]. Firewall forensics is a dedicated effort aimed at analyzing firewall logs with the specific objective of gaining useful insights regarding the nature of attacks identified and blocked by the network firewall. While the contents of a firewall log file might be difficult to decode, it is noteworthy to provide essential information that will eventually help a cyber-forensics investigator apprehend a suspected cybercrime offender.

2) Honeypot Forensics: A honeypot is a specialized part of a computer or network system that is designed is such a way that it appears and seems to have critical and sensitive information. At the same time, in a real sense, it is mainly isolated from the main network. An elaborate illustration of how a honeypot device(s) is placed in a network is indicated in Fig. 1. It is worth noting that most of Honeypot's services are secret though it is difficult to assert their suspicious nature [18]. Honeypots are considered to be essential components that help to enhance the security of an organization [19]. Having a honey port within a network makes it possible for a forensic investigator to conduct a comprehensive analysis of all the possible network-related activities and logs carried throughout the honeypot device. Additionally, network forensic investigators are in a good position to perform a comparative analysis of the data obtained from the Honeypot with similar data extracted from other network devices. A network forensic investigator must perform a comprehensive analysis of the existing honeypots in a network whereby the interaction level can be categorized as low, medium, or high level.

Network forensics is restricted to the analysis of firewalls and honeypots systems, but instead, it is widely applicable among most popular network devices. IDS and IPS are perhaps some of the most common types of devices and systems that are commonly targeted by a network forensic investigator to obtain essential cyber forensic evidence that will culminate with the apprehension of a cyber-forensic offender [19]. Routers and switches also provide essential value in that it is possible to obtain essential intrusion information from MAC address tables, ports, and routing tables, among others. Web proxies, as well as, special types of servers such as DCHP, name, and application servers also provide a network forensic investigator with rich information aimed at obtaining crucial cyber forensics evidence [19].

III. Network Forensics

Network forensics is a scientific method used to discover and retrieve information with evidential value and is used to solve a cyber-crime or apprehend a cyber-criminal. The evidence is retrieved from network and computing devices such as hard disks, routers, switches, memory devices, wireless devices, and mobile devices. Table I provides additional information related to possible viewpoints based on potential areas where the forensic investigation could be performed. Network forensics differs from intrusion detection in that the gathered evidence should be admissible in a court of law and thus should satisfy both legal and technical requirements [20]. Consequently, for forensic evidence to be accepted in a court of law, it must be authentic, relevant, complete, reliable, and believable. It is also noteworthy that the tools and techniques used to perform network forensics should also meet a court of law's legal and technical requirements.

While intrusion detection helps strengthen and improve a computer network's security, network forensics is primarily associated with the need to identify the evidence related to a security breach. In most cases, network forensics helps to solve matters related to cyber-terrorism, child pornography, narcotics, homeland security, online fraud, and corporate espionage, among others [21]. Public police mostly use the evidence obtained from network forensics and private investigators working for individuals, businesses, law enforcement agencies, and even the military [20]. It is also essential to note that business organizations and the military might also use network forensics to ensure continuity and availability of core services. In this context, network forensics help to identify vulnerabilities in corporate networks that make it convenient to implement the necessary security enhancements.

The context of the discussion offered in the paper is to explore the investigative purposes of network forensics. The investigation process starts with identifying a malicious activity upon which the evidence is then collected and preserved. The forensic activity proceeds to examine and analyze the evidence to establish the source and the nature of the malicious activity. Finally, the evidence is reported and presented to the relevant stakeholders and eventually used to make the required decision. All the essential processes involved in network forensic investigation are strategically executed using OSCAR principles that are explained in the next section.

IV. Network Forensics Methodology (OSCAR)

To ensure that forensic evidence is both accurate and reproducible, the OSCAR methodology of Network Forensics Investigation is applied. OSCAR [22] is an acronym that stats for,

- O for Obtaining Information
- S for Strategizing
- C for Collecting Evidence
- A for Analyzing Evidence
- R for Reporting

Fig. 2 illustrates the flow chart model for the OSCAR methodology.

A. Obtaining Information

This stage is associated with obtaining information regarding the incident itself and the environment in which the event took place. It is essential to collect as much information about an event to know exactly what took place. Usually, it is advisable to collect information on the description of the incident, time, date, and how it was discovered [15]. Other entities related to the event include the systems, persons, and devices involved and the summary of actions taken after the incidence discovery. It is also essential to note details about the review of discussions made, any legal issues, and the identity of the incident manager. The environment helps the forensic investigator have a good understanding of the organization's response towards an incident and the stakeholders who should be involved in the investigation process [23]. It is thus vital to collect as much information related to the organization as

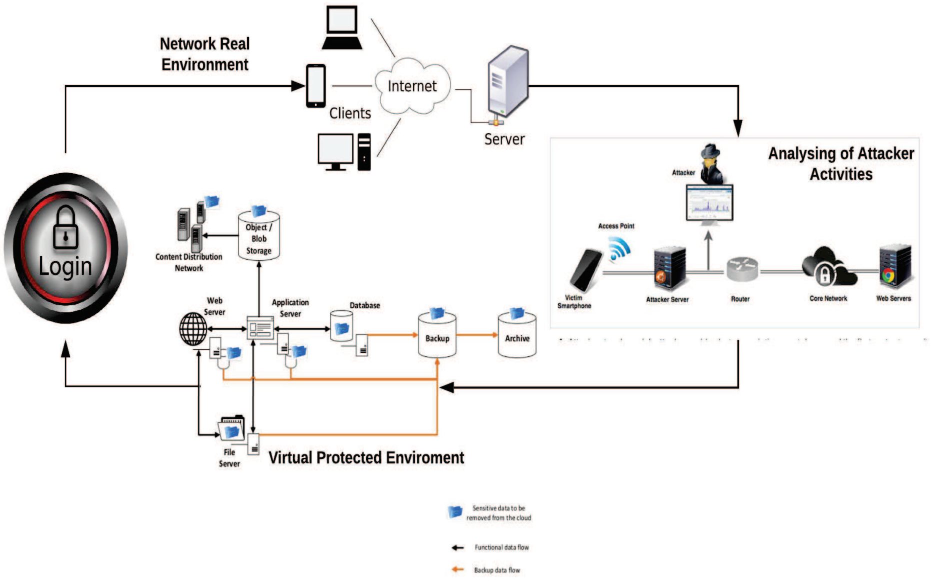


Fig. 1. Logical Placement of a Honeypot within a Network.

TABLE I. Presents Additional Information Related to Possible Viewpoints based on Possible Areas where Forensic Investigation Could be Performed

View Point	Nature of Forensics
Application	Internet browser, email, register files, application software, virus, worm, Trojans, and files (slack, erased, and swap)
System	UNIX, Windows, log system, and audit system
Hardware	PC, PDA, printer, router, switches, firewall, and IDS
Processing	Victim's, intermediate's, and attacker's side

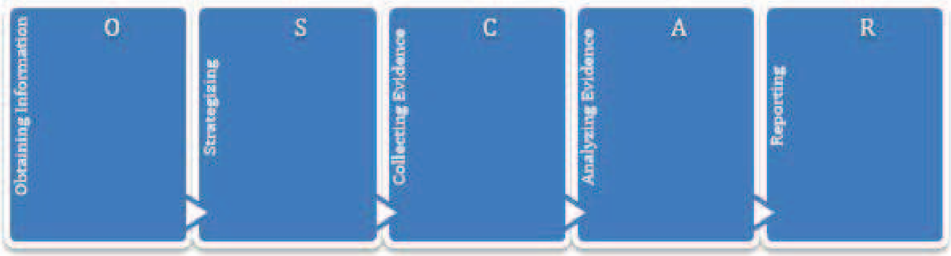


Fig. 2. Network Forensics Investigation Methodology (OSCAR).

possible. Relevant information includes the business model, any legal issues, available resources, communication systems, network topologies, and the procedures and processes used for incidence response management.

B. Strategy

Strategy requires the formulation of a detailed plan on how to carry out the investigation. Strategizing also details how evidence will be acquired [15]. This should be done using various criteria, mainly because pieces of evidence from different sources have varying levels of volatility. As indicated in Table II, the acquisition of proof should be based on several parameters such as source, the effort required, volatility, and the expected value. Evidence prioritization is vital because it helps the forensic to establish the priority of assigning personnel and resources required in network forensics. An important aspect worth noting is that each organization has different policies associated with data retention, access, and configurations [22]. Consequently, the evidence prioritization should be based on specific organizational policies. When formulating an evidence acquisition strategy, it is paramount to consider the following tips.

- Understand the goal of investigation and time frame
- List of your recourses
- Prioritize your evidence acquisition
- Estimate the value and cost of obtaining evidence
- Identify sources of evidence
- Plan to initial analysis
- Keep in mind that network forensics is a process that can be performed reiteratively

C. Collecting Evidence

The strategizing step requires the formulation of an acquisition plan and prioritization of evidence sources. Evidence used in network forensics can be obtained either from the end or intermediate devices [22]. In the former, the evidence can be gathered from the attacker's or the victim's devices, while in the latter, evidence can be obtained from third-party devices and networks. A summary of the probable sources of evidence is provided in Table III.

The next step is to collect evidence from the identified sources using the established priority. Consequently, three vital components must be considered, notably documentation capture, and store or transport.

Documentation: This means that all actions, including a list of all systems, files, and resources, should be carefully logged. It is also essential to maintain self-descriptive notes that make it easy to identify the collected evidence. The descriptive content should contain the date, time source, investigating officer, and the method used to acquire the evidence. Ensure that all devices accessed and all actions were taken during the gathering of evidence are kept to a careful log. Your notes must be kept appropriately and can be cited in court. If the case is not going to court, the notes will also be very helpful during the review. Make sure to document the date, time, source, acquisition process, investigator name(s), and custody chain.

