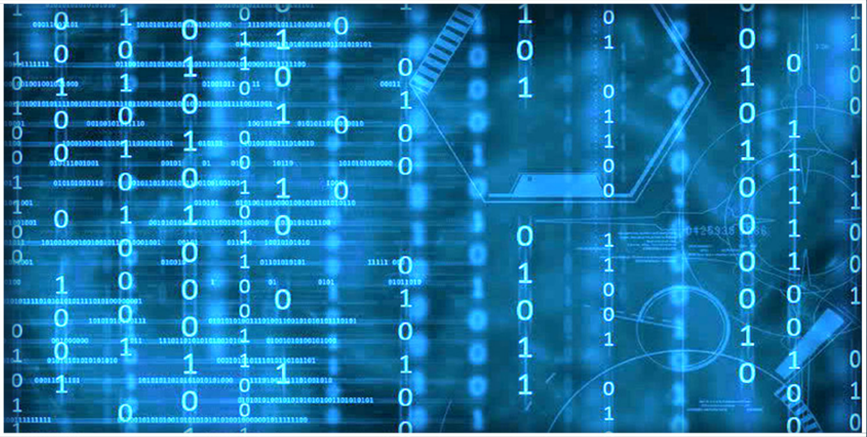
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 9 Issue 12

December 2018



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!

Managing Editor
IJACSA
Volume 9 Issue 12 December 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

Marwan Alseid

Mazin S. Al-Hakeem

Md Ruhul Islam

Md. Al-Amin Bhuiyan

Mehdi Bahrami

Mehdi Neshat

Messaouda AZZOUZI

Milena Bogdanovic

Miriampally Venkata Raghavendra

Mirjana Popovic

Miroslav Baca

Moamin Mahmoud

Moeiz Miraoui

Mohamed AbdelNasser Mohamed Mahmoud

Mohamed Salah SALHI

Mohamed A. El-Sayed

Mohamed Abdel Fatah Ashabrawy

Mohamed Ali Mahjoub

Mohamed Eldosoky

Mohamed Hassan Saad Kaloup

Mohamed Najeh LAKHOUA

Mohamed SOLTANE Mohamed

Mohammad Abdul Qayum

Mohammad Ali Badamchizadeh

Mohammad Azzeh

Mohammad H. Alomari

Mohammad Haghighat

Mohammad Jannati

Mohammad Zarour

Mohammed Abdulhameed Al-shabi

Mohammed A. Akour

Mohammed Ali Hussain

Mohammed Sadgal

Mohammed Shamim Kaiser

Mohammed Tawfik Hussein

Mohd Ashraf Ahmad

Mohd Helmy Abd Wahab

Mokhtar Beldjehem

Mona Elshinawy

Monir Kaid

Mostafa Mostafa Ezziyyani

Mouhammd sharari sharari alkasassbeh

Mounir Hemam

Mourad Amad

Mudasir Manzoor Kirmani

Mueen Uddin

Muhammad Adnan Khan

Muhammad Abdul Rehman

Muhammad Asif Khan

Muhammad Hafidz Fazli Bin Md Fauadi

Muhammad Naeem

Muhammad Saeed

Muniba Memon

MUNTASIR AL-ASFOOR

Murphy Choy

Murthy Sree Rama Chandra Dasika

MUSLIHAH WOOK

Mustapha OUJAOURA

MUTHUKUMAR S SUBRAMANYAM

N.Ch. Sriman Narayana Iyengar

Nadeem Akhtar

nafiul alam siddique

Nagy Ramadan Darwish

Najeed Ahmed Khan

Najib A. Kofahi

Namrata Dhanda

Nan Wang

Naseer Ali Alquraishi

Nasrollah Pakniat

Natarajan Subramanyam

Natheer Gharaibeh

Nayden V. Nenkov

Nazeeh Ghatasheh

Nazeeruddin Mohammad

Neeraj Kumar Tiwari

NEERAJ SHUKLA

Nestor Velasco-Bermeo

Nguyen Thanh Binh

Nidhi Arora

NILAMADHAB MISHRA

Nilanjan Dey

Ning Cai

Niraj Singhal

Nithyanandam Subramanian

Nizamud Din

Noura Aknin

Obaida M. Al-Hazaimeh

Olawande Justine Daramola

Oliviu Matei

Om Prakash Sangwan

Omaima Nazar Al-Allaf

Omar A. Alzubi

Omar S. Gómez

Osama Ali Awad

Osama Omer

Ouchtati Salim

Ousmane THIARE

P.V. Praveen Sundar

Paresh V Virparia

Parminder Singh Kang

PAUL CELICOURT

Peng Xia

Ping Zhang

CONTENTS

Validation of the Proposed Hardness Analysis Technique for FPGA Designs to Improve Reliability and Fault-Tolerance

Abdul Rafay Khatri[1], Ali Hayek[2], Josef Börcsök[3]
Department of Computer Architecture and System Programming,
University of Kassel, Kassel, Germany

***Abstract*—Reliability and fault tolerance of FPGA systems is a major concern nowadays. The continuous increase of the system's complexity makes the reliability evaluation extremely difficult and costly. Redundancy techniques are widely used to increase the reliability of such systems. These techniques provide a large area & time overheads which cause more power consumption and delay, respectively. An experimental evaluation method is proposed to find critical nodes of the FPGA-based designs, named "hardness analysis technique" under the proposed RASP-FIT tool. After finding the critical nodes, the proposed redundant model is applied to those locations of the design and the code is modified. The modified code is functionally equivalent and is more hardened to the soft-errors. An experimental set-up is developed to verify and validate the criticality of these locations found by using hardness analysis. After applying redundancy to those locations, the reliability is evaluated concerning failure rate reduction. Experimental results on ISCAS'85 combinational benchmarks show that a min-max range of failure reduction (14%-85%) is achieved compared to the circuit without redundancy under the same faulty conditions, which improves reliability.**

Keywords*—Dependability; fault injection; fault tolerance; reliability; single event effects

I. Introduction

Field Programmable Gate Array (FPGA) has been involved in various applications in the last couple of decades, such as aerospace, biomedical instrumentation, safety-critical systems, and spacecraft, due to their remarkable features. These features include parallelism, reconfiguration, separation of functions, self-healing capabilities, overall availability, low cost and low design turn-around time [1], [2]. Therefore, FPGA has become the core of many embedded applications. SRAM-based FPGA devices are sensitive to Single Event Effects (SEE), which can be caused by various sources, such as α-particles, cosmic rays, atmospheric neutrons, heavy-ion radiations and electromagnetic radiations (x-rays or gamma rays) [2], [3], [4]. When a charged particle hits a critical node of FPGA-based design, it generates the transient pulse which can produce a bit-flip effect. This phenomenon is known as Single Event Upsets (SEU). The failure rate for a component or system is the number of failures that occur per unit time.

Verilog HDL is one of the most widely used languages for implementing the design structure for Application Specific Integrated Circuit (ASIC) and FPGA-based designs [5]. Verilog HDL describes designs in various abstraction style, for example, gate, data-flow, and behavioural levels. For small designs, gate abstraction style is used, and testing & verification processes can directly and easily be applied to the designs. At this level, designs look more similar to the actual hardware design. For large designs, data-flow and behavioural abstraction styles are adopted to develop and implement the specification of the design in an HDL code [6].

The reliability of Integrated Circuits (ICs) is profoundly affected due to technology scaling. Due to shrinkage size of components, the reliability of the device is a challenge nowadays. One way to improve the reliability of these designs is redundancy, but it increases the area and time overheads. The reliability can be defined as "It is the probability that the circuit output is correct even in the presence of faults" [7]. Several SEE mitigation techniques have been presented in the literature to protect the FPGA-based designs from SEE effect. The reliability of the FPGA systems is improved by various error mitigation schemes such as multiple-redundancy with voting, Triple Modular Redundancy (TMR), hardened memory cell level, and Error Detection And Correction (EDAC) coding. Among all SEU mitigation techniques, TMR has become the most common practice because of its straightforward implementation and reliable results [8], [2], [9], [10], [11]. These mitigation methods reduce the failure rate (SER) in combinational logic in integrated circuits and improve the reliability.

An experimental set-up is presented to find the critical nodes of the designs. This set-up works on the code-level of the design, i.e. Verilog HDL. The proposed hardness analysis technique is developed under the tool "RASP-FIT" in a continuation of the previous work [6], [1]. In this work, the primary objective is to validate the proposed hardness technique which is used to find the most critical nodes of the design and to increase the fault tolerance capability and reliability of the FPGA-based designs by reducing the soft error failure rate. The redundancy of these sensitive locations is achieved by modifying the fault injection control unit. However, improving reliability by applying redundancy to the sensitive locations by code modification, area-overhead calculations, power and delay analyses are in progress. These locations are the most sensitive nodes to the permanent and transient faults. Few fault models such as bit-flip and stuck-at (1/0) are involved in the experimentation. Various benchmark circuits are considered & evaluated, and authors found that a significant improvement in reliability is achieved. The proposed approach is a simple, straightforward, and easy to use.

The remainder of this paper is organised as follows: The related work is presented in Section II. Section III presents an overview of the RASP-FIT tool and explains the proposed methodology to find sensitive locations in the design. Section IV describes the experimental set-up and their components in detail. Results are discussed in Section V. Finally, Section VI concludes the paper and presents some directions for future work.

II. Related Work

Reliability concerning soft errors has become a crucial issue in digital circuits due to technology scaling nowadays. Soft errors are transient errors that can cause digital circuits to operate incorrectly. If a soft error occurs in the combinational logic, it results in a Single Event Transient (SET). On the other hand, if it occurs in the memory cell itself, it results in a Single Event Upset (SEU). Both SET and SEU have a significant impact on circuit operation, and they should be adequately treated [12]. Soft Error Rate (SER) is a measurement evaluation metric for the sensitivity of the digital design to soft errors. SER estimation can be evaluated using two methods, i.e. dynamic and static. Fault injection and logic simulation techniques practice dynamic methods [13].

Authors in [14] demonstrated that, with increasing technology generations, soft errors caused more effects in logic devices than memory devices. Therefore, soft errors are becoming a major concern in digital systems. To overcome the effect of soft errors in combinational circuits, several fault tolerance techniques have been introduced in the literature. Fault tolerance techniques for combinational circuits are classified into three main categories: hardware redundancy-based, synthesis-based, and physical characteristics-based techniques [12].

The reliability of the FPGA systems is improved by various error mitigation scheme such as Triple Modular Redundancy (TMR). The problem with TMR technique is that it requires high area overhead nearly more than 200%. In [15], authors proposed a method for reducing the effect of SEU and called it "Selective Triple Modular Redundancy (STMR)". In this method, the conventional TMR technique is applied to selective gates of the whole design. These gates are more sensitive than other gates in the design. The signal probabilities of its inputs determine the sensitivity of a gate to an SEU. Soft error mitigation scheme based on logic implication is proposed in [16]. According to this method, the selective functionally redundant wires are attached to the combinational logic of a circuit. The procedure to find these functionally redundant wire is illustrated in this work.

Authors in [17] described the co-hardening technique, which is a technique that tries to reduce protection overheads complementing software mitigation techniques with hardware techniques in a selective way. Probabilistic Transfer Matrix (PTM) is a gate-level approach for the accurately measured reliability of combinational designs. The drawback of PTM is long simulation runtime and memory usage. Therefore, this technique is upgraded and called Efficient Computation PTM [18]. Authors in [13] proposed a new method for SER estimation based on vulnerability evaluation of the design gates. They introduced a probabilistic vulnerability window concept which considered three masking factors (i.e. logical, electrical and timing).

In this work, the validation of the proposed hardness analysis technique is presented in detail, which shows the reduction in failure rate and hence achieves the improvement in reliability. In comparison with the above works in general, our experimental set-up provides more enhancement in reliability with both permanent and transient fault models.

III. RASP-FIT Tool and Hardness Analysis

The RASP-FIT (RechnerArchitektur and SystemProgrammierung)–German name of the institute– Fault Injection Tool has the capability to instrument FPGA-based designs, for fault simulation and emulation of designs. This tool is designed specifically for the FPGA-based designs, which are written in Verilog at different abstraction levels. The tool consists of three major functions, namely, `fault_injection()`, `static_compaction()` and `hardness_analysis()`. RASP-FIT tool, with its Graphical User Interface (GUI), is developed in Matlab. All these functions are developed in Matlab under the function `RASP_FIT()`.

A. Verilog Code Modifier under RASP-FIT tool

The RASP-FIT tool has the capability to instrument FPGA-based designs, written in Verilog at different abstraction levels. This tool modifies or instruments the code by inserting faults. At each abstraction level, the way of modification of the code is different and also fault models are defined at that abstraction level [6]. In the modification process, it also adds Fault Injection, Selection and Activation (FISA) control unit in the target design. The FISA unit selects and activates the particular fault in the whole design.