Capturing: evidence involves ensuring that the data or network traffic packets, as well as logs, are written to a hard, CD, or removable hard drive. Network forensics tools such as Wireshark and tcpdump are used to capture data packets [15].

Store/Transport: implies that the evidence should be stored in a secure place to maintain the chain of custody. It is essential to keep updated and signed log containing the details of all the parties who have obtained access to the evidence. Care should also be exerted when handling and disposing of evidence to maintain its integrity, reliability, and admissibility before a court of law.

TABLE II. PRESENTS EXAMPLE OF PRIORITIZATION OF EVIDENCE THAT LIST POSSIBLE SOURCES OF PROOF IN THE CASES, THE PROBABLE VALUE, LIKELY EFFORT OF OBTAINING AND THE EXPECTED VOLATILITY. FOR EVERY INVESTIGATION THESE PRINCIPLES WERE SELECT DISTINCT

Source of Evidence	Likely value	Effort	Volatility	Priority
Web Proxy Cache	High	Low	Medium	1
Firewall logs	High	Medium	Low	2
ARP tables	Low	Low	High	3

In summary, the following tips are crucial during the process of evidence collection.

- Obtaining evidence as soon as possible.
- Make verifiable steganography copies of collected evidence.
- Use reliable and reputable tools
- Document everything, which helps you later.
- Keep secure your notes and hide the original under restricted custody and access.

D. Analyze

The strategizing step requires the formulation of an acquisition plan and prioritization of evidence sources. Evidence used in network forensics can be obtained either from the end or intermediate devices [22]. In the former, the evidence can be gathered from the attacker's or the victim's devices, while in the latter, evidence can be obtained from third-party devices and networks. A summary of the probable sources of evidence is provided in Table III.

The next step is to collect evidence from the identified sources using the established priority. Consequently, three vital components must be considered, notably documentation capture, and store or transport.

Documentation: This means that all actions, including a list of all systems, files, and resources, should be carefully logged. It is also essential to maintain self-descriptive notes that make it easy to identify the collected evidence. The descriptive content should contain the date, time source, investigating officer, and the method used to acquire the evidence. Ensure that all devices accessed and all actions were taken during the gathering of evidence are kept to a careful log. Your notes must be kept appropriately and can be cited in court. If the case is not going to court, the notes will also be very helpful during the review. Make sure to document the date, time, source, acquisition process, investigator name(s), and custody chain.

Capturing: evidence involves ensuring that the data or network traffic packets, as well as logs, are written to a hard, CD, or removable hard drive. Network forensics tools such as Wireshark and tcpdump are used to capture data packets [15].

Store/Transport: implies that the evidence should be stored in a secure place to maintain the chain of custody. It is essential to keep an updated and signed log containing the details of all the parties who have obtained access to the evidence. Care should also be exerted when handling and disposing of evidence to maintain its integrity, reliability, and admissibility before a court of law.

In summary, the following tips are crucial during the process of evidence collection.

E. Report

This is perhaps the most crucial aspect of forensic investigation primarily because it helps to convey the results to the concerned parties. Thus, it is vital to present the report in a manner that can be understood by a lay and non-technical

TABLE III. Provides Evidence used in Network Forensics that can be Obtained Either from the End or Intermediate Devices

Affiliation	Source
End side (attacker and/or victim side)	Operation system audit trail, system event log, application event log, alert log, recovered data, and swap files
Intermediate	Traffic data packets, firewall log, IDS log, router log, and access control log

TABLE IV. Tools & Devices Use for Various Testing Applications

Device/Tool	Usage	Software/OS Version	Company/Developed
Mac-Book Air	Create a test network, host proxies	macOS Siera (10.12.6)	Apple
iPad	Test device connected to test network	iOS 11.2.6	Apple
Charles Proxy	Capture/save live network traffic	4.2.5	Karl von Randow
Wireshark	Capture/save live network trafficv	2.6.0	Wireshark
Burp Suite	Capture live network traffic	1.7.33	PortSwigger Security
Windows Laptop	Network forensics of iOS apps	Windows 10	Windows
NetworkMiner	Analyze network traffic	2.3.1	NETRESEC Erik Hjelmvik

audience. Additionally, the report should be not only factual but also contains defensible details. The report's technical information and results should be explained thoroughly to aid in the decision-making process.

V. Network Forensic Tools

Network forensic tools help in network investigation to gather essential information about an intrusion activity. These tools are used to analyze network traffic to identify the nature and type of activities within the network over a specific duration [45]. The forensic tools are designed so that they are compatible with network hardware devices such as firewalls, thereby making it possible to collect and preserve network traffic.

Additionally, these tools are equipped with the ability to perform a quick analysis of network traffic. Network forensics tools can be categorized based on either host-based or network-wide-based. Additional categories include general-purpose tools, specific tasks tools, or libraries/framework tools [46]. A review of the most frequently used network forensic tools is summarized in Table IV. The following subsections discuss them comprehensively.

A. Wireshark

Wireshark is an open-source graphical user interface application software tool designed to capture, filter, and analyze network traffic. It is easy to use, and thus it is helpful in the analysis of network forensics data. Wireshark has more packet filtering capabilities, decoding protocol features, and packets detail markup language (PDML). In Wireshark, it is possible to view network packets as they are captured in real-time. Wireshark also shows the results of lost pockets due to CPU power [47]. Wireshark can be used as several instruments in one Anwendung. Program. You will use it to evaluate the structure of Network traffic checking for potential security flaws And assaults on health. This can detect other types of Encapsulation, isolation, and show of all fields in the Packet network. You have all those powerful capabilities. Do you think Wireshark's hard to know? For specific instances, Respect it, but you can quickly learn how to use it, the filters with the app, and how to use them Packets unique to the network. Filters in WireShark refer to Berkeley Packet Filters. That is simply a language for microprogramming Compiled against packets and executed at run time Taken off by software like tcpdump and Wireshark. Primarily, filters are used to separate a Quite small parcel set among a large number of Packets focused on search criteria. The filter is compiled to run as best Quality, significant when you are doing a quality Real-time grab. Filtering is for others WireShark's most essential features since it makes Achieving two purposes: selectively collecting the packets From the network, and to locate interested parties Packages [47][48] [49].

B. Tshark

Tshrak is a command-line tool used for data network protocol analysis. It helps to capture traffic from a live data network and read traffic information from saved packet data files. It can also print a decoded form of network packets to a quality output or writes the packet to a pcap file. For instance, tshark can capture data traffic on the network interface "eth1" filtering out all traffic from port 22 and sorting the results in the file "test1.pcap. # tshark I eth1 w test1.pcap" not port 22. Capture on eth1 235. Tshark is a packet capture application that can potent-sensing and explain pcap scrutiny functionality. It captures packet-data from an alive network or inspects packets from an earlier trapped file and decodes those packets' form into the standard output file. The default capture file format built into TShark is pcap. Weka consists of data pre-processing, classification, regression, clustering, correlation and visualization methods that are well-suited to the creation of new schemes [22] [50] [51].

C. Dumpcap

The dumpcap is a network traffic analysis tool, which is designed to capture data packets. It is a Wireshark distribution tool, which comes in command-line. The tool captures traffic from a live network and is equipped to write the output in a pcapng file format. Dumpcap has the added advantage of using fewer system resources, making it possible to boost the capture capabilities. Table V provides a summative analysis of popular tools used for network forensics [47].

D. Network Forensic Analysis Tools (NFATs)

Network Forensic Analysis Tools (aka NFATs) allow network investigators and system administrators to track networks and gather any anomalous or malicious traffic information. Such tools synergize with network infrastructure and network

TABLE V. MOST COMMONLY USED TOOLS TO SUPPORT VARIETY OF NETWORK FORENSIC INVESTIGATIONS

Tools	Open Sourece/ Proprietary software	Plateform	Website	Attributes
TCPDump Win dump [24], [25]	Open Source	Unix/Windows	www.tcpdump.org	F
Ngrep [26], [27]	Open Source	Unix	http ://ngrep.sourceforge.net	F
Wireshark [28] [29] [28]	Open Source	Unix/Windows	www.Wireshark.org	F
Driftnet [28]	Open Source	Unix/Windows	www.backtrack-linux. Org/backtrack-S-releue [Release 3, August 2012]	F F
NetworkMiner [30] [31]	Open Source /Prop	Windows	www.netresec.com/?page=NetworkMiner	F
Airmon-ng. Airodump-ng & Aireplay-ng. [32] [33]	Open Source	Unix	www.backtrack-linux. Org /backtrack-S-releue [Release 3, August 2012]	F L R C F L R C
Kismet [33]	Open Source	Unix/Windows	www.kismetwireless.net	F
NetStumbler [34]	Open Source	Windows	www.netstumbler.com	F
Xplico [35]	Open Source	Unix	http://packetstormsecuity.org/files/tags/forensics	F
DeepNines [35]	Proprietary	Unix	www.deepnines.com	F
Sleuth Kit [36]	Open Source	Unix	www.sleuthkit.org	F R C
Argus [33]	Open Source	Unix	www.qosient.com/argus	F L
Fenris [31]	Open Source	Unix	http://camtuf.coredump.cx/fenris/whatis.shtml	F
Flow-Tools [30]	Open Source	Unix	www.splintered.net/sw/flowtools	F L
EtherApe [31]	Open Source	Unix	http ://etherape.sourceforge.net	F
Honeyd [37] [38]	Open Source	Unix	www.citi.umich.edu/u/provos/honeyd	F
SNORT [24], [25]	Open Source	Unix/Windows	www.snort.org	F
Omnipeek/ /EtherPeek [37]	Proprietary	Windows	www.wildpackets.com	F L R
Savant [31]	Proprietary	Appliance /Windows	www.intrusion.com	F R
Forensic Log Analysis-GUI [31]	Open Source /Prop	Unix	http://sourceforge.net/projects/pyflag	L
Dragon IDS [39] [40]	Proprietary	Unix	www.enterasys.com	F R L C
Infinistream [40]	Proprietary	Appliance	www.netscout.com	F R C
RSA En Vision [31]	Proprietary	Appliance /Windows	www.emc.com/security/rsa-envision.html	F L R C A
NetDetector [41] [42]	Proprietary	Appliance	www.niksun.com	F R C A
NetIntercept [43]	Proprietary	Appliance	www.nikson.com/sandstom.php	F R C A
NetWitness [44]	Proprietary	Windows	www.netwitness.com [www.rsa.com]	F L R C A

appliances, such as firewalls and IDS, to make it possible to maintain long-term network traffic records. NFATs allow for rapid analyzes of patterns detected by network security tools.

VI. SYSTEM TYPES ARE USED TO GATHER DATA / TRAFFIC FROM THE NETWORK

Two types of Network traffic collecting data systems can be “stop, look and listen” or “catch-it-as-you-can.”

“Catch-it-as-you-can”: All packets are sent to the database through a traffic point where they are stored in. The analysis is then conducted on stored data. Data from the analysis is also stored in the database. The data saved can be preserved for future review. Nevertheless, it should be noted that this type of device demands a considerable storage capacity.

The “stop, look and listen” method is different from the “catch-it-as-you-can” approach because only data is stored in the database needed for analysis. The incoming traffic in memory is filtered and processed in real-time, meaning this device needs less storage and a much faster processor.

Since the two systems need ample storage space, it is necessary to weigh and address privacy issues with the “catch-it-as-you-can” system. This program also collects user data; however, ISPs are prohibited from receiving or revealing information without user permission.

VII. CHALLENGES RELATING TO NETWORK EVIDENCE

Network-based evidence faces specific challenges in many fields, including collection, storage, content, privacy, confiscation, and admissibility. Below we’ll cover some of the significant issues Below.

Collection : Within a network environment, clear proof can be hard to locate. Networks include as many bits of data as possible; from wireless devices to web proxies to big log servers; which often makes it difficult to determine the proof’s correct position. Even if you know where a specific piece of evidence exists, it can be difficult for political or technological purposes to access it.

Storage: Commonly, the network of computers can not use permanent or secondary data. As a result, the data they hold can be so fragile they won’t survive a computer reset.

Content: Unlike files, management to contain all file contents and their metadata, network devices with the desired degree of granularity may or may not store information. Network computers also have minimal storage capacity, instead of full data records that have crossed the network, only selected transaction or data transfer metadata are typically retained.

Privacy: Legal problems related to personal privacy occur unique to computer network-based retrieval techniques, depending on the jurisdiction.

Sezure: Seizing a hard disk may disturb a person or an organization. Nonetheless, it is also possible to design and implement a replica of the original, so that critical operations can continue with minimal disruption. Seizing a networked device can be even more damaging. A whole part of the network can be downgraded indefinitely for more extreme situations. Investigators can, however, minimize the impact on computer network operations in such circumstances.

Admissibility: For criminal and civil cases, evidence-based on file systems is now widely acknowledged. So long as the evidence stored on the file system is legitimate collected, adequately interpreted, and relevant to the case, there are clear precedents for the processing and presenting the evidence in court. In comparison, network forensics is a modern approach to automated investigations. There are often contradictory or even non-existent legal precedents for accepting different kinds of facts based on the digital network. With time, digital network-based testimony becomes more prevalent, setting precedents for the case and standardizing them.

VIII. Conclusion

Network forensic investigation is an essential process that helps a cyber-forensics investigator to obtain, analyze, evaluate, categorize, and identify crucial evidence. It ultimately makes it possible to apprehend a cyber-criminal or any person suspected of committing a cyber-criminal offense. Consequently, it is paramount for a network forensic investigator to consider adopting and utilizing an efficient and robust forensic network investigation methodologies that ultimately help improve the investigation process. As intimated in this research, the OSCAR methodology provides a forensic investigator with essential tools and guidelines that determines the approach, methods, and strategies used to obtain, strategize, collect, analyze, and report the findings of a network forensics investigations. It is also paramount for the network forensic investigation process to follow and be executed using essential tools such as Wireshark, tshark, Burpe Suite, and tcpdump that tends to help in simplifying and improving the forensics investigation process. Future work: To developed a tool-kits that parse various network protocols commonly used in various sorts of different networks are required. And, because most data in networks is volatile, it may be necessary to preserve or document it selectively in advance to speed up the forensic process.

References

[1] M. Matsalu *et al.*, “Digitaalse ekspertiisi tööjõu pädevuse arendamine eesti kaitseliidu näitel,” Ph.D. dissertation, 2019.

[2] G. S. Chhabra and P. Singh, “Distributed network forensics framework: A systematic review,” *International Journal of Computer Applications*, vol. 119, no. 19, 2015.

[3] G. A. Pimenta Rodrigues, R. de Oliveira Albuquerque, F. E. Gomes de Deus, G. A. De Oliveira Júnior, L. J. García Villalba, T.-H. Kim *et al.*, “Cybersecurity and network forensics: Analysis of malicious traffic towards a honeynet with deep packet inspection,” *Applied Sciences*, vol. 7, no. 10, p. 1082, 2017.

[4] D. Chang, M. Ghosh, S. K. Sanadhya, M. Singh, and D. R. White, “Fbhash: A new similarity hashing scheme for digital forensics,” *Digital Investigation*, vol. 29, pp. S113–S123, 2019.

[5] L. Liebler, P. Schmitt, H. Baier, and F. Breitinger, “On efficiency of artifact lookup strategies in digital forensics,” *Digital Investigation*, vol. 28, pp. S116–S125, 2019.

[6] K. Karampidis, E. Kavallieratou, and G. Papadourakis, “A review of image steganalysis techniques for digital forensics,” *Journal of information security and applications*, vol. 40, pp. 217–235, 2018.

[7] F. Akhtar, J. Li, M. Azeem, S. Chen, H. Pan, Q. Wang, and J.-J. Yang, “Effective large for gestational age prediction using machine learning techniques with monitoring biochemical indicators,” *The Journal of Supercomputing*, pp. 1–19, 2019.

[8] J. Li, D. Zhou, W. Qiu, Y. Shi, J.-J. Yang, S. Chen, Q. Wang, and H. Pan, “Application of weighted gene co-expression network analysis for data from paired design,” *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[9] F. Akhtar, J. Li, Y. Pei, A. Imran, A. Rajput, M. Azeem, and Q. Wang, “Diagnosis and prediction of large-for-gestational-age fetus using the stacked generalization method,” *Applied Sciences*, vol. 9, no. 20, p. 4317, 2019.

[10] A. Imran, J. Li, Y. Pei, J.-J. Yang, and Q. Wang, “Comparative analysis of vessel segmentation techniques in retinal images,” *IEEE Access*, vol. 7, pp. 114 862–114 887, 2019.

[11] J. Li, L. Liu, J. Sun, H. Mo, J.-J. Yang, S. Chen, H. Liu, Q. Wang, and H. Pan, “Comparison of different machine learning approaches to predict small for gestational age infants,” *IEEE Transactions on Big Data*, 2016.

[12] A. Almulhem, “Network forensics: Notions and challenges,” in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2009, pp. 463–466.

[13] T. Kilpatrick, J. Gonzalez, R. Chandia, M. Papa, and S. Shenoi, “An architecture for scada network forensics,” in *IFIP International Conference on Digital Forensics*. Springer, 2006, pp. 273–285.

[14] K. Wang, M. Du, Y. Sun, A. Vinel, and Y. Zhang, “Attack detection and distributed forensics in machine-to-machine networks,” *IEEE Network*, vol. 30, no. 6, pp. 49–55, 2016.

[15] A. Kurniawan and I. Riadi, “Detection and analysis cerber ransomware based on network forensics behavior,” *International Journal of Network Security*, vol. 20, no. 5, pp. 836–843, 2018.

[16] R. Messier, *Network forensics*. John Wiley & Sons, 2017.

[17] H. Bensefia and N. Ghoualmi, “An intelligent system for decision making in firewall forensics,” in *International Conference on Digital Information and Communication Technology and Its Applications*. Springer, 2011, pp. 470–484.

[18] S. Krasser, G. Conti, J. Grizzard, J. Gribschaw, and H. Owen, “Real-time and forensic network data analysis using animated and coordinated visualization,” in *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*. IEEE, 2005, pp. 42–49.

[19] Q. Al-Mousa and Z. Al-Mousa, “Honeypots aiding network forensics: Challenges and notins,” *Journal of Communication*, vol. 8, no. 11, pp. 700–707, 2013.

[20] J. Llano Tejera, “Herramientas forenses para la respuesta a incidentes informáticos,” Ph.D. dissertation, Universidad Central” Marta Abreu” de Las Villas, 2014.

[21] W. Ren, “Modeling network forensics behavior,” *Journal of Digital Forensic Practice*, vol. 1, no. 1, pp. 57–65, 2006.

[22] S. Davidoff and J. Ham, *Network forensics: tracking hackers through cyberspace*. Prentice hall Upper Saddle River, 2012, vol. 2014.

[23] J. Buric and D. Delija, “Challenges in network forensics,” in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2015, pp. 1382–1386.

[24] P. Arlos and M. Fiedler, “A comparison of measurement accuracy for dag, tcpdump and windump,” *available online at Blekinge Institute of Technology (Sweden)¡ www. its. bth. se/staft/pca*, 2007.

[25] P. Goyal and A. Goyal, “Comparative study of two most popular packet sniffing tools-tcpdump and wireshark,” in *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2017, pp. 77–81.

[26] D. Dittrich, “Dissecting distributed malware networks,” *Availabel from:¡ http://security. isu. edu/ppt/pdfppt/Core02. pdf*, 2002.