Test and reliability evaluation using fault injection techniques require the modification of the design. Modification of Verilog code for various FPGA-based designs and obtaining the compact test vectors for maximum fault coverage using static compaction technique, also hardness analysis provides the information about the critical nodes of design are presented in previous work [19], [1], [6]. Fig. 1 shows the modified code of the original design with fault injection control unit. The RASP-FIT tool injects faults in every possible location in the design. This example shows the insertion of a bit-flip fault model in the design. For understanding purpose, we call it "full-faulty module" in this paper. The FISA control unit of the full-faulty module can select and activate all faults in the design. More detail about the RASP-FIT tool is described in [20].

B. Hardness Analysis: Identification of Sensitive Nodes

The sensitive location is the location in a System Under Test (SUT), where the occurrence of any fault results in a failure. The sensitive locations of the SUT are obtained using the proposed hardness analysis approach. According to this approach, these locations are more or less equally sensitive to bit-flip and stuck-at (1/0) faults. In this method, we obtained the more efficient test vectors/patterns which detect more faults than others. We selected a set-point value for each design and each fault model. The experimental way to obtain the sp-value is described in [21]. The procedure for

```
// Original design
module c17 (N1,N2,N3,N6,N7,N22,N23);

input N1,N2,N3,N6,N7;
output N22,N23;
wire N10,N11,N16,N19;

nand NAND2_1 (N10, N1, N3);
nand NAND2_2 (N11, N3, N6);
nand NAND2_3 (N16, N2, N11);
nand NAND2_4 (N19, N11, N7);
nand NAND2_5 (N22, N10, N16);
nand NAND2_6 (N23, N16, N19);

endmodule
```

```
// Compilable faulty design
module c17_1 (select,N1,N2,N3,N6,N7,N22_f1,
    N23_f1);
input N1,N2,N3,N6,N7;
output N22_f1,N23_f1;
wire N10,N11,N16,N19;
input[3:0] select;
reg f0,f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11;
always @ (select) begin
if (select == 4'd0) begin
f0=fis;f1=0;f2=0;f3=0;f4=0;f5=0;f6=0;f7=0;f8
    =0;f9=0;f10=0;f11=0;end
else if (select == 4'd1) begin
f0=0;f1=fis;f2=0;f3=0;f4=0;f5=0;f6=0;f7=0;f8
    =0;f9=0;f10=0;f11=0;end
.
.
.

else if (select == 4'd11) begin
f0=0;f1=0;f2=0;f3=0;f4=0;f5=0;f6=0;f7=0;f8
    =0;f9=0;f10=0;f11=fis;end
else begin
f0=0;f1=0;f2=0;f3=0;f4=0;f5=0;f6=0;f7=0;f8
    =0;f9=0;f10=0;f11=0;end
end
nand NAND2_1 (N10,f0^ N1,f1^ N3);
nand NAND2_2 (N11,f2^ N3,f3^ N6);
nand NAND2_3 (N16,f4^ N2,f5^ N11);
nand NAND2_4 (N19,f6^ N11,f7^ N7);
nand NAND2_5 (N22_f1,f8^ N10,f9^ N16);
nand NAND2_6 (N23_f1,f10^ N16,f11^ N19);
endmodule
```

Fig. 1. Original code (left) & instrumented compilable design code (right) by RASP-FIT.

obtaining the qualified test patterns is outlined in algorithm 1. Once the efficient test patterns are obtained, then hardness analysis is performed on them using the RASP-FIT tool under hardness analysis function. The sequel describes the procedure to calculate hardness.

Algorithm 1 Dynamic compaction algorithm in a nutshell

```
1: Input patterns are applied using LFSR from total vector space $T_S$
2: Fault detections are counted for each pattern applied
3: Sum of detections are compared with set-point value
4: if Is sum greater than or equal to set-point value then
5:     Stored pattern as qualified test vectors $T_q$
6:     Increment the number of $T_q$ count
7: else
8:     Apply new pattern to the SUT
9: end if
10: Go to step 2
11: Stop simulation when $T_q$ count reaches 500
```

The whole experimental set-up is shown in Fig. 2 in order to find sensitive locations. In the first step, a Verilog design file is applied as an input to the RASP-FIT tool. The RASP-FIT tool generates the faulty copies of the original design with evenly distributed faults in them. The user selects the fault models and the number of defective modules. This tool generates a file (top module design) which contains the instantiations of different modules, comparator logic, dynamic compaction logic to select efficient test vectors and memory logic to store responses. Once the faulty modules are generated, the Xilinx ISE and Modelsim tools are used to create the project and simulate the designs. Repeat this procedure for every fault models, e.g. bit-flip, stuck-at 0 and stuck-at 1 fault models. The simulation results are stored in text (*.txt) file which is further applied as an input to the RASP-FIT tool to perform the hardness analysis. The simulation set-up generator, shown in Fig. 4, is explained in Section IV to validate the hardness analysis technique. The technique is described in the sequel.

Hardness or Hard to Detect (HTD) is the characteristic of those faults which can be detected very rarely. The hardness analysis receives text files as input, reads them and stores the data in a matrix form. Authors call it Fault Matrix (FM), which is an arrangement of qualified input patterns and the detection of faults for the input patterns given in Eq. 1,

$$\text{FM} = \begin{bmatrix} P_1 & F_{1,1} & F_{2,1} & \cdots & F_{N,1} \\ P_2 & F_{1,2} & F_{2,2} & \cdots & F_{N,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_i & F_{1,i} & F_{2,i} & \cdots & F_{N,i} \end{bmatrix} \qquad (1)$$

Where P_1 to P_i are qualified input patterns obtained during fault-experiment, and the array of detected faults for a particular pattern are placed in a row of the matrix. When

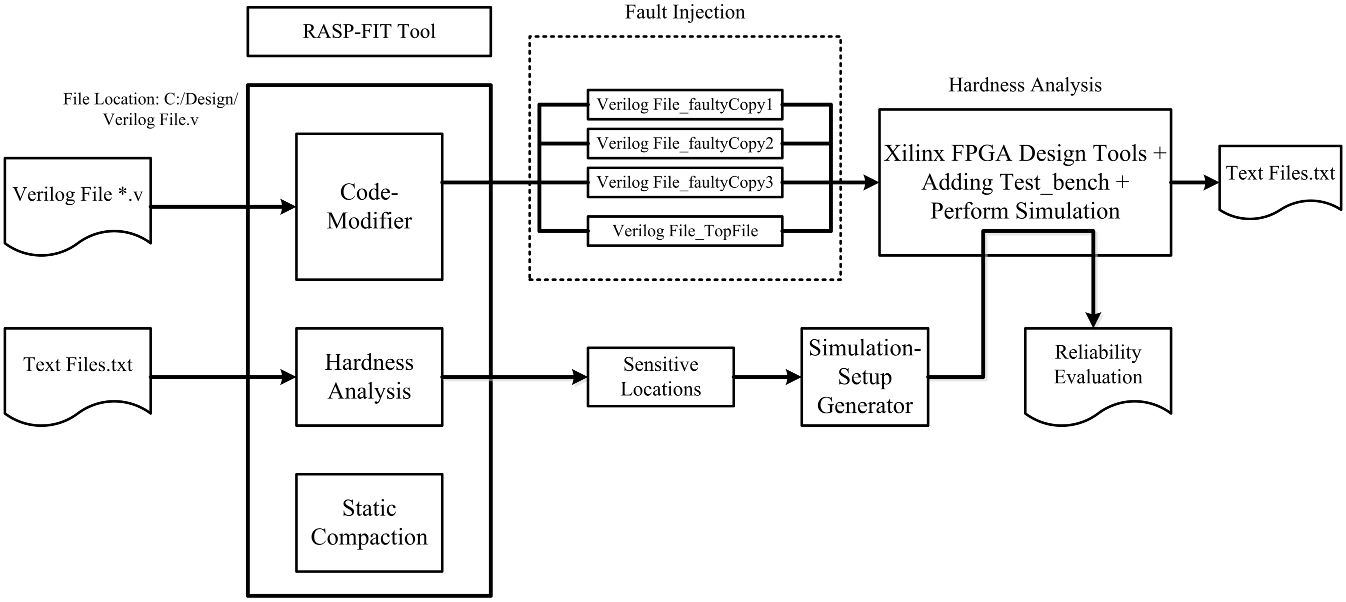


Fig. 2. Overall flow of RASP-FIT tool, hardness analysis and simulation set-up generator.

the specific fault is detected, it gets value '1', otherwise gets value '0'. Hardness *(H)* of individual fault is calculated by Eq. 2,

$$H = \left(1 - \frac{\text{No. of Fault Detections}}{\text{Total Patterns in FM}}\right) X100 \qquad (2)$$

If the hardness of a fault results in 100%, it means the fault is not detectable for any input; hence, it is called an un-testable or undetectable fault. On the other side, a hardness of 0% shows the detection of fault for all test vectors, which means that the portion of the circuit where the fault has appeared is very critical to fault attacks. The hardness of each fault for every fault model is calculated and placed in a matrix, named Hardness Matrix (HM) as shown in Eq. 3,

$$\text{HM} = \begin{bmatrix} H_{f_1,bf} & H_{f_2,bf} & ... & H_{f_N,bf} \\ H_{f_1,sa0} & H_{f_2,sa0} & ... & H_{f_N,sa0} \\ H_{f_1,sa1} & H_{f_2,sa1} & ... & H_{f_N,sa1} \end{bmatrix} \qquad (3)$$

Hardness values of each fault for every fault model are arranged column-wise. Each column is compared with the different threshold values one after the other for each fault to get the number of sensitive locations. Threshold values are used to get the number of the most sensitive areas to less vulnerable places. There are four levels considered for obtaining sensitive locations empirically, i.e. 35%, 55%, 75% and 95%. Once, we identify the sensitive locations for different threshold values, we will apply the proposed redundant model on to those locations only and obtained the modified code for the design under experiment. The modified code is more hardened than the original design. The proposed redundant model is described in the sequel.

Development of Redundant Model: The sensitive location is merely a wire/net either connecting two gates or input of gates. Redundant model is developed based on the Duplication With Comparison (DWC) technique, in which the sensitive location is duplicated, and their outputs are compared using XOR gate. The output of the XOR is feed to the select input of a multiplexer, which routes the correct output. It is noted that this redundant model is used for the validation of the proposed hardness analysis technique. These sensitive locations should be made hardened to improve reliability and reduce the soft error rate. The sensitive node is replaced by the proposed redundant model as shown in Fig. 3, for the wire/net without fan-out. This work is in progress to modify the original code according to the redundant model to add hardening.

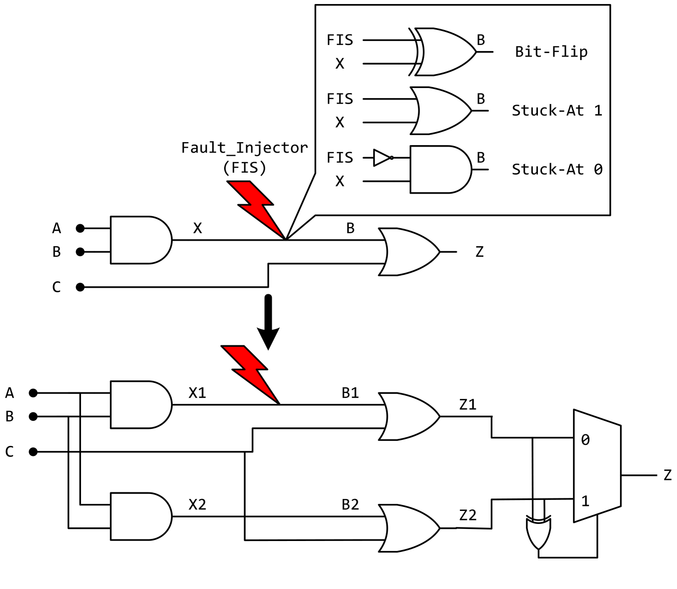


Fig. 3. Redundant fault model for validation of approach.

IV. Experimental Set-Up

To validate the efficiency of the proposed hardness analysis technique in the process of improving fault tolerance capabilities and reliability of FPGA-based designs, we created an experimental set-up. Various benchmark designs are evaluated, and the results are presented in this work. Fig. 4 shows the experimental set-up for the validation of the proposed hardness analysis technique. Fig. 4 shows the description of each components in the sequel.

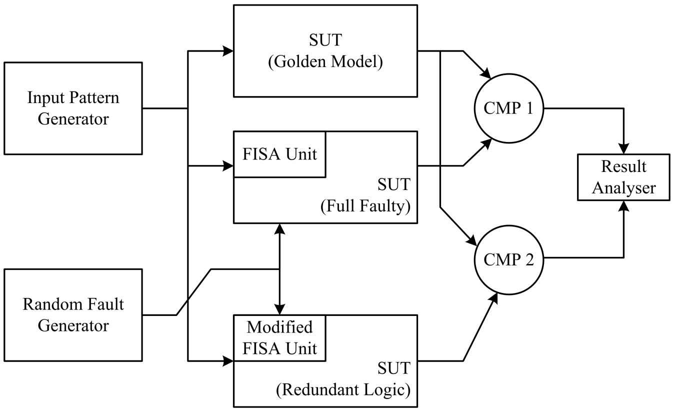


Fig. 4. Experimental set-up for the validation for the hardness analysis technique.