[27] J. R. Binkley and S. Singh, “An algorithm for anomaly-based botnet detection.” *SRUTI*, vol. 6, pp. 7–7, 2006.

[28] U. Banerjee, A. Vashishtha, and M. Saxena, “Evaluation of the capabilities of wireshark as a tool for intrusion detection,” *International Journal of computer applications*, vol. 6, no. 7, pp. 1–5, 2010.

[29] L. Chappell, "Wireshark 101: Essential skills for network analysis-wireshark solution series," *Laura Chappell University, USA*, 2017.

[30] R. Chowdhary, S. L. Tan, J. Zhang, S. Karnik, V. B. Bajic, and J. S. Liu, "Context-specific protein network miner–an online system for exploring context-specific protein interaction networks from the literature," *PLoS One*, vol. 7, no. 4, p. e34480, 2012.

[31] R. Umar, I. Riadi, and B. F. Muthohirin, "Live forensics of tools on android devices for email forensics," *Telkomnika*, vol. 17, no. 4, pp. 1803–1809, 2019.

[32] P. Čisar and S. M. Čisar, "Ethical hacking of wireless networks in kali linux environment," *Annals of the Faculty of Engineering Hunedoara*, vol. 16, no. 3, pp. 181–186, 2018.

[33] O. Barybin, E. Zaitseva, and V. Brazhnyi, "Testing the security esp32 internet of things devices," in *2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T)*. IEEE, 2019, pp. 143–146.

[34] S. Ekhator, "Evaluating kismet and netstumbler as network security tools & solutions." 2010.

[35] J.-N. Hilgert, M. Lambertz, and D. Plohmann, "Extending the sleuth kit and its underlying model for pooled storage file system forensic analysis," *Digital Investigation*, vol. 22, pp. S76–S85, 2017.

[36] J.-N. Hilgert, M. Lambertz, and S. Yang, "Forensic analysis of multiple device btrfs configurations using the sleuth kit," *Digital Investigation*, vol. 26, pp. S21–S29, 2018.

[37] N. Provos, "Honeyd-a virtual honeypot daemon," in *10th DFN-CERT Workshop, Hamburg, Germany*, vol. 2, 2003, p. 4.

[38] R. Chandran, S. Pakala *et al.*, "Simulating networks with honeyd," *online], Technical paper, Paladion Networks, December*, 2003.

[39] P. Kazienko and P. Dorosz, "Intrusion detection systems (ids) part 2-classification; methods; techniques," *WindowsSecurity. com*, 2004.

[40] J. Kipp *et al.*, "Using snort as an ids and network monitor in linux," *GIAC*, pp. 1–4, 2001.

[41] P. Lin, K. Ye, and C.-Z. Xu, "Netdetector: an anomaly detection platform for networked systems," in *2019 IEEE International Conference on Real-time Computing and Robotics (RCAR)*. IEEE, 2019, pp. 69–74.

[42] Y. R. Wang and A. Kanemura, "Designing lightweight feature descriptor networks with depthwise separable convolution," in *????????????? ? 34 ????? (2020)*. ?????? ??????, 2020, pp. 2K1ES204–2K1ES204.

[43] R. Joshi and E. S. Pilli, "Network forensic tools," in *Fundamentals of Network Forensics*. Springer, 2016, pp. 71–93.

[44] T. A. Moore, M. E. Longworth, B. Girardi, and D. Love, "Apparatus and method for network analysis," Dec. 15 2009, uS Patent 7,634,557.

[45] M. H. Mate and S. R. Kapse, "Network forensic tool–concept and architecture," in *2015 Fifth International Conference on Communication Systems and Network Technologies*. IEEE, 2015, pp. 711–713.

[46] A. Lazzez, "A survey about network forensics tools," *Int. J. Comput. Inf. Technol*, vol. 2, no. 1, 2013.

[47] R. Hunt and S. Zeadally, "Network forensics: an analysis of techniques, tools, and trends," *Computer*, vol. 45, no. 12, pp. 36–43, 2012.

[48] S. Wang, D. Xu, and S. Yan, "Analysis and application of wireshark in tcp/ip protocol teaching," in *2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT)*, vol. 2. IEEE, 2010, pp. 269–272.

[49] V. Ndatinya, Z. Xiao, V. R. Manepalli, K. Meng, and Y. Xiao, "Network forensics analysis using wireshark," *International Journal of Security and Networks*, vol. 10, no. 2, pp. 91–106, 2015.

[50] Y. Lee and Y. Lee, "Toward scalable internet traffic measurement and analysis with hadoop," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 1, pp. 5–13, 2012.

[51] R. Menon and O. G. MENON, "Mining of textual databases within the product development process," Ph.D. dissertation, 2004.

Cloud Computing in Remote Sensing: Big Data Remote Sensing Knowledge Discovery and Information Analysis

Yassine SABRI[1], Fadoua Bahja[2]
Laboratory of Innovation in Management
and Engineering for Enterprise (LIMIE),
ISGA Rabat, 27 Avenue Oqba,
Agdal, Rabat, Morocco

Aouad Siham[3]
Mohammed V University of Rabat
Smart Systems Laboratory (SSL)
ENSIAS, Morocco

Aberrahim Maizate[4]
RITM- ESTC/CED -ENSEM,
University Hassan II
Km7, El jadida Street, B.P.
8012, Oasis, Casablanca 8118

***Abstract*—With the rapid development of remote sensing technology, our ability to obtain remote sensing data has been improved to an unprecedented level. We have entered an era of big data. Remote sensing data clear showing the characteristics of Big Data such as hyper spectral, high spatial resolution, and high time resolution, thus, resulting in a significant increase in the volume, variety, velocity and veracity of data. This paper proposes a feature supporting, salable, and efficient data cube for time-series analysis application, and used the spatial feature data and remote sensing data for comparative study of the water cover and vegetation change.The spatial-feature remote sensing data cube (SRSDC) is described in this paper. It is a data cube whose goal is to provide a spatial-feature-supported, efficient, and scalable multidimensional data analysis system to handle large-scale RS data. It provides a high-level architectural overview of the SRSDC.The SRSDC offers spatial feature repositories for storing and managing vector feature data, as well as feature translation for converting spatial feature information to query operations.The paper describes the design and implementation of a feature data cube and distributed execution engine in the SRSDC. It uses the long time-series remote sensing production process and analysis as examples to evaluate the performance of a feature data cube and distributed execution engine. Big data has become a strategic highland in the knowledge economy as a new strategic resource for humans. The core knowledge discovery methods include supervised learning methods data analysis supervised learning, unsupervised learning methods data analysis unsupervised learning, and their combinations and variants.**

***Keywords*—Remote sensing; data integration; cloud computing; big data**

I. Introduction

In recent decades, the remarkable developments in Earth observing(EO) technology provided a significant amount of remote sensing(RS) data openly available [1]. This large observation dataset characterized the information about the earth surface in space, time, and spectral dimensions [2][3]. Apart form these dimensions, these data also contain many geographic features, such as forests, cities, lakes and so on, and these features could help researchers to locate their interested study areas rapidly. Now these multidimensional RS data with features have been widely used for global change detection research such as monitoring deforestation [4] and detecting temporal changes in floods [35]. However, the conventional geographic information system (GIS) tools are inconvenient for scientists to process the multidimensional RS data, because they lack appropriate methods to express multidimensional data models for analysis. And researchers have to do additional data organization work to conduct change detection analysis. For a more convenient analysis, they need a multidimensional data model which could support seamless analysis in space, time, spectral and feature [5].

Recently, many researchers have proposed using a multidimensional array model to organize the RS raster data [6][7]. Subsequently, they achieved the spatio-temporal aggregations capacity used in spatial on-line analytical processing (SOLAP) systems [8][9], as a data cube. Using this model, researchers can conveniently extract the desired data from the large dataset for analysis, and it reduces the burdens of data preparation for researchers in time-series analysis. However, in addition to extracting data with simple three-dimensional (3D) space-time coordinates, researchers occasionally need to extract data with some geographic features [10][11][12], which are often used to locate or mark the target regions of interest. For example, flood monitoring often needs to process multidimensional RS data which have the characteristics of large covered range, long time series and multi bands [13]. If we built all the analysis data as a whole data cube which has the lakes or river features, we could rapidly find the target study area we need by feature and analyse the multi bands image data to detect the flood situation with the time series. That makes researchers focus on their analysis work without being troubled by the data organization. This study aims to develop the spatial-feature remote sensing data cube(SRSDC), a data cube whose goal is to deliver a spatial feature-supported, efficient, and scalable multidimensional data analysis system to handle the large-scale RS data. The SRSDC provides spatial feature repositories to store and manage the vector feature data, and a feature translation to transform the spatial feature information to a query operation. To support large-scale data analysis, the SRSDC provides a work-scheduler architecture to process the sliced multidimensional arrays with dask [14].

The remainder of this paper is organized as follows. Section 2 describes some preliminaries and related works. Section 3 presents an architectural overview of the SRSDC. Section 4 presents the design and implementation of a feature data cube and distributed execution engine in the SRSDC. Section 5 uses the long time-series remote sensing production process and analysis as examples to evaluate the performance of a feature data cube and distributed execution engine. Section 6 concludes this paper and describes the future work prospects.

II. Preliminaries and Related Work

A. Knowledge Discovery Categories

In the following, we discuss four broad categories of applications in geosciences where knowledge discovery methods have been widely used and have aroused impressive attention. For each application, a brief description of the problem from a knowledge discovery perspective is presented.