A. Input Pattern Generator

This block is used to generate different input patterns which are simultaneously applied to both the original and faulty copies of the SUT. For small circuits having few input ports, we generate and apply all possible combination of inputs. For a large number of inputs, random input patterns are generated and applied to the golden, full-faulty and redundant SUT simultaneously. These patterns are generated using the Linear Feedback Shift Register (LFSR), defined in the test-bench. Fig. 5 shows the code for the generation of input patterns randomly for the SUT `(c432.v)` having 36 inputs.

```
// Random input generator code for 36
    input design

reg [35:0] e = 36'hF59611d09; //seed

always @(clk) begin
e = { e[34:0],  e[35] ^ e[34] };
{N1,N4,N8,N11,... ,N73,N76} = e ;
end
```

Fig. 5. Generate random input patterns in test-bench.

B. Random FISA (Fault Generator)

In this experimental set-up, random faults are generated, selected and activated in both the full-faulty module and the redundant module. Poisson distribution is used to generate random faults because of the following reasons:

1) The event is something that can be counted in whole numbers.
2) Occurrences are independent. Therefore, one occurrence neither diminishes nor increases the chance of another.
3) The average frequency of occurrence for the period in question is known.
4) It is possible to count how many events have occurred.

Poisson distributed numbers can be generated using this code in the test-bench. During the simulation, for each pattern, 40 faults are selected and activated using Random FISA unit in each copy of the SUT shown in Fig. 6 for each pattern applied. The total of 500 patterns is stored with their responses in a text file for the further calculations.

```
always
  begin
    #5; select = $dist_poisson(seed, mean);
  end
```

Fig. 6. Part of test_bench to generate random faults for FISA unit.

C. Golden Model (SUT)

Golden SUT is the original module without faults, and it is a reference design for the comparison between the faulty SUT and the SUT with redundant logic. Various combinational logic circuits from ISCAS'85 benchmark are considered for the validation of the proposed hardness analysis technique.

D. Full-Faulty SUT with FISA Unit

Faulty SUT is the modified benchmark design by injecting faults in every possible location. The RASP-FIT tool generates the faulty SUT. This tool is capable of instrumenting the Verilog design, written in all abstraction levels, by injecting various types of permanent and transient faults in all possible location. The user can choose between the types of fault models for analysis. Fig. 1 depicts the original code and modified code generated by RASP-FIT tool.

E. Redundant SUT with Modified FISA Unit

Once the information about the critical nodes is obtained, redundancy is applied to them. Fig. 3 shows the proposed redundant model. When the fault occurs on this sensitive location, the multiplexer logic masks it. However, we modified the FISA unit included in the SUT with redundant logic in this work. With this modification, the FISA unit does not activate the particular sensitive fault in the design, even though, it is selected by random FISA input generator.

F. Comparators

One comparator logic is used to compare the responses of golden SUT with the full-faulty SUT, whereas another comparator logic compares the responses of golden SUT with the SUT having modified FISA unit for sensitive locations. The value of comparator is logic ‘1’ when both responses are different from each other and logic ‘0’ in another case.

G. Result Analyser

Result analyser is developed in Matlab. The failure rate for both modules is stored during simulation in a text file. Result analyser program reads the text file containing responses and calculates the number of fault detections for both designs.

Reliability Improvement Calculations: Based on the definition of reliability mentioned earlier in the introduction section, the reliability of the combinational logic system can be improved by reducing the failure rate. There are two main methods which are most widely used, i.e. error detection and retry and error masking. In the proposed technique, full-faulty and the module with redundancy are run under the same conditions. Both modules are compared with the golden reference module design which is a fault-free design. Several faults are randomly selected and activated, and fault detections for both modules are stored in a text file. Reliability evaluation program is written in Matlab which takes responses of the experimental set-up stored in a text file and count the number of errors occurred for both modules. Improvement in reliability is calculated, in term of failure rate reduction, using the Eq. 4,

$$FR = \frac{SER_{fullFaulty} - SER_{redundant}}{SER_{fullFaulty}} X100\% \quad (4)$$

Where FR is a failure rate reduction (SER reduction) achieved after making the critical locations as redundant by using modified FISA unit, $SER_{fullFaulty}$ is a fault detection for full-faulty SUT and $SER_{redundant}$ is a fault detection for redundant SUT with modified FISA unit.

V. Results and Discussion

The primary objective of the work is to obtain the high reliability with applying the selective redundancy to some sensitive locations. In this way, a cost-effective solution is obtained with high reliability. These following steps describe the procedure used to perform the proposed technique:

1) Using the RASP-FIT tool, first, perform the fault injection analysis to generate faulty code of the design using Verilog code modifier function.
2) Repeat the step 1 for all fault models, e.g. bit-flip, stuck-at 0 and stuck-at 1.
3) Create a simulation set-up using Xilinx ISE and Modelsim tool.
4) Using the RASP-FIT tool, perform hardness analysis and find the sensitive locations of the design.
5) Modify the fault control unit for the most sensitive places as described earlier.
6) Create a simulation set-up as explained in Section IV.
7) Run simulations for all fault models and save the simulation results in text files separately.
8) Import all these files in Matlab and evaluate reliability.

To explain the proposed method, we consider a simple benchmark design `c17.v` from ISCAS'85 designs as an example (Fig. 1). In this design, the RASP-FIT finds and injects a total of 12 faults in the whole design. Hardness analysis is performed, and it is found that the 4 locations are most sensitive (at Threshold = 55%) and we have to apply for the redundancy on these locations. Fig. 7 shows the critical nodes for various benchmark designs at different threshold values. The redundant module is proposed in Section III, but it is not used here because, in this work, we are validating the proposed hardness analysis technique regarding failure rates. In this work, we modified the FISA unit such that these four faults never selected and activated. Total number of fault selected and activated during the experiment can be obtained using the Eq. 5,

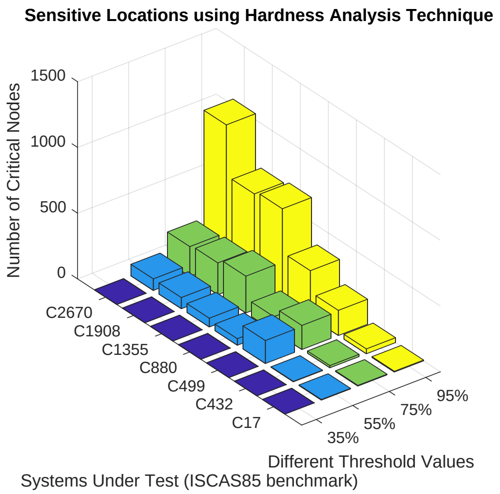


Fig. 7. Number of critical nodes for different threshold.

$$T_{fault} = P_i \times nSec \times F_{copy} \quad (5)$$

Where T_{fault} is the total number of fault injected per experiment for each fault model, P_i is the number of patterns considered, i.e. 500, nSec is the number of faulty copies of SUT and F_{copy} is the number of faults selected and activated per copy of SUT during the experiment.

We perform the simulation of a full-faulty SUT and the SUT with modified FISA unit. Soft error rate (failure rate) for the full-faulty module is observed with random fault injection. Similarly, the detection rate for the module with redundancy is also recorded for each fault model separately. Reliability improvement is calculated and presented in Table I. The results validate that the obtained locations using the proposed hardness analysis technique are susceptible. By applying redundancy or masking/hardening techniques to these locations, it shows the significant improvement in the reliability as shown in the last column of Table I.

VI. Conclusion

In this paper, the authors introduced a method to increase the fault tolerance capability and the reliability of combinational circuits. The idea is based on first finding the sensitive locations of the design using the proposed hardness analysis technique and then applying redundancy/hardening to those nodes only. In this way, the overall fault tolerance of the original circuit is enhanced and the failure rate is also reduced as well. Experimental results on ISCAS'85 combinational benchmarks show that we can get a maximum reliability improvement of 85%. The approach is straightforward and easy to use.

Future work includes the analysis of area-overhead due to redundant logic applied to the sensitive locations. Also, the delay and power consumption analyses are the core areas after

TABLE I. IMPROVEMENTS IN RELIABILITY AND FAULT TOLERANCE CAPABILITY OF FPGA-BASED DESIGNS

SUT	Fault Model	Total Fault (T_{fault}) Injected/Experiment	Number of Sensitive Nodes	Faulty Module Detection Rate	Redundant Module Detection Rate	SER Reduction (%)
c17	Bit-flip	20,000	4/12	9759	3435	64.802
	Stuck-at 0			6104	1973	67.68
	Stuck-at 1			3689	1403	61.79
c432	Bit-flip	60,000	17/336	5641	4441	21.28
	Stuck-at 0			2581	2208	14.95
	Stuck-at 1			3005	2180	27.45
c499	Bit-flip	60,000	177/408	24689	3607	85.39
	Stuck-at 0			7058	1118	84.15
	Stuck-at 1			17602	2484	85.88
c880	Bit-flip	60,000	155/729	22780	16442	27.823
	Stuck-at 0			11703	8411	28.13
	Stuck-at 1			11077	8203	25.95
c1355	Bit-flip	60,000	67/1064	10604	8454	20.28
	Stuck-at 0			2930	2276	22.32
	Stuck-at 1			7674	6178	19.50
c1908	Bit-flip	100,000	242/1498	32962	21874	33.64
	Stuck-at 0			18765	12171	35.14
	Stuck-at 1			14197	9703	31.65

applying redundancy to the critical/sensitive nodes in the future work.

REFERENCES

[1] A. R. Khatri, A. Hayek, and J. Börcsök, *Applied Reconfigurable Computing*, vol. 9625 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016.

[2] W. Xin, "Partitioning Triple Modular Redundancy for Single Event Upset Mitigation in FPGA," in *2010 International Conference on E-Product E-Service and E-Entertainment*, (Henan), pp. 1–4, IEEE, Nov 2010.

[3] M. Desogus, L. Sterpone, and D. M. Codinachs, "Validation of a tool for estimating the effects of soft-errors on modern SRAM-based FPGAs," in *2014 IEEE 20th International On-Line Testing Symposium (IOLTS)*, (Platja d'Aro, Girona, Spain), pp. 111–115, IEEE, Jul 2014.

[4] L. A. C. Benites and F. L. Kastensmidt, "Automated design flow for applying Triple Modular Redundancy (TMR) in complex digital circuits," in *2018 IEEE 19th Latin-American Test Symposium (LATS)*, pp. 1–4, IEEE, Mar 2018.

[5] H. Ben Fekih, A. Elhossini, and B. Juurlink, *Applied Reconfigurable Computing*, vol. 9040 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2015.

[6] A. R. Khatri, A. Hayek, and J. Börcsök, "Validation of the Proposed Fault Injection , Test and Hardness Analysis for Combinational Dataflow Verilog HDL Designs under the RASP-FIT Tool," in *2018 IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong.*, (Athens, Greece), pp. 544–551, IEEE Comput. Soc, 2018.

[7] G. dos Santos, E. Marques, L. d. B. Naviner, and J.-F. Naviner, "Using error tolerance of target application for efficient reliability improvement of digital circuits," *Microelectronics Reliability*, vol. 50, pp. 1219–1222, Sep 2010.

[8] A. R. Khatri, A. Hayek, and J. Börcsök, "RASP-TMR: An Automatic and Fast Synthesizable Verilog Code Generator Tool for the Implementation and Evaluation of TMR Approach," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 590–597, 2018.

[9] P. Balasubramanian, K. Prasad, and N. E. Mastorakis, "A Fault Tolerance Improved Majority Voter for TMR System Architectures," *WSEAS Transactions on Circuits and Systems*, vol. 15, pp. 108–122, 2016.

[10] S. Müller, T. Koal, M. Schölzel, and H. T. Vierhaus, "Timing for Virtual TMR in Logic Circuits," in *IEEE 20th InternationalOn-Line Testing Symposium (IOLTS)*, pp. 190–193, 2014.

[11] S. Di Carlo, G. Gambardella, P. Prinetto, D. Rolfo, P. Trotta, and A. Vallero, "A novel methodology to increase fault tolerance in autonomous FPGA-based systems," in *2014 IEEE 20th International On-Line Testing Symposium (IOLTS)*, (Girona, Spain), pp. 87–92, IEEE, Jul 2014.

[12] A. H. El-Maleh and K. A. K. Daud, "Simulation-Based Method for Synthesizing Soft Error Tolerant Combinational Circuits," *IEEE Transactions on Reliability*, vol. 64, pp. 935–948, Sep 2015.

[13] M. Raji, H. Pedram, and B. Ghavami, "Soft error rate estimation of combinational circuits based on vulnerability analysis," *IET Computers & Digital Techniques*, vol. 9, pp. 311–320, Nov 2015.