Detecting objects and events in geoscience data is important for a number of reasons. For example, detecting spatial and temporal patterns in climate data can help in tracking the formation and movement of objects such as cyclones, weather fronts, and atmosphere rivers, which are responsible for the transfer of precipitation, energy, and nutrients in the atmosphere and ocean. Many novel pattern mining approaches have been developed to analyze the spatial and temporal properties of objects and events, e.g., spatial coherence and temporal persistence that can work with amorphous boundaries. One such approach has been successfully used for finding spatio-temporal patterns in sea surface height data, resulting in the creation of a global catalogue of Mesoscale Ocean eddies. The use of topic models has been explored for finding extreme events from climate time series data. Given the growing success of deep learning methods in mainstream machine learning applications, it is promising to develop and apply deep learning methods for a number of problems encountered in geosciences. Recently, deep learning methods, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to detect geoscience objects and events, such as detecting extreme weather events from a climate model.

Knowledge discovery methods can contribute to estimating physical variables that are difficult to monitor directly, e.g., methane concentrations in air or groundwater seepage in soil, using information about other observed or simulated variables. To address the combined effects of heterogeneity and small sample size, multi-task learning frameworks have been explored, where the learning of a model at every homogeneous partition of the data is considered as a separate task, and the models are shared across similar tasks.

The sharing of learning is able to help in regularizing the models across all tasks and avoid the problem of over fitting. Focusing on the heterogeneity of climate data, online learning algorithms have been developed to combine the ensemble outputs of expert predictors and conduct robust estimates of climate variables such as temperature. To address the paucity of labeled data, novel learning frameworks such as semi-supervised learning, active learning, have huge potential to improving the state-of-the-art in estimation problems encountered in geoscience applications. Forecasting long-term trends of geoscience variables such as temperature and greenhouse gas concentrations ahead of time can help in modeling future scenarios and devising early resource planning and adaptation policies. Some of the existing approaches in knowledge discovery for time-series forecasting include exponential smoothing techniques, the auto regressive integrated moving average model and probabilistic models, such as hidden Markov models and Kalman filters. In addition, RNN-based frameworks such as long-short-term-memory (LSTM) have been used for long-term forecasting geoscience variables.

An important problem in geoscience application is to understand the relationships in geoscience data, such as periodic changes in the sea surface temperature over the eastern Pacific Ocean and their impact on several terrestrial events such as floods, droughts and forest fires. One of the first knowledge discovery methods in discovering relationships from climate data is a seminal work, where graph-based representations of global climate data were constructed. In the work, each node represents a location on the Earth and an edge represents the similarity between the eliminated time series observed at a pair of locations. The high-order relationships could been discovered from the climate graphs. Another kind of method for mining relationships in climate science is based on representing climate graphs as complex networks, including approaches for examining the structure of the climate system, studying hurricane activity. Recently, some works have developed novel approaches to directly discover the relationships as well as integrating objects in geoscience data. For example, one work has been implemented to discover previously unknown climate phenomena. For causality analysis, the most common tool in the geosciences is bivariate Grange analysis, followed by multi-variate Granger analysis using vector auto regression (VAR) models.

B. Knowledge Discovery Methods

As a new strategic resource for human beings, big data has become a strategic highland in the era of knowledge economy. It is a typical representative of the data-intensive scientific paradigm following experience, theory and computational models, since this new paradigm mainly depends on data correlation to discover knowledge, rather than traditional causality. It is bringing about changes in scientific methodology, and will become a new engine for scientific discovery.

Knowledge discovery of remote sensing big data lies at the intersection of earth science, computer science, and statistics, and is a very important part of artificial intelligence and data science. Its aims at dealing with the problem of finding a predictive function or valuable data structure entirely based on data and will not be bothered by the various data types and, is suitable for comprehensively analyzing the Earth's big data.

The core knowledge discovery methods include supervised learning methods, unsupervised learning methods, and their combinations and variants. The most widely used supervised learning methods use the training data taking the form of a collection of (x, y) pairs, and aims to produce a prediction y' in response to a new input x' by a learned mapping f(x), which produces an output y for each input x (or a probability distribution over y given x). There are different supervised learning methods based on different mapping functions, such as decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers. In recent years, deep networks have received extensive attention in supervised learning. Deep networks are composed of multiple processing layers to learn representations of data with multiple levels of abstraction, and discover intricate structures of the big earth data by learning its internal parameters to compute the representation in each layer. Deep networks have brought about breakthroughs in processing satellite image data, forecasting long-term trends of geoscience.

Unlike supervised learning methods, unsupervised learning involves the analysis of unlabeled data under assumptions about structural properties of the data. For example, the dimension reduction methods make some specific assumptions that the earth data lie on a low-dimensional manifold and aim to identify that manifold explicitly from data, such as principal components analysis, manifold learning, and auto encoders. Clustering is another very typical unsupervised learning algorithm, which aims to find a partition of the observed data, and mine the inherent aggregation and regularity of data. In recent years, much current research involves blends across supervised learning methods and unsupervised learning. Semi supervised learning is a very typical one, which makes use of unlabeled data to augment labeled data in a supervised learning context considering the difficulty of obtaining some geoscience supervision data. Overall, knowledge discovery of the big earth data needs to leverage the development of artificial intelligence, machine learning, statistical theory, and data science.

C. Related Work

With the growing numbers of archived RS images for Earth observation, an increasing number of scientists are interested in the spatiotemporal analysis of RS data. Many researchers proposed combining online analytical processing (OLAP) [15] technology with the GIS [16] to build a data cube. They built the multidimensional database paradigm to manage several dimension tables, periodically extracting the dimension information from the data in GIS, and achieved the ability to explore spatiotemporal data using the OLAP spacetime dimension aggregation operation. Sonia Rivest et al. deployed a spatial data warehouse based on GIS and spatial database to acquire, visualize, and analyze the multidimensional RS data [17]. Matthew Scotch et al. developed the SOVAT tool [18], using OLAP and GIS to perform the public health theme analysis with the data composed of spatiotemporal dimensions. These tools can facilitate researchers extracting data with spatiotemporal dimensions; however, their multidimensional data model is unsuitable for complicated scientific computing. Further, they did not adopt an appropriate architecture for large data processing [19][20]. Therefore, their ability to handle large-scale data is limited.

Owing to natural raster data structure of Earth observation images, the time-series imagery set can be easily transformed to multidimensional array. For example, a 3D array can represent the data with spatiotemporal dimensions. This data type is suitable for parallel processes, because a large array can be easily partitioned into several chunks for distributed storing and processing. In addition, the multidimensional array model enables a spatiotemporal auto-correlated data analysis; therefore, researchers need not be concerned about the organization of discrete image files. Thus, much research is focused on developing new analysis tools to process the large RS data based on the multidimensional array model; e.g., Gamara et al.[21] tested the performance of spatiotemporal analysis algorithms on array database architectures - SCIDB [22], which described the efficiency of spatiotemporal analysis based on the multidimensional array model, Assis et al.[23] built a parallel RS data analysis system based on the MapReduce framework of Hadoop [24], describing the 3D array with key/value pairs. Although these tools have significantly improved the computation performance of RS data analysis, they also contain some deficiencies. First, many of them focused only on analyzing the RS raster image data located by geographic coordinates, and did not provide the support of spatial feature, thereby limiting their ability to use these geographic objects in the analysis application. Next, some of these tools require analysers to fit their algorithms into specialised environments, such as Hadoop MapReduce framework [25]. This will be user unfriendly to researchers who only desire to focus on their analysis application.

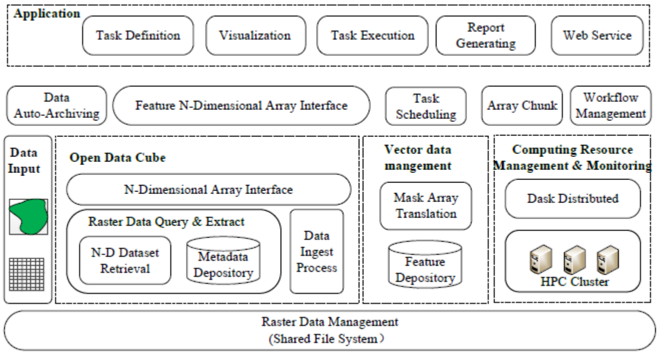


Fig. 1. The Architecture of the SRSDC.

III. Architecture Overview

A. Target Data and Environment

The SRSDC system is designed for providing the services of large RS data time-series analysis with spatial features, and it aims to manage and process the large-scale spatial feature data and satellite images seamlessly. Based on the open data cube (ODC) [26], which is a popular data cube system used for spatial raster data management and analysis, we archived large amounts of satellite data within China. These data came from different satellites including Landsat, MODIS, GaoFen(GF) and HuanJing(HJ). In addition, the SRSDC also contains many features data within China, such as lakes, forests and cities. These spatial vector data were downloaded from the official web site of OpenStreetMap [27]. Before obtaining these satellite images in the SRSDC, the geometric correction and radiometric correction for these images must be ensured. This can ensure the comparability between the images in different time, space and measurements; subsequently, the global subdivision grid can be used to partition the data into many tiles(grid files). These tiles were stored as the NetCDF format [28], which supports many analysis libraries and scientific toolkits.

B. FRSDC Architecture Overview

The SRSDC system adopted the relational database and file system to manage the spatial data. It is designed to be scalable and efficient and provide feature support for time-series analysis. Compared with the ODC system [29], which only supports spatial raster data management and analysis as a data cube, the SRSDC supports the extraction of target satellite data as a multidimensional array with the geographic object. It could perform the spatial query operation with geographic objects, instead of locating data with only geographic coordinates. Therefore, the dataset built for analysis has geographic meaning. Thus, if researchers desire to obtain the target dataset, they only need to query data with the geographic meaning of the analysis themes, without knowing specific geographic coordinates. As shown in Fig. 1, the system is primarily composed of the data management and distributed execution engine (DEE). Data management consists of two parts, raster data management and vector data management. For the raster data, the SRSDC will archive it into a shared network file system and extract its metadata information automatically; these metadata will be stored into the metadata depository managed by ODC. For the vector data, the SRSDC stores them as geographic objects in the spatial database. After the data management, an N-Dimensional array interface is responsible for transforming the raster data and vector data to an N-Dimensional array that has the spatial feature information. Xarray [28] is used for array handing and computing. DEE is responsible for providing the computing environment and resources on high performance computing(HPC) clusters. The SRSDC use dask [24], which is a parallel computation library with blocked algorithms, for the task scheduling, distributed computing, and resource monitoring. It could help researchers to execute the analysis tasks in parallel.