[14] P. Shivakumar, M. Kistler, S. Keckler, D. Burger, and L. Alvisi, "Modeling the effect of technology trends on the soft error rate of combinational logic," in *Proceedings International Conference on Dependable Systems and Networks*, pp. 389–398, IEEE Comput. Soc, 2002.

[15] P. Samudrala, J. Ramos, and S. Katkoori, "Selective triple Modular redundancy (STMR) based single-event upset (SEU) tolerant synthesis for FPGAs," *IEEE Transactions on Nuclear Science*, vol. 51, pp. 2957–2969, Oct 2004.

[16] S. Almukhaizim and Y. Makris, “Soft Error Mitigation Through Selective Addition of Functionally Redundant Wires,” *IEEE Transactions on Reliability*, vol. 57, pp. 23–31, Mar 2008.

[17] F. Kastensmidt and P. Rech, *FPGAs and Parallel Architectures for Aerospace Applications*. Cham: Springer International Publishing, 2016.

[18] H. Cai, K. Liu, L. A. de Barros Naviner, Y. Wang, M. Slimani, and J.-F. Naviner, “Efficient reliability evaluation methodologies for combinational circuits,” *Microelectronics Reliability*, vol. 64, pp. 19–25, Sep 2016.

[19] A. R. Khatri, A. Hayek, and J. Börcsök, “ATPG method with a hybrid compaction technique for combinational digital systems,” in *2016 SAI Computing Conference (SAI)*, (London, UK), pp. 924–930, IEEE, Jul 2016.

[20] A. R. Khatri, A. Hayek, and J. Börcsök, “RASP-FIT : A Fast and Automatic Fault Injection Tool for Code-Modification of FPGA Designs,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 30–40, 2018.

[21] A. R. Khatri, A. Hayek, and J. Börcsök, “Validation of selecting SP-values for fault models under proposed RASP-FIT tool,” in *2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*, (Karachi, Pakistan), pp. 1–7, IEEE, Nov 2017.

Utilization of a Neuro Fuzzy Model for the Online Detection of Learning Styles in Adaptive e-Learning Systems

Luis Alfaro[1], Claudia Rivera[2], Jorge Luna-Urquizo[3],
Elisa Castañeda[4]
Universidad Nacional de San Agustín
Arequipa, Perú

Francisco Fialho[5]
Universidade Federal de Santa Catarina
Florianópolis, Brazil

***Abstract*—After conducting a historical review and establishing the state of the art of the various approaches regarding the design and implementation of adaptive e–learning systems —taking into consideration the characteristics of the user, in particular their learning styles and preferences in order to focus on the possibilities for personalizing the ways of utilizing learning materials and objects in a manner distinct from what e–learning systems have traditionally been, which is to say designed for the generic user, irrespective of individual knowledge and learning styles— the authors propose a system model for the classification of user interactions within an adaptive e–learning platform, and its analysis through a mechanism based on backpropagation neural networks and fuzzy logic, which allow for automatic, online identification of the learning styles of the users in a manner which is transparent for them and which can also be of great utility as a component of the architecture of adaptive e–learning systems and knowledge-management systems. Finally, conclusions and recommendations for future work are established.**

***Keywords*—e-Learning; learning style identification; backpropagation neural network; fuzzy logic; neuro fuzzy systems**

I. Introduction

Learning is not a process of accumulation of representations of the external environment. Rather, it is a continuous process of behavioral transformation by way of continuous change in the capacity of the nervous system to synthesize new information. Memory does not depend on the indefinite retention of an invariant structure representing an entity (e.g., an idea, image, or symbol), but rather the functional ability of the system to create when certain recurrence conditions are given, for example, a behavior that satisfies recurrent demands or one that the observer would classify as a reactivator of a previous behavior.

For Dittus and Vasquez [1], "one commonly has the idea that the nervous system is an instrument which obtains information from the environment that the organism subsequently utilizes in order to construct a model of the world". Every autopoietic unit is unique because it is characterized by the phylogenetic inheritance of its ancestors, as well as its life history or ontogeny [2], defined as "the history of structural change of a unit, without which it loses its organization". It is in this way that different human beings possess different ways of learning. Some can construct knowledge in a more optimal way when they receive information through auditory pathways, while others do so visually or through other senses. However, for Alshammari [3], e-learning systems do not consider the diversity of learning types, learners' abilities, learners' knowledge, or the learning context.

Nowadays, in many teaching/learning activities, be they traditional or the utilization of technological resources or emergent media, learner differences in the classroom, such as the different learning styles [4], they may possess, tend to be ignored or simply not considered. This fact tends to result in the stardardization of methodologies, strategies, and techniques for different kinds of students, and this is understandable because it is extremely difficult for a teacher to apply multiple teaching strategies in the classroom. Currently, adaptive e-learning provides new ways of focusing traditional models of education, thus making possible the personalization of characteristics and educative experiences for each type of learner [5]. Among these characteristics, learning styles are one of the most important factors in learning. Adaptive systems focus on the transformation of learning from the passive receptor of information to collaborator in the educative process.

For Joy and Kolb [6], the types of learning styles indicate the differences in perspective regarding learning, based on individual preferences and considering the dialectical combination of those modalities. Learning styles are the cognitive, affective, and physiological traits which serve as relatively-stable indicators of how students perceive interactions and respond to their learning environments. They describe a learner in terms of the educative conditions which are more favorable for his/her learning. In this sense, the identification of student's learning styles is considered a vital element, and diverse perspectives for the identification of them have been developed.

In this paper, a model is proposed based on an analysis of user interactions within an e-learning platform, utilizing the concepts of 'fuzzy logic' and 'neural networks' with the objective of identifying individual learning styles and adapting the contents of the platform to the demands, preferences, and learning styles of each user.

II. Theoretical Framework

In this section, the authors conduct a historical review, identifying the state of the art and exploring the theoretical foundation of the research project.

A. e- Learning systems

The diversity and heterogeneity of resources available on the internet, the newest trends in methodologies and teaching/learning tools, and the current needs of the users make it indispensable to have at one's disposal virtual learning environments which possess the characteristics of adaptation and content personalization, as well as virtual assistants, among other tools. In this context, one of the lines of research which has recently seen a lot of activity is that of e-learning. According to H. Hashim and Z. Tasir [7], "an e-learning platform is that which applies and utilizes electronic media and information and communication technology (ICTs)". E-learning can imply other alternative terms, such as 'online education', 'computer-based, e-learning systems', and others. If the root of the word is taken as a reference, e-learning is translated as electronic learning, and in such a way, in its broadest conceptualization, it can encompass virtually any educative activity that utilizes electronic media in order to realize all or part of a learning process. This particular reference has arisen due to other online services, for example, e-business or e-commerce.

The following are components of an e-learning course, complementary to the instructional strategy: objectives; study cases; readings; centers of knowledge; conceptual maps; complementary, instructional materials and elements of interactivity and evaluation, for example, animations, simulations, interactive tasks, glossaries, biographies, self-evaluation exercises, and open-ended question exercises; material format: slides, media clips, linear text, multimedia, graphics, digital video, and audio; navigation tools, such as arrows for going forward or back; print copies; online help; site maps; filters; chat applications; forums; and email; some of which can better attend to the demands and preferences of the users.

Morales [8] establishes e-learning participants and their respective responsibilities in the following manner:

1) Teachers/tutors: Their role is to facilitate learning, for which they have to supply the tools so that the student learns autonomously and is capable of constructing his/her own knowledge in an active and responsible manner.

2) Students: The students need to have planning capability; flexibility to adapt to new and different ways of learning, as well as the traditional modalities; the capability to participate/integrate in the virtual group; technical competency in the navigation and use of new technologies, as well as a favorable attitude towards them; and time availability for learning within or apart from the work schedule, depending on the case.

According to Alshammari et al. [3], in e-learning systems, the learner may be overwhelmed by the great quantity of information that he/she encounters. The student might make poor decisions in relation to the subjects or material under study. Learning may demand a lot of time or create confusion and/or frustration, so for this reason, it might not be very effective. One of the modifications for the development of e-learning systems consists of being familiar with the differences among students, as well as their indvidual needs, with the objective of providing a personalized learning system that gives better relevance to the instructional material in accordance with the demands and needs of the student.

Adaptive e-learning systems based on different learning styles generally use different learning-style models. This raises the issue of which models and theories are most suitable and effective as components of these environments. An adaptive e-learning system based on knowledge level and learning style has been designed and implemented by Alshammari et al. [3]. This system facilitates personalized-learning pathways through the organization of material links according to their relevance to a particular learner; it also provides adaptive guidance and feedback to support learner-system interaction goals. Using a standard usability instrument, an experimental evaluation concerning learners' perception of usability was conducted to compare the adaptive e-learning system with a non-adaptive version, which yielded favorable results for the former.

In other words, understanding the needs of the students and identifying their learning patterns and preferences is crucial with regard to the design of e-learning-systems material in accordance with distinct learning styles, in this manner closing the resulting breach in relation to the members of the triangular community, which is to say the students, instructors, and adaptive contents online. It is necessary to establish what is required in order to capture the attention of each student and satisfy the demands and needs of his/her natural learning style so that what is learned is retained over the long term. Therefore, for Abdullah [9], identifying the learning styles is considered a vital element in the design of e-learning systems.

Lo and Shu [10] point out that the majority of authors in the field concur that the consideration of learning styles in the pedagogical process can increase the efficiency and efectiveness of learning. In this sense, diverse approaches have been developed for the identification of learning styles. Particularly, in the current paper, individual learning styles are identified in order to focus on subsequently the adaptation of platform content in accordance with the demands, preferences, and learning and thinking styles of each user, attempting in this way to supply the particular learning resources and objects that the student prefers.

B. Proposal for e-Learning Adaptive System Architecture

The authors propose an adaptive-system architecture based on autonomous intelligent agents for the implementation of a virtual-learning platform, given that this has proven itself to be the approach with the most potential in the field. Among their principle advantages are:

- They permit the modelling of individual profiles for each student, thus facilitating tasks such as the search for information and contents.
- They facilitate the incorporation of a knowledge-representation model and can facilitate the tasks of adaptation and personalization of contents in the proposed platform.
- They permit the incorporation of machine-learning characteristics in conjunction with other approaches and techniques of artificial intelligence.
- They can be equipped with various characteristics, such as autonomy, initiative, mobility (including among distinct platforms), and adaptability, among others.

The architecture of the proposed mulit-agent model, as with the description of its components, can be found in Alfaro et al. [11], which is shown in Figure 1.

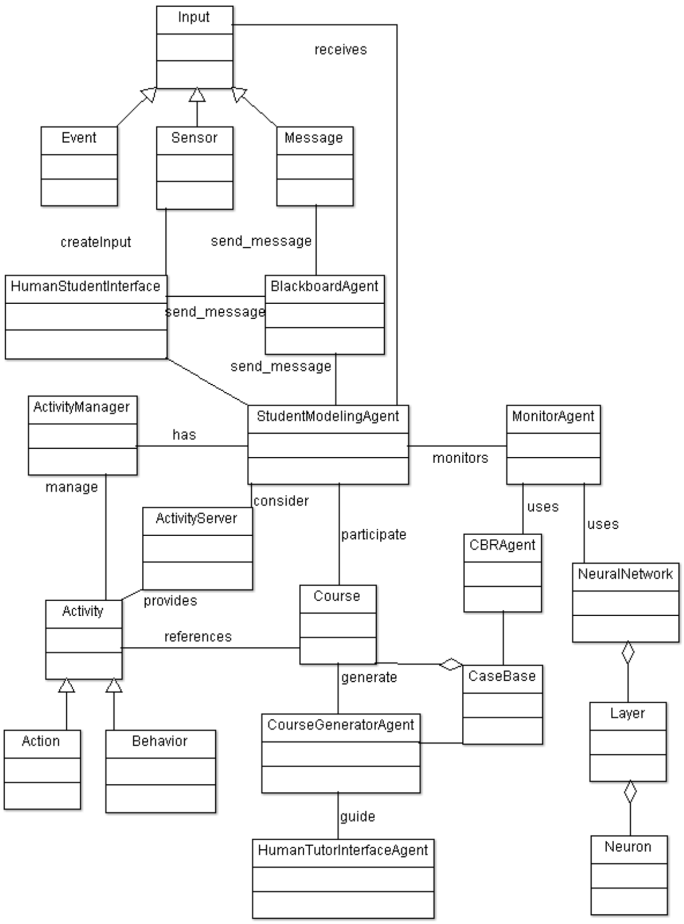


Fig. 1. Architecture of the multi-agent system [11]

The implementation of the proposed intelligent agents was realized utilizing the JADE platform, which is an agent platform distributed with a container for each host, in which the agents are executed and which possesses storage for diverse languages and ontologies, complying with FIPA (Foundation for Intelligent Physical Agents) specifications, for which developed agents can easily be integrated in other languages and platforms, including owners.