IV. Design and Implementation

A. Feature Data Cube

1) Spatial feature object in FRSDC: Spatial feature is a geographic object that has special geographic meaning. It is often important for RS application, because researchers occasionally need to process the RS image dataset with geographic objects, such as the classification of an RS image with spatial features [30][31][32]. However, many RS data cube systems only provide the multidimensional dataset without features. Hence, researchers are required to perform additional work to prepare the data for analysis. For example, the ODC system [24] and the data cube based on SCIDB [22] could only query and locate the study area by simple geographic coordinates, so researchers must transform their interested spatial features to coordinate ranges one by one if they want to prepare the analysis ready data. To solve this problem, the SRSDC combined the basic N-Dimensional array with geographic objects to provide the feature N-Dimensional array for researchers, and researchers could easily organise the analysis ready N-Dimensional dataset by their interested features. Within the SRSDC, now we primarily archive the forest and lake features of China, and store them as geographic objects in a PostGIS database [33][34]. The unified modeling language(UML) class diagram in the SRSDC is shown in Fig. 2, and the description of these classes is as follows:

1) The feature class is provided for users to define their spatial feature of interest with a geographic object. It contains the feature type and the geographic object.
2) The feature type class represents the type of geographic object, such as lakes, forests, cities and so on. It contains the description of the feature type and analysis algorithm names that are suitable for the feature type.
3) The geographic object class describes the concrete vector data with geographic meaning, such as Poyang lake (a lake in China). It contains the vector data type to illustrate its geometry type.
4) The raster data type class is used to describe the type of satellite data. It contains the satellite platform, product type, and bands information.
5) The feature operation class is used to extract the feature data-cube dataset from the SRSDC. It contains the feature object, raster data type, and time horizon to build the target feature N-Dimensional array. It also provides some operation functions for users.

2) Data management: As mentioned above, data management consists of raster data management and vector data management. Because of satellite data's large volume and variety, the SRSDC uses the file system to store the raster data and manage the metadata by a relation database that contains NoSQL fields.

In the metadata depository, the SRSDC uses NoSQL fields to describe the metadata information instead of a full relation model. This is because the number of satellite sensors is increasing rapidly, and if the full relation model is used, the database schema must be expanded frequently to meet the new data sources. In contrast, NoSQL fields are more flexible in describing the metadata of satellite data that originate from different data sources. The NoSQL fields contain the time, coordinate, band, data URL, data type, and projection. Among these fields, some are used for data query, such as the time, band, coordinates, and data type. Some other fields are used for loading the data and building the multidimensional array in memory, such as data URL and projection. In addition, comparing the vector data volume (GB level) we download from the OpenStreetMap [26] with the raster data volume (TB level) we archived from several satellite data centers [32], we found that most spatial features that are represented by the vector data are not as large as the raster images; therefore we established a feature depository instead of file system to store and manage them as geographic objects. These objects may contain different geographic meanings, and we defined them as feature types.

The runtime implementation of feature data cube building and processing is shown in Fig. 3. First, the SRSDC receives the user's data request from the web portal and obtains the target geographic object by querying the feature depository. Next, it conducts a feature translation to transform the geographic object into a mask array and obtains the minimum bounding rectangle(MBR) of the feature. Subsequently, with the vertex coordinates of MBR and time horizon, the SRSDC searches for the required raster data's metadata to locate physical URLs of the raster data. Next, ODC's N-Dimensional array interface will load the raster data set from the file system and build a multidimensional array in memory. Subsequently, the mask array will be applied to masking the multidimensional array, and a new multidimensional array with features for analyzing and processing will be obtained. Finally, the SRSDC will

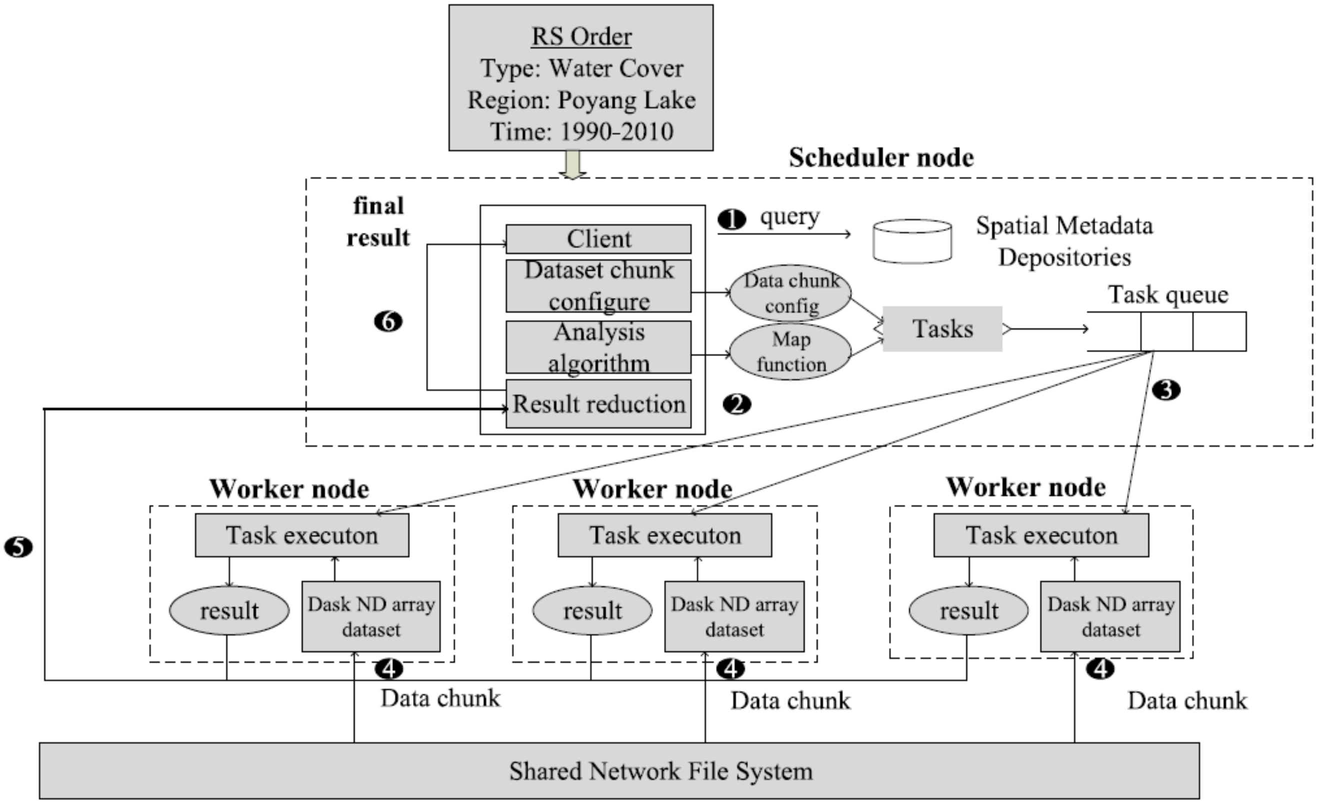


Fig. 2. Large-scale RS Analysis Processing with Distributed Executed Engine.

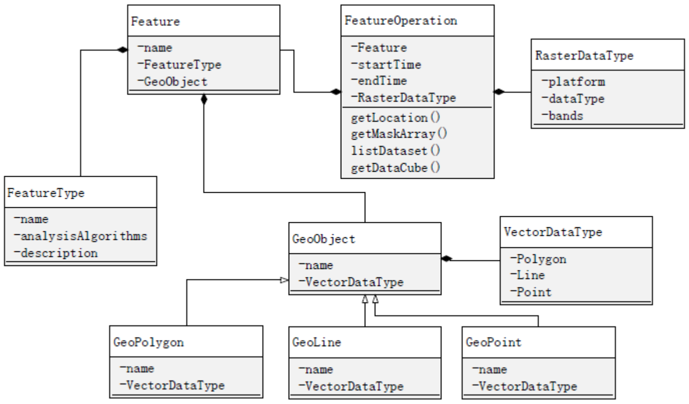


Fig. 3. The UML Class Diagram in the SRSDC.

process the data with the relevant algorithm and return the analysis results to the user.

B. Distributed Executed Engine

As an increasing number of RS applications need to process or analyze the massive volume of RS data collections, the stand-alone mode processing can not satisfy the computation requirement. To process the large-scale RS data efficiently, we built a distributed executed engine using the dask a distributed computing framework focusing on scientific data analysis. Compared with the popular distributed computing tools such as Apache Spark, dask supports the multidimensional data model natively and has a similar API with pandas and numpy. Therefore, it is more suitable for computing an N-Dimensional array. Similar to Spark, dask is also a master-slave system framework that consists of one schedule node and several work nodes. The schedule node is responsible for scheduling the tasks, while the work nodes are responsible for executing tasks. If all the tasks have being performed, these workers' computation results would be reduced to the scheduler and the final result would be obtained.