The originality of this hybrid proposal resides in the fact that it incorporates diverse artificial-intelligence techniques, such as ‘intelligent agents’, a ‘backpropagation neural network’, ‘fuzzy logic’, and ‘case-based reasoning’. It also incorporates the ‘learning-based-on-projects’ paradigm.

The current proposal principally focuses on the model possessing a high degree of adaptability to the student’s demands.

C. Backpropagation Neural Networks - RNAs

Backpropogation is a training method used for a multi-layer neural network. It can be thought of as a generalization of the “Delta Rule” for direct networks with more than two layers. In this case, at least one layer of neurons is not involved with the input or output and is, therefore, internal to the network. This layer and its connections, when it learns to effectuate a function, acts as if there were an internal representation of the problem’s solution. Without going in to detail, backpropagation is a supervised learning rule. If an example is presented to the network, and the network output is verified, it is compared to the expected output, yielding an error. The gradient of this error is calculated in relation to to the synaptic values of the output layer, which is then updated by a selected step. The output error of the penultimate layer can, therefore, be calculated, and in this manner, starting at the front, the error (origin of the name ‘backpropagation’) propagates backwards through all the connection layers.

RNAs possess innumerable algorithms for pattern recognition: Kohonen, Perceptron, Adaline, and many others, each with its own specificities. For R. Lanelhas [12], the principle advantage of using RNA backpropagation is that it works with multiple layers and solves ‘non-linearly-separable’ problems that some algorithms cannot solve. Therefore, they can be counted among the networks proposed by Honey and Munford for the identification of learning styles.

Another important characteristic is that backpropagation is feedforward, which means that the connnection between neurons is not cyclic.

The RNA backpropagation is multi-layered, as it has a minimum of three layers. There are many calculations involved in the process so that the weight is adequately readjusted.

The RNA theory has provided an alternative to classical computation for those problems for which traditional or common methods have yielded not-very-convincing or disappointing results. This project in particular focuses on the online identification of learning styles, which has to do with pattern recognition within imprecise limits. Its degree of complexity made posible the incorporation of ‘fuzzy logic’, which, upon conducting the experimentation and corresponding tests, permitted the obtainment of superior results, as is discussed in the corresponding section.

D. Fuzzy Logic

For Timothy [13], fuzzy logic can be seen as a formalization mode of imprecise reasoning that represents certain human capacities to make approximate inferences and judgements within conditions of uncertainty.

According to Ozdemir et al. [14], determining the learning style most adequate to the individual capacities of the student is very important for quick, easy, and effective learning. However, the quantification of said capacities and the rules to follow in order to determine the most convenient learning style are of an imprecise nature, for which any approach one wishes to follow should incorporate fuzzy-logic techniques. In the particular case of the perspective developed by the aforementionded authors, an ‘expert system’ is proposed, in which the membership or belonging functions, as well as each one of the inputs and outputs of the inferrence rules, employ concepts of fuzzy logic.

Palomino et al. [15] part from the premise from which it is possible to define much more practical mechanisms adjusted to the real educative action for the detection of learning styles, utilizing techniques associated with fuzzy logic. The proposed approache is based on the concept of learning pathways as a way to establish the type of preference that the learners possess with respect to how they perceive and process information, where the inputs are defined by fuzzy combinations.

Stathacopoulou et al. [16] point out that the neuro-fuzzy approaches are capable of handling imprecise information in a fashion superior to computational methods, for which this approach is utilized for diverse tasks.

E. Learning Styles

For Alonso et al. [4], “learning styles are the cognitive, affective, and physiological traits which serve as relatively-stable indicators of how students perceive interactions and respond to their learning environments”. “The learning style describes a learner in terms of the educative conditions that are more conducive to favoring his/her learning. (. . . .) certain educative approximations are more effective than others for him/her”.

The learning style can predict the behavior of the student and, in this way, constitute itself as a good indicator of effective distance learning. The majority of the research that has been conducted in this area is based on learning styles because these are more dynamic, and they yield superior results if they are adequately attended to.

Cognitive traits have to do with the way in which students structure contents, form and utilize concepts, interpret information, solve problems, select representational modalities (e.g., visual, auditory, kinesthetic), etc. Affective traits are linked to the motivations and expectations that influence learning, while physiological traits are related to the biotype and biorhythm of the student [17].

For Joy and Kolb [6], the learning-style types indicate the differences in approaches with regard to learning, based on individual preferences and considering the dialectical combination of those modalities. There are four learning-style modalities, which are: divergent, assimilating, convergent, and accommodating. Divergent learners prefer to make greater use of concrete experiences and reflexive observation. Those of the assimilating type prefer to learn by way of reflexive observation and abstract conceptualization. Those of the convergent type prefer to engage in abstract conceptualization and active experimentation, while those of the accommodating type utilize active and concrete experimentation.

Not all learning-style models are ideal for the development of educative materials within adaptive, hyper-media systems. The approach most used by many adaptive-system researchers is Honey and Mumford’s model [18], because it is centered on how information is perceived and processed. Nonetheless, other models are based on aspects that are not very relevant to development in web environments.

In practice, the majority of learners tend to display the characteristics of one style without either affirming or setting aside the other styles. According to the preferred style, the same content will turn out to be easier (or more difficult) to learn, depending on how it is presented to the learner and how it is dealt with in the classroom.

Optimal learning requires the four stages of Kolb’s wheel, for which it is necessary to enforce discipline in such a way that activities that cover all the stages are guaranteed. With this, on the one hand, the learning of all students will be facilitated, whichever their preferred style may be, and, moveover, the stages will be strengthened for those who are less comfortable with the content. The stages mentioned are: active, reflexive, teoric, and pragmatic.

The review realized in this part of the paper allows the authors to provide the teoric basis that is required for the development of the online model for the identification of students’ learning styles, which is presented in the next section of this paper.

F. Related Work

Research into e-learning systems is currently poised for continued growth due to the fact that there are currently important educative-system demands, which require high degrees of adaptation and intelligence from those systems to be able to provide students with more personalized attention according to their particular requierments. In this part of the paper, the authors attempt to establish the state of the art regarding the research subject.

For Maldonado-Perez [19], “in the learning model based on projects one finds the essence of problematic teaching, thus showing the student the way towards the obtainment of concepts.” The contradictions that arise and the ways leading to their solution contribute to this object of pedagogical influences becoming an active subject. This learning model demands that the professor be a creator as well as a guide, who stimulates the students to learn, discover, and feel satisfied by the accumulated and adequately-operated and utilized knowledge, which can be achieved if teaching-based-on projects is correctly applied.

It is worth pointing out the majority of e-learning tools found on the market and based on web platforms are not naturally compatible with the Project-Based-Collaborative-Learning paradigm (PBCL), for which Abdallah et al. [20] proposes a general meta-model that permits the adapatation of existing platforms to this paradigm, taking as a case study the adaptation of the Moodle platform.

It is worth indicating that the traditional approaches already mentioned were based on the previous identification of the learning style of each participant through the application of surveys and other tests. Nevertheless, there are currently techniques for the automatic identification of the learning style of each individual, such as the proposal of Klansja-Milicevic et al. [21], based on a hybrid, recommendation system that combines clustering and data-mining techniques, and also that of Lo and Shu [10], in which neural networks are utilized for the identification of learning styles starting from the monitoring of the user’s behavior on the platform.

On the other hand, the possibility of integrating diverse types of actors with well-defined roles and their capacity to handle heterogeneous resources has been addressed principally through approaches based on multi-agent systems, where variations have been observed, such as the execution of interactions through the use of the XML standard [22], delivery of contents and distribution of roles in a dynamic and adaptive way [23]. According to Azambuja and Vicari [24], the application of multi-agent architectures allows for improvement in the interactivity of e-learning platforms, such as described in their proposal based on JADE architecture.

There is, moreover, a set of techniques from diverse areas that can be applied to the improvement of the proposed models. One example is implementation of the use of rubrics for the evaluation of complex, imprecise, and subjective areas [25]; the utilization of reasoning-based-on-cases techniques [26], applied to the evaluation and selection of projects according to the characteristics of the audience and learning environments, among other factors.

The vast quantity of projects in the area makes it possible to establish as fact that there are different proposals for the architecture and modelling of adaptive, e-learning systems that utilize diverse artificial-intelligence approaches in order to develop systems with a high degree of personalization and a high capacity to adapt to the learners' personal requirements and expectations.

III. Model System Development

Next, we will describe the different elements that were developed, as well as the procedural steps that were followed in order to build the adaptative e-learning-model system.

A. *Traditional Detection of Learning Styles*

The proposed model utilizes as a reference the classification of learning styles proposed by Honey and Mumford [18]. For the experimental data collection, a survey was applied to a group of 34 pregraduate students from the Professional School of Marketing at Saint Augustine National Univeristy of Arequipa, Peru during the second academic semester of 2017.

The student responses were systematized and tabulated on a digital spreadsheet designed for the purpose with the objective of facilitating the couting of responses, independently from the number of students, considering future tests and the possible scaling of the developed platform. Table I shows the summary of responses from the group of students, where the following facts must be considered:

- The cells highlighted in yellow correspond to students for which one solitary preferred learning style can be identified in a clear manner.
- The cells highlighted in grey correspond to cases in which it is not possible to identify only one learning style, either due to the possible mixed preferences of some students or deficiencies in the application of the aforementioned survey.

These facts, although they will only be pointed out here, are analyzed in greater detail in section IV, where they possess greater relevance with regard to the implementation and tests of the proposed model. However, it should be mentioned that traditional methods (e.g., surveys), although they are the most accepted, are also far from infallible in terms of the identification of learning styles or the other cognitive characteristics of the students.

TABLE I. Data obtained from the traditional method

Student	Activist	Reflector	Theorist	Pragmatist
1	11	16	13	10
2	15	15	13	15
3	13	13	12	9
4	13	16	12	17
5	9	18	17	15
6	13	18	14	14
7	15	15	17	17
8	12	9	8	13
9	16	13	10	17
10	14	17	15	12
11	13	18	13	12
12	11	17	15	13
13	10	17	19	10
14	13	17	14	13
15	13	18	13	16
16	12	17	16	17
17	14	15	16	15
18	13	14	15	18
19	11	16	17	13
20	13	18	11	13
21	13	17	17	13
22	10	18	18	18
23	14	16	15	17
24	17	16	11	17
25	16	15	18	20
26	15	10	7	16
27	12	15	13	15
28	10	18	17	15
29	17	14	11	15
30	10	15	12	17
31	19	16	13	14
32	13	12	10	13
33	19	12	10	15
34	10	11	18	18
Total	3	10	3	8

The distribution of the student's preferences for distinct learning styles is shown in Figure 2, where it can be appreciated that the number of cases in which it is not possible to determine only one preferred learning style is one of the most significant groups, which represents an important fact from the point of view of educative technologies, because it indicates that the students currently adapt better to ditinct types of materials, resources and contents, as well as learning environments.

Finally, it should be indicated that the data obtained through the traditional method (surveys) will be utilized in order to compare them with the results obtained from the neural network as a way to validate them, for which the data have been divided into two sets of equal cardinality (equal number of registers), procuring that both sets retain similarity regarding the percentatge of learning styles identified in each group. These sets will be utilized as training and test sets for the implementation of the neural network.

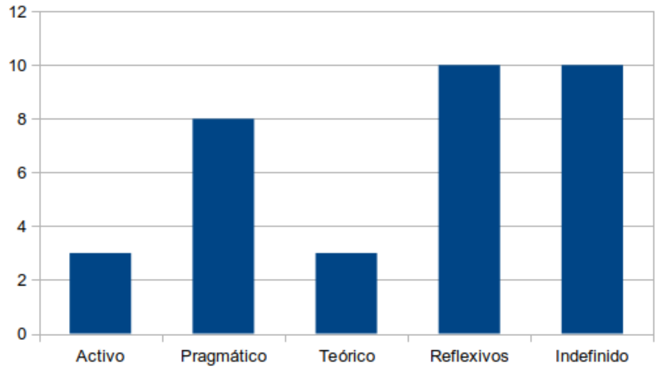


Fig. 2. Resuls of the test data obtained [27]

B. Establishment of Resource Categories and their Relation to Learning Styles

In order to classify user interactions, a list of 20 resource categories was defined, considering the types of resources existing in the Moodle platform, which was utilized for the implementation of the platform due to the fact that it is an opensource platform, subsequently relating each resource category with each one of the learning-style categories, utilizing general sets defined for that purpose (Table II).

TABLE II. RESOURCE CATEGORIES AND THEIR RELATION TO LEARNING STYLES

	Resource Type	Activist	Reflector	Theorist	Pragmatist
1	Content (Textual)	Low	High	High	Medium
2	Content (Mixed)	Medium	High	High	High
3	Content (Multimedia)	High	Medium	Medium	High
4	Content (Simulation)	High	High	Medium	High
5	Content (Url's)	Low	High	High	Low
6	Case Study (Textual)	Medium	High	High	Medium
7	Case Study (Multimedia)	High	Medium	Medium	High
8	Examples (Textual)	Medium	High	High	High
9	Examples (Multimedia)	High	High	Medium	High
10	Examples (Url's)	Medium	High	High	Low
11	Glossary (Reading)	Low	High	High	Medium
12	Glossary (Writing)	Null	High	High	Low
13	Wiki (Reading)	Medium	High	Medium	Medium
14	Wiki (Writing)	Medium	Medium	Low	Low
15	Forum (Reading)	Medium	Medium	Medium	Medium
16	Forum (Writing)	Medium	Low	Medium	Medium
17	Chat (Reading)	High	Low	Medium	Medium
18	Chat (Writing)	High	Low	Low	Medium
19	Self-assessments	Medium	High	High	Low
20	Conceptual maps	Nulo	High	High	Medium

It is important to mention that the source code of the Moodle platform was modified with the purpose of adding a log of user activity and interaction, which would store the resource selections realized by the user (by way of 'clicks') according to the type of category, which is to say a log of user behavior in the platform.

In this manner, the proposed approach permitted the attainment of the inputs (i.e., user interactions) and expected outputs of the model (i.e., learning styles identified beforehand), manually dividing the obtained data into two similar sets, making it possible for them to have the same cardinality (i.e., number of individuals) and also for them to be equally representative of the distinct learning styles in the same proportion in which they are found in the test data obtained with the traditional method.

C. Proposal and Implementation of the Neuro-Fuzzy System

Towards the identification of the learning style, a back-propagation neural network model was proposed, composed of an input layer, a hidden layer, and an output layer, such as shown in Figure 3. Upon implementation of the neural network, for the activation of the neurons, the sigmoidal function was chosen due to the fact that it permits the modeling of temporal progressions, which go from beginning levels—in which the contents are more or less generic and do not require sophisticated knowledge on the part of the users—to advanced levels, which with the passage of time, as content personalization is refined, permit the attainment of the required knowledge for a more precise identification of the user type (Figure 4).

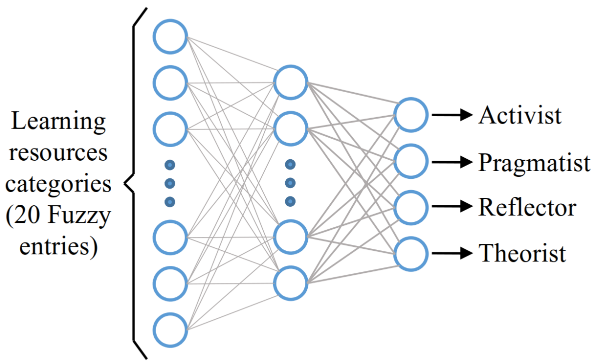


Fig. 3. Neural Network architecture

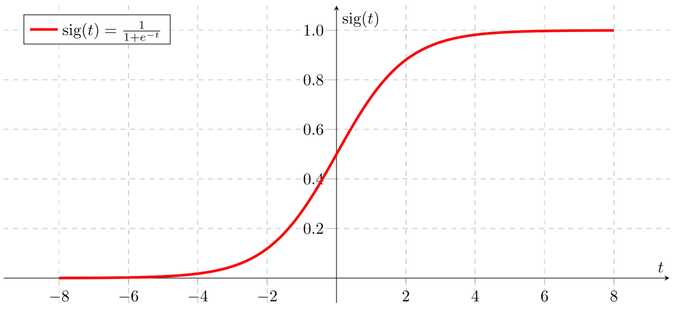


Fig. 4. Sigmoidal function

The input neurons represent each one of the platform resource categories previously defined (Figure 3), while the input values represent user preference for each of these categories. In this manner, 20 input neurons have been identified.

In the 'neural network', the hidden layer increases the processing capacity, and the number of neurons in the hidden layer directly affects the capacity of the 'neural network' for learning. In the proposed case, during the initial experimentation phase, before pre-processing the inputs (a process which will be described later), tests were conducted with distinct numbers of neurons in the hidden layer. Moreover, tests were also conducted with two intermediate layers without achieving favorable results. In the final design of the 'neural network', a hidden layer with an equal number of neurons as the input layer was utilized, which was able to achieve the best results in an acceptable period of time.

The output layer was implemented with four output neurons that represent the four learning styles proposed by Honey [18], while the output values indicate user compatibility with said learning style.

During the initial runs of the model, some inconveniencies were found in the performance of the network, which, according to the analysis conducted, arose due to two causes:

- The level of 'noise' present in the input sets – For example, the user identified as number 23 in Table I —whose number of responses in relation to the learning styles is shown in Table III, according to the method proposed by Honey [18]—corresponds to the 'pragmatic' learning style. Nevertheless, the values are very close in each of the learning styles, complicating the establishment of a clear differentiation. This can be due to, for example, the selection of some indispensable contents for the achievement of objectives or, rather, the recommendations of other users which were finally selected by the user but do not necessarily represent the predominate learning style for him and, consequently, should not have a noticeable influence in the model.

TABLE III. Illustration of the Noise Level

Activist	Reflector	Theorist	Pragmatist
14	16	15	17

- The input sets for which it is impossible to define an output –For example, as in the case of the user identified as number 2 in Table I, whose response log in relation to the learning styles is shown in Table IV, where clearly, according to the scheme proposed by Honey [18], it is not possible to identify the learning style.

TABLE IV. Specific Input Set

Activist	Reflector	Theorist	Pragmatist
15	15	13	15

In order to focus on these inconveniences, especially the first one, the authors opt for conducting a pre-processing of inputs through the definition of the second group of general sets, which are oriented towards achieving a better categorization of user preferences for a certain resource category, according to the percentage and relevance of their interactions in each resource category, such as shown in Figure 5.

This decision is based on the premise proposed by Palomino [15], who points out that determining a student's learning style becomes a problem of a general nature because evaluative situations and characteristics must be taken into consideration, albeit with a certain level of imprecision. These require a treatment appropriate to the nature of the problem, such that corresponds to the proposed case. Generally, the application of the concept of 'fuzzy logic' permits the modeling of situations which do not have precise limits, which makes for a much more realistic model, especially when having to do with cognitive and subjective aspects.

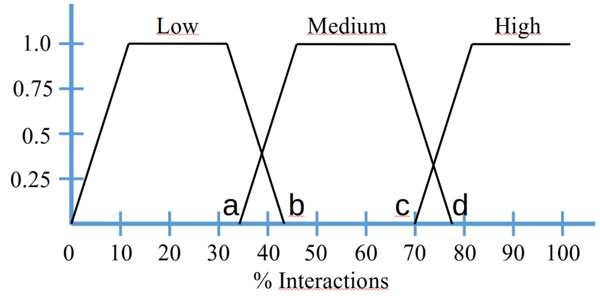


Fig. 5. Fuzzy sets for user preference for a certain resource category [27]

To this mechanism of 'fuzzification' —starting from the output data of the platform logs, previous to being considered as input data for the 'neural network'—the name 'preprocessing stage' of inputs is given. For the definition of the 'fuzzy sets', a trapezoidal function was utilized, defined as shown in Equation 1.

$$\mu B(x) = trapezoidal(x; a, b, c, d) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c}, & c \leq x \leq d \\ 0, & d \leq x \end{cases} \tag{1}$$

Subsequent to implementing the specified criteria for this stage, the results obtained by the model are notably improved, and a superior design of the 'neural network' could be determined, with only one hidden layer remaining, as was explained previously.

It should be noted that the developed model, under an opencode platform, will allow for, in a subsequent stage, the development of a resource selection and adaptation mechanism within the platform, based on the learning style of each individual. What is more, this mechanism will also permit the validation and refinement of knowledge regarding user preferences, as well as the creation of more sophisticated user profiles.

IV. Results and Discussion

For the experimentation and runs with the proposed model, the students were asked to perform some activities as part of a university course throughout the semester. It is important to note that the identification must be made throughout a period of platform-utilization time, given that the data analyzed in just one session might be seen as influenced by the time available for the identification of the style, the emotional state of that particular moment, problems in the environment, etc., making it possible for errors in the perception and identification of the learning style to arise.

For example, Figure 6, shows the identification of the learning styles of four students throughout each week of the

20-week duration of the semester, where it can be appreciated that, for example, for the 'case-1' student, the identification realized in weeks 3 and 20 might indicate that the student fits into second category. However, when the general panorama is observed, it is clear that this student instead fits into the first category.

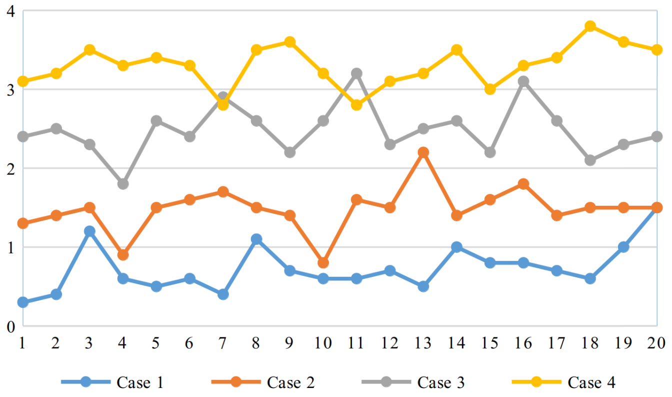


Fig. 6. Variation of user identifications through several learning sessions [27]

This phenomenon is relatively normal, given the proximity among some learning styles and the mixed preferences of some students, which also can be appreciated in Table I and Figure 6, where, for example, the set of students with mixed preferences is the second most representative, in some cases the task of identifying just one learning style being very complex.

In this sense, the most feasible option would be to identify the learning styles during some introductory course —such as 'study strategies' or previous activities before beginning to perform a content adaptation—and later validate and refine this identification in the subsequent activities or courses.

Finally, the results obtained by the neural network demonstrated a 76.5% coincidence with those obtained through the traditional method, which is to say that the learning styles of 26 of the 34 students were obtained correctly. For this reason, it can be said that the proposed model reached 76.5% efficiency with respect to the manual method proposed by Honey [18].

Table 5 shows a reasonable comparison of the different approaches for automatic, online detection of learning styles, considering the classification of learning styles used in each approach. It should be noted that in the case of approaches where efficiency was calculated for each of the learning dimensions or styles, the average of these has been considered in order to facilitate comparison with other more general approaches, such as the one proposed in the present research.

TABLE V. COMPARISON OF EVALUATED MODELS

Evaluated models	Learning Styles	Efficiency
Bayesian networks	Felder & Silverman	66%
NBTree y CRB	Felder & Silverman	67.5%
Genetic algorithms and K-NN	Ad-hoc	96%
Monitoring of interactions	Felder & Silverman	79.6%
Learning objects and time estimation	Felder & Silverman	69.6%
Neural networks and navigation maps	Vincent & Ross	90%
Stochastic models	Felder & Silverman	70%
NeuroFuzzy model (Proposal)	Honey & Mumford	77.1%

V. Conclusion

It has been proposed the design of an adaptive e-learning system, still in the process of implementation, which considers the characteristics of the users, in particular their learning styles and preferences, to focus on the possibility of personalizing the ways of using the objects of learning, in a different way to traditional systems, designed for generic users.

The authors also proposed a System model for the classification of user interactions, within an adaptive e-learning System platform, and whose analysis through a mechanism based on a backpropagation neural network and fuzzy logic, it allows the automatic online identification of learning styles, in a transparent way for the user, which is very useful within the platform of the adaptive e-learning system.

The determination of a student's learning style is a problem of a general nature because evaluative situations and characteristics must be considered (with a certain level of imprecision to be expected), requiring an appropriate treatment to the nature of this problem.

The identification of learning styles can not be based on a single session or user access, which can lead to errors of interpretation, but rather must be made over several sessions in order to achieve an adequate accuracy in the identification.

It is important to propose and develop the system model, which using Case Based Reasoning (CBR), establishes the correspondence between the preferences of students of different learning styles, detected with the proposed neuro-fuzzy system, with the problems formulated with the project-based learning approach, which contains the learning objects with the greatest significance for them.

It is important to continue contributing to the area with research into learning and thinking styles, whose theoretical bases must be taken seriously as important elements in adaptive, e-learning systems.

Acknowledgment

The authors would like to thanks to Saint Augustin National University (UNSA) for supporting this research.

References

[1] R. Dittus and C. Vásquez, *Abriendo la autopoiesis: Implicancias para el estudio de la comunicación organizacional*. Cinta moebio, vol. 56, pp. 136-146. doi: 10.4067/S0717-554X2016000200002, 2016.

[2] H. Maturana and F. Varela, *El árbol del conocimiento. Las bases biológicas del entendimiento humano*. Buenos Aires: LUMEN Editorial Universitaria, 208 p., ISBN: 987-00-0358-3, 2003.

[3] M. Alshammari and R. Anane and R. Hendley, *Adaptivity in e-learning systems*. In 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems. IEEE, 2014.