In the SRSDC, we could index the satellite image scenes by adding their metadata information to the database, and then obtain the data cube dataset (N-Dimensional arrays) from the memory for computing. However, to compute the large global dataset, we should slice the large array into the fixed-size sub-arrays called chunks for computing in the distributed environment. The SRSDC partitions these native images into seamless and massive tiles based on a latitude/longitude grid. The tile size is determined by the resolution of satellite images. For example, in the SRSDC, the Landsat data (each pixel 0.00025°) was partitioned into tiles of size 1°x 1°, and the tiles (4000x4000 pixels array) can be easily organized as a data chunk, which is suitable for the memory in the worker node. By configuring the grid number and time horizon, the chunk could be built. Further, with these data chunks, the SRSDC can transform the big dataset (N-Dimensional arrays) to several sub-arrays loaded by different worker nodes. After all the data chunks have been organized, the scheduler will assign the chunks to the workers and map the functions for

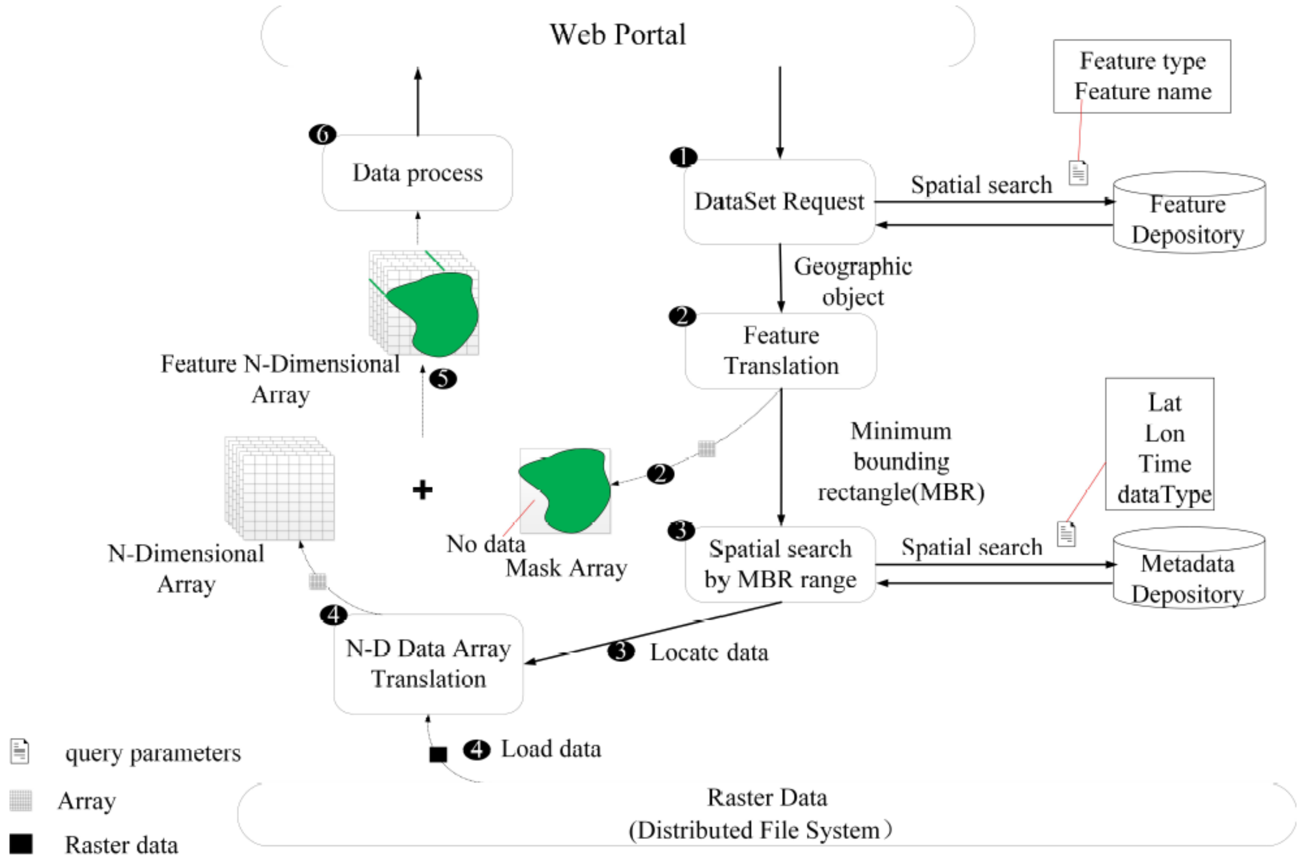


Fig. 4. Runtime Implementation of Feature Data Cube Building.

computing.

As shown in Fig. 4, the processing of large-scale time series analysis by the distributed executed engine is as follows:

1) Organize the data cube dataset by multidimensional spatial query.
2) Configure the appropriate parameters (grid number or time horizon) to organized the data chunks for workers, manage the chunks' ids with a queue.
3) Select the analysis algorithm and data chunks to compose the tasks, and assign these tasks to the worker node.
4) Check the executing state of each task in the workers; if failure occurs, recalculate the result.
5) Reduce all the results to the scheduler and return the analysis result to the client.

V. Experiments

To verify the ability of multidimensional data management and large-scale data analysis in the SRSDC, we conducted the following time-series analysis experiments focusing on spatial feature regions and compared the performance of GEE and stand-alone mode processes on the target dataset.

In this experiment, two RS application algorithms for time-series change detection have been used: NDVI for vegetation change detection and water observation from space (WOfs) for the water change detection. We built the distributed executed engine with four nodes connected by a 20 GB Infiniband network; one node for the scheduler and tree nodes for the workers. Each node was configured with Inter(R) Xeon(R)E5-2460 CPU(2.0GHz) and 32GB memory. The operating system is CentOS 6.5, and the python version is 3.6. To test the performance of the feature data cube, we selected two study regions with the special features as examples. One region for the NDVI is Mulan hunting ground, Hebei Province, China(40.7°-43.1°N, 115.8°-119.1°E), and another region for WOfs is Poyang Lake, Jiangxi Province, China. We built two feature data cube datasets for 20 years(1990-2009) with Landsat L2 data and the geographic objects. The data volume for Mulan hunting ground is approximately 138 GB and the data volume for Poyang Lake is about 96 GB. Figure 5 shows the percentage of observations detected as water for Poyang Lake over the 20-year time series. The red area represents the frequent or permanent water, and the purple area represents the infrequent water. From the result, the shape area of Poyang Lake can be observed clearly. Fig. 6 shows the annual average NDVI production on the Mulan hunting ground; Fig. 7 shows the NDVI time series result of the sampling site(41.5620°N, 117.4520°E) over 20 years. As shown, the values during 2007-2008 were abnormally below the average. This is because the average annual rainfall during this time is lower than that in normal years.

To test the processing performance of the DEE for different amounts of data, we tested time consumed by processing 6.3 GB, 12.8 GB, 49.6 GB, 109.8 GB, 138.6GB input data for NDVI production. These data have been partitioned into 4000x4000 pixels tiles mentioned above, with which we compared the performances of the stand-alone model and DEE models:

1) stand-alone model: organize the dataset as data chunks, and process these data chunks serially with a single server.
2) DEE model: organize the large dataset as data chunks, and assign different workers to read these data chunks to process them in parallel with the distributed executed engine, which consists of one schedule node and three work nodes. As shown from the experimental results in Fig. 8, the DEE mode is much faster than the stand-alone mode because it can use the shared memory of clusters nodes and process the large dataset in parallel. As the process data amount increases, we also observed that the time consumed will grow nonlinearly.

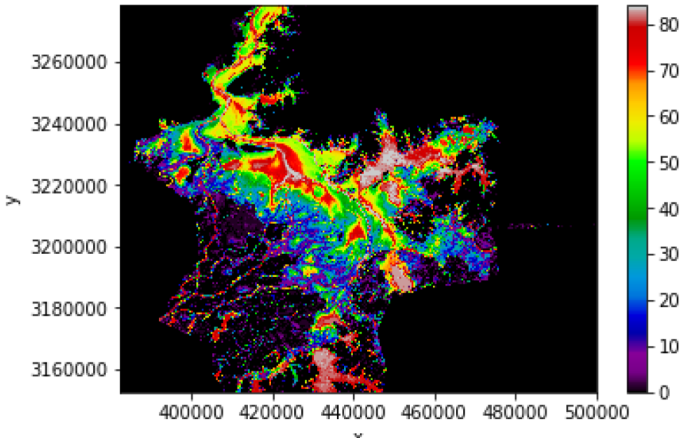


Fig. 5. Water Area of Poyang Lake over 20-year Time Series.

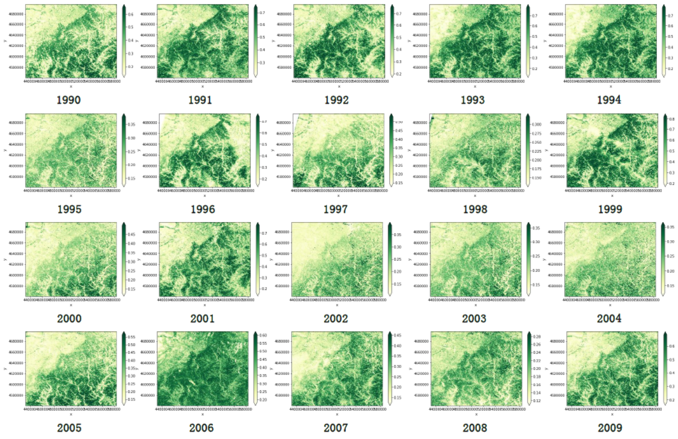


Fig. 6. The Annual Average NDVI of Mulan Hunting Ground for 20 Years.

This is due to the IO limit of the shared network file system and scheduling overhead. The speedup performance when generating the NDVI production with increasing numbers of work nodes also proved this point, as shown in Fig. 9. Therefore, we conclude that the SRSDC has a certain capacity to process the massive data, which is unsuitable for the memory.