[4] C. Alonso and D. Gallego and P. Honey, *Los estilos de aprendizaje. Procedimientos de diagnóstico y mejora*. Bilbao: Mensajero, 1994.

[5] T. Wang and K. Wang and Y. Huang, *Using a style-based ant colony system for adaptive learning*. ESWA. 2449-2464, 2008.

[6] S. Joy and D. Kolb, *Are There Cultural Differences in Learning Style?*. Available in https://weatherhead.case.edu/departments/organizational-behavior/workingPapers/WP-07-01.pdf, 2007.

[7] M. Fadzil and L. Andreasen and M. Buhl and T. Amina, *The concept of ubiquity and technology in lifelong learning*. ASEM Education and Research Hub for Lifelong Learning, pp. 1-18, ISBN: 978-89-20-01127-6 93370, 2012.

[8] E. Morales, *Gestión del conocimiento en sistemas e-learning, basado en objetos de aprendizaje, cualitativa y pedagógicamente definidos*. Ph.D. dissertation, Universidad de Salamanca, 2007.

[9] M. Abdullah, *The Impact of Learning Styles on Learner's Performance in E-Learning Environment*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 9, 2015.

[10] J. Lo and P. Shu, *Identi fication of learning styles online by observing learners browsing behaviour through a neural network*. British Journal of Educational Technology, vol. 36, no. 1, pp. 43–55, jan 2005.

[11] L. Alfaro and C. Rivera and J. Luna-Urquizo and E. Castañeda and F. Fialho, *Fuzzy neural System Model for Online Learning Styles Identification, as an Adaptive Hybrid ELearning System Architecture Component*. 2018 Latin American and Caribbean Consortium of Engineering Institutions Conference, available in: http://laccei.org/LACCEI2018-Lima/full_papers/FP259.pdf., 2018.

[12] R. Lanhellas, *Redes neurais artificias. Algoritmo de backpropagation*. Available in: https://www.devmedia.com.br/redes-neurais-artificiais-algoritmo-backpropagation/28559, 2013.

[13] R. Timothy, *Fuzzy Logic With Engineering Aplications*. Wesley, 2010.

[14] A. Ozdemir and A. Alaybeyoglu and N. Mulayim and K. Filiz, *Performance evaluation of learning styles based on fuzzy logic inference system*. Computer Applications in Engineering Education, Vol 24, No. 6, pp. 853-865. DOI: 10.1002/cae.21754, 2016.

[15] M. Palomino and M. Strefezza and L. Contreras, *Sistema difuso para la detección automática de estilos de aprendizaje en ambientes de formación web*. Ciencia, Docencia y Tecnología, vol. 27, no. 52, pp. 269-294. ISSN: 1851-1716. 2016.

[16] R. Stathacopoulou and G. Magoulas and M. Grigoriadou and M. Samarakou, *Neuro-fuzzy knowledge processing in intelligent learning environments for improved student diagnosis*. Information Sciences, 170(2-4), 273-307, 2005.

[17] B. Maraza and L. Alfaro and O. Alejandro and C. Vivanco and Y. Jimenez and S. Choquehuayta and J. Herrera and N. Caytuiro, *Biofeedback of states of anxiety through automated detection processes using different technologies*. International Journal of Engineering & Technology. 7 (3) 1609-1614, available in: www.sciencepubco.com/index.php/IJET, DOI: 10.14419/ijet.v7i3.15041, 2018.

[18] P. Honey and A. Mumford, *The Manual of Learning Styles*. Maidenhead, Berkshire: P. Honey, Ardingly House, 1986.

[19] M. Maldonado-Pérez, *Aprendizaje basado en proyectos colaborativos. Una experiencia en educación superior*. Revista de Educación Laurus, vol. 14, no. 28, pp. 158-180, 2008.

[20] F. Abdallah and C. Toffolon and B. Warin, *Models transformation to implement a project-based collaborative learning (PBCL) scenario: Moodle case study*. 2008 Eighth IEEE International Conference on Advanced Learning Technologies, ICALT08, pp. 639-643, 2008.

[21] A. Klasnja-Milicevic and B. Vesin and M. Ivanovic and Z. Budimac, *E-learning personalization based on hybrid recommendation strategy and learning style identification*. Computers and Education, vol. 56, no. 3, pp. 885-899, 2011.

[22] A. Garro and L. Palopoli, *An xml multi-agent system for e-learning and skill management*. Agent Technologies Infrastructures, Tools and Applications for E-Services, Springer Lecture Notes in Computer Science, vol. 2592, pp. 283-294., 2003.

[23] M. Hameed and N. Akhtar and M. Missen, *Role based multi-agent system for e-learning (MASEL)*. International Journal of Advanced Computer Science and applications, vol. 7, no. 3, pp. 194-200, 2016.

[24] R. Azambuja and R. Vicari, *Improving interactivity in e-learning systems with multi-agent architecture*. International Conference on Adaptative Hypermedia and adaptative web-based systems, vol. 2347, no. 1, pp. 466-471, 2002.

[25] H. Googrich, *What do we mean by results: Using rubrics to promote thinking and learning*. Educational Leadership, vol. 57, pp. 13-18, 2000.

[26] J. Kolodner, *Case-Based Reasoning: Foundational Issues*. California, USA: Morgan Kaufmann Publishers, 1994.

[27] L. Alfaro and C. Rivera and J. Luna-Urquizo and E. Castañeda and F. Fialho, *Online learning styles identification model, based on the analysis of user interactions within an e-learning platform, using neural networks and fuzzy logic*. International Journal of Engineering and Technology (UAE). Vol 7 No 4.17 (2018): Special Issue 17 - Published: 2018-10-27.

An Optimal Control Load Demand Sharing Strategy for Multi-Feeders in Islanded Microgrid

Muhammad Zahid Khan[1], Muhammad Mansoor Khan[2], Xu Xiangming[3], Umar Khalid[4], Muhammad Ahmed Usman Rasool[5]

[1,2,4,5]School of Electronic, Information and Electrical, Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China
[3]State Grid Zhenjiang Power Supply Company, Jiangsu, China

***Abstract*—For the operation of autonomous microgrid (MG), an essential task is to meet the load demand sharing using multiple distributed generation (DG) units. The conventional droop control methods and its numerous variations have been developed in the literature in order to realize proportional power sharing amongst such multiple DG units. However, the conventional droop control strategies are subjected to power sharing error because of non-trivial feeder impedances of medium-voltage MGs. Further, complex MGs configurations (mesh or looped networks) usually make to reactive power sharing and system voltage regulation more challenging. This paper presents an optimal control strategy in order to perform the proportional power sharing and voltage regulation for multiple feeders in islanded AC MGs. The case study simulation for optimizing the power sharing and voltage regulations in proposed control strategy has been verified through using MATLAB/Simulink systems.**

Keywords—Optimal control; power sharing; voltage regulation; MG

I. Introduction

The application of distributed power generation such as wind turbine, photovoltaics' and fuel cell has been experienced a fast development in the past decades [1] [2]. DG units as compare to conventional centralized power generation, provides more clean and renewable power close to consumer's end [3][4]. Therefore, it can ease the stress of numerous traditional transmission and distribution framework[5]. Power electronics converters are interfaced between DG units and the grid, are the vital elements of the MGs [6] , and perform the flexibility of islanded or grid connected operation.

On the other side, high infiltration of power electronics based DG units presents couple of issues, such as voltage deviations, frequency and power flow variations [7]. In order to sort out these aforementioned problems, the idea of MG has been emerged, which is based on the control of multiple DG units. As compare to a solitary DG, MG can accomplish predominant power management within its distribution network [3]. The MG can operate either in grid-connected mode or islanding mode. In grid connected mode, the MG is connected with the main grid at the point of common coupling (PCC) and according to dispatched references every DG unit provides proper active and reactive power. The most used control strategies are reported in [8] for grid connected inverters.

In the operation of islanded MG mode, the load demand should be appropriately shared by DG units according to their respective ratings and availability of power from either their respective prime movers or energy sources [3]. Communication based power sharing control strategies include master/slave control, concentrated control, and distributed control[8], while control strategies without communication are usually based on the droop concept, which can be classified into four main categories: 1) conventional and variants of the droop control[9]; 2) construct and compensate based strategies[10]; 3) the hybrid droop/signal injection strategy; 4) In [11] virtual framework structure-based method is developed.

Traditional frequency and voltage magnitude droop control approaches are adopted for interfacing inverters in a decentralized mode to attain power sharing and voltage regulation [12]. However, a little while back is observed that conventional droop control strategy in low voltage MG has led to have few power control stability issues, as the DG feeders have largely resistive (high R/X ration) behavior [7]. It can also be observed that active power at the steady state is usually proportionally shared among DG units, while the reactive power sharing deteriorates due to mismatched of DG units output and feeder impedances [7]. The impedances of transmission line be asymmetrical due to distinctive separations amongst DG units while the design of LCL filters are depends on varying system conditions and design considerations which leads to dissimilar DG unit's output filters impedances [13] [7]. In addition, the presence of local loads and the complex network MG configuration usually further increase load sharing performance.

To resolve the power control problems, few enhanced droop control strategies [11] and [14] have been reported in previous literature. In [15], an accurate power sharing control approach has been reported to restore the load point voltage with the decreased voltage deviation. Author proposed an enhanced reactive sharing strategy in[5]. Aforementioned, these two strategies are, however, attained at cost of inverter terminal voltage deviation. Furthermore, the droop strategies based on virtual impedances methods are seen as a promising strategy to handle power sharing issue. The virtual frequency-voltage frame and virtual power idea were reported in [14] [14]and [11], that enhance the stable operation of the MG system. However, these strategies cannot subdue the reactive power sharing errors in the meantime. In addition, appropriate power sharing among inverter and electric machine is subject to challenging in these strategies, when small synchronous generators are included into the MG. Although the author addressed the power sharing issue in[16], but the respective steady state voltage distortion deteriorates the overall power quality of MG. Author proposed an “Q-V dot droop” strategy in [13], but it is noticed that reactive power sharing enhancement is not evident when the local loads are incorporated. Author in [17] used additional PCC voltage measurement in order to mitigate the error of reactive power sharing. In [18], an enhanced virtual impedances control method has been reported for a DG unit, that is able to compensate for the unequal feeder impedances. Although, the power sharing can be enhanced by virtual impedance based droop control strategies but voltage droop and virtual impedance deteriorates the inverter terminal voltage quality [19].

In order to reduce the tradeoff among reactive power sharing and bus voltage deviations in multi feeders a recent control strategy is developed in [20] where a Kalman filter-based state estimator used which required high bandwidth date rate. In addition, feeders can be located at considerable distance from each other, therefore it increases complexity and reduce the reliability and flexibility of MG operation. Therefore, this paper proposed an optimal control load demand sharing strategy for multi-feeders which is directly based on load estimation and optimal regulator as shown in Fig. 3. The salient contribution of this work can summarized as follows:1) The load is estimated at respective feeders which reduces the bandwidth data requirements; 2) The proposed optimal control strategy achieved task of proportional power sharing and system voltage regulation for multiple buses simultaneously.

The rest of this paper is organized as follows. In Section II, the operation of MG is discussed. The operating principal of a proposed control approach is given in section III. The simulations results are presented in section IV and finally section V concludes the paper.

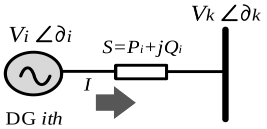


Fig. 1. An i^{th} Inverter Connected with k^{th} AC Bus.

MGs are consisting on considerable number of DG units and connected load as shown in Fig. 2. Every DG unit is connected to the MG with an interfaced inverter where DG inverters connected to the PCC via their corresponding feeders. The statues of main grid and MG are controller by the MG central controller. Depending on operations requirements, the main grid can be connected or disconnected from the MG by switching the state of static transfer switch STS at the common bus coupling point. In the grid-connected operation mode, the active and reactive reference usually are allocated by central controller and in order to track the power the conventional droop control strategy can be used. PI regulation for the voltage magnitude control used to mitigate the steady stated reactive power tracking errors. So, during grid connected mode the power sharing is not concern. When micro grid is operating in islanded mode, the load demand of MG should be properly shared by DG units. In this mode of operation, the DG units can operate using conventional power frequency droop control strategies as

$$\omega_i = \omega_i * - D_{Pi} . P_i \tag{1}$$

$$V_i = V_i * - D_{Q_i} . Q_i \tag{2}$$

Where, $V_i *, \omega_i, D_P$ and D_Q are the nominal voltage magnitude, nominal frequency, real and reactive power slops, respectably for i^{th} DG unit.