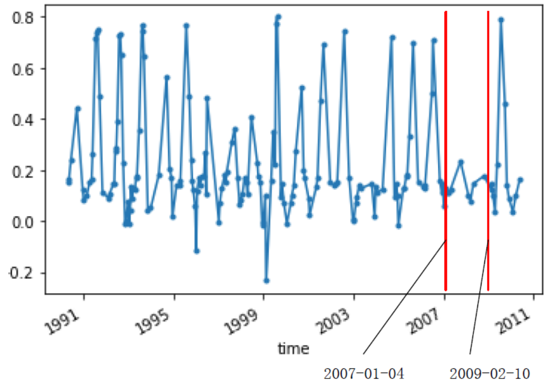


Fig. 7. A NDVI Time Series of Sampling Site on Mulan Hunting Ground.

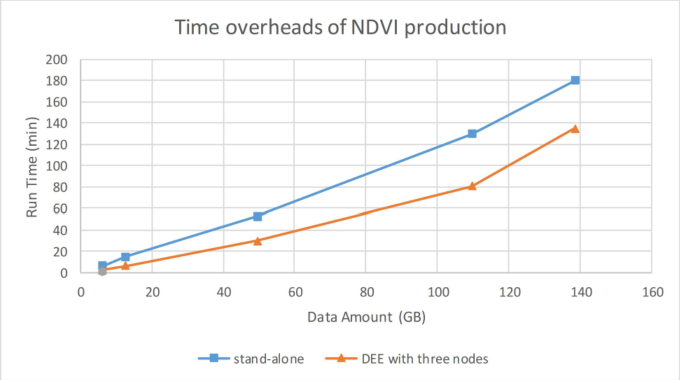


Fig. 8. Runtime of NDVI with the Increase of Data Volume.

VI. Conclusions

We have designed and tested a feature supporting, scalable, and efficient data cube for time-series analysis application, and used the spatial feature data and remote sensing data for comparative study of the water cover and vegetation change. In this system, the feature data cube building and distributed executor engine are critical in supporting large spatiotemporal RS data analysis with spatial features. The feature translation ensures that the geographic object can be combined with satellite data to build a feature data cube for analysis. Constructing a distributed executed engine based on dask ensures the efficient analysis of large-scale RS data. This work could provide a convenient and efficient multidimensional data services for many remote sensing applications [33][34]. However, it also has some limitations; for example, the image data is stored in the shared file system, and its IO performance is limited by the network.

In the future, more work will be performed to optimize the system architecture of the SRSDC, such as improving the performance of the distributed executed engine, selecting other storage methods which could ensure the process data locality, adding more remote sensing application algorithms, etc.

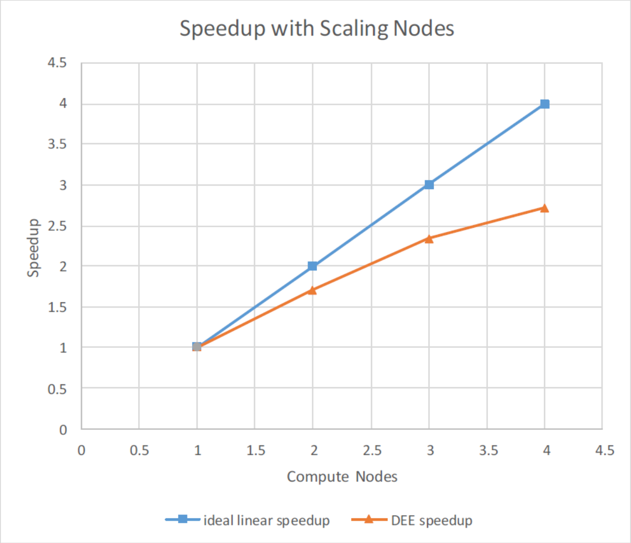


Fig. 9. Speedup for the Generation of NDVI Products with Increasing Nodes.

References

[1] Huadong Guo, Lizhe Wang, Fang Chen, et al. Scientific big data and digital earth. Chinese Science Bulletin, 59(35):50665073, 2014.

[2] Weijing Song, LizheWang, Peng Liu, et al. Improved t-sne based manifold dimensional reduction for remote sensing data processing. Multimedia Tools and Applications, Feb 2018.

[3] Weijing Song, Lizhe Wang, Yang Xiang, et al. Geographic spatiotemporal big data correlation analysis via the HilbertHuang transformation. Journal of Computer and System Sciences, 89:130 (141), 2017.

[4] Robert E. Kennedy, Zhiqiang Yang, and Warren B. Cohen. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. landtrendr — temporal segmentation algorithms. Remote Sensing of Environment, 114(12):2897 2910, 2010.

[5] Toshihiro Sakamoto, Nhan Van Nguyen, Akihiko Kotera, et al. Detecting temporal changes in the extent of annual flooding within the Cambodia and the Vietnamese Mekong Delta from modis time-series imagery. Remote Sensing of Environment, 109(3):295 313, 2007.

[6] Zhang L.F., Chen H., Sun X.J., et al. Designing spatial-temporal-spectral integrated storage structure of multi-dimensional remote sensing images. Journal of Remote Sensing, 21(1):62 73, 2017.

[7] Assis, Luiz Fernando, Gilberto Ribeiro, et al. Big data streaming for remote sensing time series analytics using map-reduce. In XVII Brazilian Symposium on GeoInformatics, 2016.

[8] D.B. Gonzalez and L.P. Gonzalez. Spatial data warehouses and solap using open-source tools. In 2013 XXXIX Latin American Computing Conference (CLEI), pages 112, Oct 2013.

[9] T.O. Ahmed. Spatial on-line analytical processing (solap): Overview and current trends. In 2008 International Conference on Advanced Computer Theory and Engineering, pages 10951099, Dec 2008.

[10] Lizhe Wang, Weijing Song, and Peng Liu. Link the remote sensing big data to the image features via wavelet transformation. Cluster Computing, 19(2):793810, Jun 2016.

[11] LizheWang, Jiabin Zhang, Peng Liu, et al. Spectral spatial multi-featurebased deep learning for hyperspectral remote sensing image classification. Soft Computing, 21(1):213221, Jan 2017.

[12] Weitao Chen, Xianju Li, Haixia He, et al. Assessing different feature sets' effects on land cover classification in complex surface-mined landscapes by Ziyuan-3 satellite imagery. Remote Sensing, 10(1), 2018.

[13] K.O. Asante, R.D. Macuacua, G.A. Artan, et al. Developing a flood monitoring system from remotely sensed data for the Limpopo basin. IEEE Transactions on Geo science and Remote Sensing, 45(6):17091714, June 2007.

[14] Matthew Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. In Kathryn Hu and James Bergstra, editors,Proceedings of the 14th Python in Science Conference, pages 130 136, 2015.

[15] Erik Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley Sons, Inc., 2002.

[16] Z. Yijiang. The conceptual design on spatial data cube. In 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), pages 645648, April 2012.

[17] Sonia Rivest, Yvan Bedard, Marie-Josee Proulx, Martin Nadeau, Frederic Hubert, and Julien Pastor. Solap technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. ISPRS Journal of Photogrammetry and Remote Sensing, 60(1):17 33, 2005.

[18] Matthew Scotch and Bambang Parmanto. Sovat: Spatial olap visualization and analysis tool. In Proceedings of the Hawaii International Conference on System Sciences, page 142.2, 2005.

[19] Junqing Fan, Jining Yan, Yan Ma, et al. Big data integration in remote sensing across a distributed metadata-based spatial infrastructure. Remote Sensing, 10(1), 2018.

[20] Jining Yan, Yan Ma, Lizhe Wang, et al. A cloud-based remote sensing data production system. Future Generation Computer Systems, 2017.

[21] Gilberto Camara, Luiz Fernando Assis, et al. Big earth observation data analytics: Matching requirements to system architectures. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, BigSpatial '16, pages 16, New York, NY, USA, 2016. ACM.

[22] SCIDB. A database management system designed for multidimensional data. http://scidb.sourceforge.net/project.html, 2017.

[23] Apache. Hadoop web site. http://hadoop.apache.org/, 2017.

[24] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, et al. The hadoop distributed file system. In IEEE Symposium on MASS Storage Systems and Technologies, pages 110, 2010.

[25] odc. Open data cube. http://datacube-core.readthedocs.io/en/latest/index.html, 2017.

[26] OpenStreetMap. the project that creates and distributes free geographic data for the world. http://www.openstreetmap.org, 2017.

[27] UCAR. Netcdf file format and api. http://www.unidata.ucar.edu/software/netcdf/, 2017.

[28] Stephan Hoyer and Joseph J.Hamman. Xarray: N-d labeled arrays and datasets in python. Journal of Open Research Software, 5(3), 2017.

[29] Weitao Chen, Xianju Li, Haixia He, et al. A review of fine-scale land use and land cover classification in open-pit mining areas by remote sensing techniques. Remote Sensing, 10(1), 2018.

[30] Xianju Li, Weitao Chen, Xinwen Cheng, et al. Comparison and integration of feature reduction methods for land cover classification with rapideye imagery. Multimedia Tools and Applications, 76(21):2304123057, Nov 2017.

[31] Xianju Li, Gang Chen, Jingyi Liu, et al. Effects of rapideye imagery's red-edge band and vegetation indices on land cover classification in an arid region. Chinese Geographical Science, 27(5):827835, Oct 2017.

[32] Jie Zhang, Jining Yan, Yan Ma, et al. Infrastructures and services for remote sensing data production management across multiple satellite data centers. Cluster Computing, 19(3):118, 2016.

[33] Ye Tian, Xiong Li, Arun Kumar Sangaiah, et al. Privacy-preserving scheme in social participatory sensing based on secure multi-party cooperation. Computer Communications, 119:167 178, 2018.

[34] Chen Chen, Xiaomin Liu, Tie Qiu, et al. Latency estimation based on traffic density for video streaming in the internet of vehicles. Computer Communications, 111:176 186, 2017.