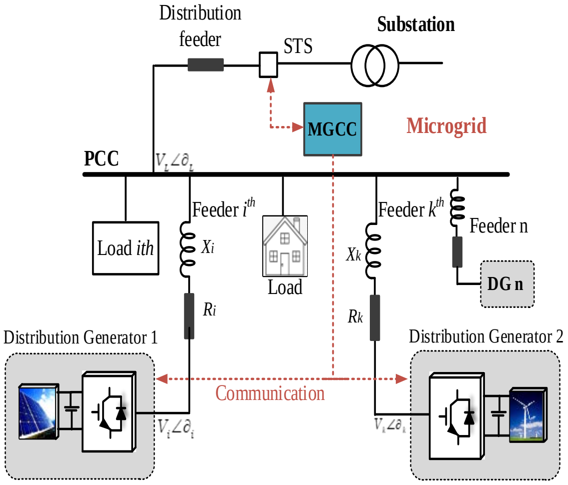


Fig. 2. Illustration of the Microgrid Configuration.Operation of MG.

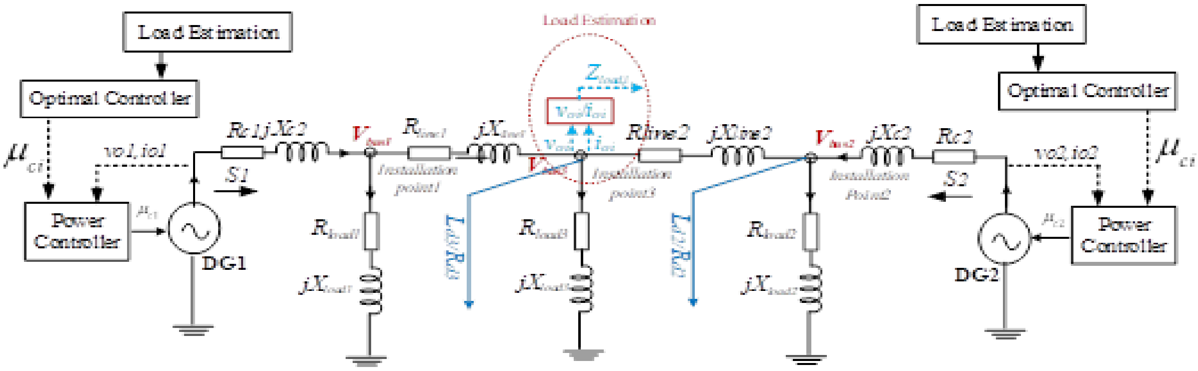


Fig. 3. Block Diagram of Proposed Optimal Control Strategy.

II. Proposed Optimal Control Strategy

A radial type feeder is used in proposed optimal control strategy as illustrated in Fig. 2. All three buses V_{bus1}, V_{bus2}, and V_{bus3} are fed through two DG units DG1 and DG2 interfaced using three phase, three wire power electronics inverters connecting through feeder impedances with three linear loads R_{load1}, R_{load2} and R_{load3}.This proposed strategy is composed on load estimation and optimal steady state estimator regulator. Load estimation strategy used in order to estimate specific feeder's impedances which have advantages that it reduces the data bandwidth requirements. Based on these load estimation values $Z_{load,i,}$ the optimal regulators are responsible to compute the optimal control command which is a cost function and send two optimal control commands u_{c1} and u_{c2} to power controllers in order to realize proportional power sharing and voltage restoration. Ld2/Rd2 and Ld3/Rd3 are disturbance load which are exerted to examine the effectiveness of this proposed strategy for both inductive and resistive MGs. Different cases of disturbance load addition are further discussed in detail in section 4.

A. Mathematical Model

In order to explain operation of a MG, a simplified circuit in Fig. 1, is illustrated where two *i*th and *k*th DG units are parallel connected. The complex power drawn towards the *k*th ac bus can be expressed as.

$$S_{ik} = P_{ik} + jQ_{ik} \tag{3}$$

Where, active power P_i and reactive power Q_i are introduced at every existing node by DG converters. If power inverters are supposed to be an ideal controllable voltage source which is connected to line impedances via mains, then the movement of real and reactive powers in transmission line impedances can be expressed as.

$$P_i = \frac{V_i}{R_i^2 + x_i^2}.[R_i V_i - R_i V_k \cos\partial_{ik} + X_i V_k \sin\partial_{ik}] \tag{4}$$

$$Q_i = \frac{V_i}{R_i^2 + x_i^2}.[-R_i V_k \sin\partial_{ik} + X_i V_i - X_i V_k \cos\partial_{ik}] \tag{5}$$

Where, i=1,2 represents the two branches in circuit. V_i is magnitude of inverter output and V_k represents the PCC bus voltage, while P_i and Q_i are the active and reactive powers flowing from ith inverter terminal to kth common bus voltage, illustrates the difference amongst the power angle phase of the output impedance.

In higher voltage HV and medium MV network the inductive elements are typically higher then resistive as shown in table 1 [11], however, HV and MV networks have inductive behavior, therefore we can neglect the resistive part. As power angle ∂ is small in such type of lines so we can assume that and the possible power flow in network can be written as

$$P_{i,R_x=0} \approx \frac{V_i V_k}{x_i}[\ \sin\partial_{ik}] \tag{6}$$

$$Q_{i,R_x=0} \approx \frac{V_i^2 V_k - V_i V_k \cos\partial_{ik}}{x_i} \tag{7}$$

Where,

$$\partial_i - \partial_k \propto P_i \tag{8}$$

$$V_i - V_k \propto Q_i \tag{9}$$

According to expression (8) and (9), it is obvious that the real power P_i drawn towards *k*th node predominately depends on power angle while reactive power Q_i injected by each DG inverter mostly controlled by voltage difference V_i -V_k of ith and kth ac bus.

B. Load Estimation and Optimal Regulator Principal

Proposed optimal control strategy used to estimate the load at specified feeder follows variable frequency local voltage based park transformation as expressed in (10) and elaborated in Fig. 4(a). This strategy, firstly sensed the voltage $V_{bus,i}$ and current $I_{bus,i}$ at local node of ith feeder, later both voltage and current signals are converted into abc—dq0, where rotating frame is aligned 90 degrees behind A axis. dq0 values converted from real-imaginary inputs to a complex valued output signal, where load impedances $Z_{ri—c}$, I is achieved by ratio of $V_i \angle \partial_i$ and $i_i \angle \partial_i$ which is also a complex valued signal as expressed in (11). Since, optimized cost function impedances which is expressed in (13), input signals should be real-imaginary valued signal, so a complex to real-imaginary block is used as shown in Fig. 4(a), and elaborated in (12), which converts the complex load impedances signals to real-imaginary valued impedances signals. Aforementioned, these impedances are estimated at respective feeders which reduce the bandwidth data requirements. Later these impedances signals are sent to the proposed optimal controller achieved task of proportional power sharing and voltage restoration for multiple buses simultaneously.

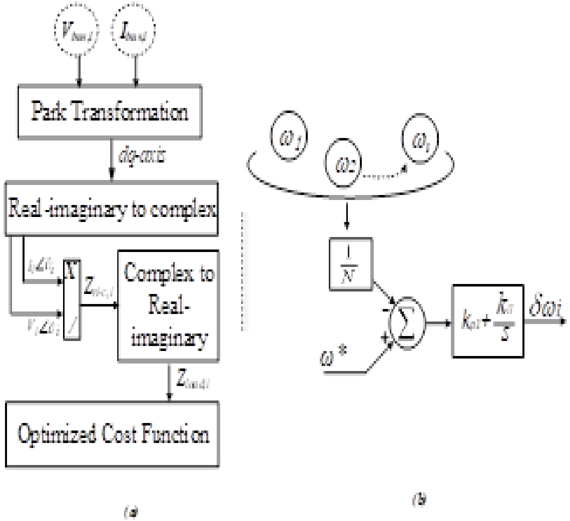


Fig. 4. (a) Load Estimation of ith Feeder (b) Frequency Regulation.

$$\left.\begin{aligned}&U_s=(u_d+j.u_q)=(u_\alpha+j.u_\beta).e^{-j\left(\omega t-\frac{\pi}{2}\right)}\\&\begin{bmatrix}u_d\\u_q\\u_0\end{bmatrix}=\begin{bmatrix}\sin(\omega t)&-\cos(\omega t)&0\\\cos(\omega t)&\sin(\omega t)&0\\0&0&1\end{bmatrix}\begin{bmatrix}u_a\\u_b\\u_c\end{bmatrix}\end{aligned}\right\}\tag{10}$$

$$\left.\begin{aligned}&V_{mag}=\sqrt{(V_{Real})^2-(V_{\mathrm{Im}ag})^2};\ \partial_i=\tan^{-1}V_{Real}/V_{\mathrm{Im}ag}\\&i_{mag}=\sqrt{(i_{Real})^2-(i_{\mathrm{Im}ag})^2}\quad;\ \partial_i=\tan^{-1}i_{Real}/i_{\mathrm{Im}ag}\end{aligned}\right\}\tag{11}$$

$$\left.\begin{aligned}&V_{Real}=V_i\cos\angle\partial_i;\quad V_{\mathrm{Im}ag}=V_i\sin\angle\partial_i\\&i_{Real}=i_i\cos\angle\partial_i;\qquad i_{\mathrm{Im}ag}=i_i\sin\angle\partial_i\end{aligned}\right\}\tag{12}$$

Once the load impedances founded, the optimal regulator is a following critical step. The proposed optimal regulators are an optimized cost function which is presented in order to compute control commands according to the estimated network impedances. Different type of optimization techniques is being used such as linear, quadratic and higher order optimization strategies in order to minimize or maximize cost function of the system. Linear optimization has wide area of application and it is easy regarding solvability but the limitation of linear optimization is that it works only with the variables that are linear as well as problem formulation is freaky. Higher order cost functions are convenient but solution is inconvenient. In this paper, quadratic optimization based cost function is used which is easy regarding solvability, problem formulation and solution is also convenient. Some cost functions have constrained reactive powers which can be supposed to equal i.e $Q_1=Q_2$ in order to vary the tradeoff among real power and inverter terminal voltages, but usually the reactive power requirement is not so stringent, so in this case, this paper used constrains real powers which are supposed to be equal $P_1=P_2$ and the tradeoff among bus voltage and reactive powers is found through minimization. Moreover, the desired control commands are acquired by computing the optimization cost function that minimize the reactive power sharing error ΔQ_i and voltage error ΔV_i at specific bus, which can be expressed as

$$\begin{aligned}&\min\ J=\left(\sum_{j=1}^{n_b}(\omega_{bj}(V_{bus_j}-V_{jref}))^2\right)+(\omega_Q(Q_i-Q_{i+1}))^2\\&constrainsts\quad P_i-P_{i+1}=0\end{aligned}\tag{13}$$

In the cost function ω_{bj} and ω_Q are the weights for network voltages at specific bus and reactive power error, respectively. V_{jref} (j=1,2,3) are set at nominal voltage 300V, while $n_b=3$ is the number of buses that has been chosen. Nevertheless, the limitation with is proposed strategy is that, it requires the accurate grid impedances of every load node which aggravates the computationally complexity with increase in feeder numbers. By considering that limitations, in future the whole grid network impedance can approximated into one load node which will reduce the computationally burden over system.

However, moving forward, after computing every optimal controller sends desired control commands to their respective power controllers and operates with the control commands until upcoming sampling update. Frequencally, the optimal regulator obtains new estimated impedances due to measurements feedback and accordingly revises its original control plan. Then, the voltage control commands are send to compensate for voltage and reactive power sharing deviations.

C. *Power Flow Control*

Proposed control strategy is illustrated in Fig. 5. The output frequency and voltages of Inverter Bridge connected with dc power source are adjusted by power, voltage and current controllers. Every individual DG unit is formulated in its d-q frame where they depend on their both individual angular frequency and angle. Every inverter interfaced with DG units are transferred to the d-q frame by using following transformation equations as

$$\begin{bmatrix}f_D\\f_Q\end{bmatrix}=\begin{bmatrix}\cos(\partial i)&-\sin(\partial i)\\\sin(\partial i)&\cos(\partial i)\end{bmatrix}\begin{bmatrix}f_d\\f_q\end{bmatrix}\tag{14}$$

The power controller block for proposed strategy is illustrated in Fig. 5, adopts P-ω droop as illustrated in (15) and provides the angle of i^{th} DG units which can be express as

$$\omega_i=\omega_i{}^*-m_{Pi}P_i\tag{15}$$

$$\partial_i=\int(\omega_i-\delta\omega_i)dt\tag{16}$$

Where ω^*_i ¬and m_{Pi} are the reference frequency and the P-ω droop coefficient, respectively. Noticeably, the voltage references u_{ci} as shown in Fig. 5, is used instead of Q-V droop, which can be calculated by optimization based cost function as discussed aforementioned in section B. However, the average real power is acquired by instantaneous power passing low-pass filters as expressed below

$$P_i=\frac{\omega_c}{s+\omega_c}p_i\tag{17}$$

Where, ω_c is cutoff frequents of low pass filters Instantaneous active power pi can be represented in d-q frame as

$$p_i=V_{odi}.i_{odi}+V_{oqi}.i_{oqi}\tag{18}$$

Here, v_{odqi} and i_{odqi} are inverter terminal voltage and current, respectively on d-q frame.

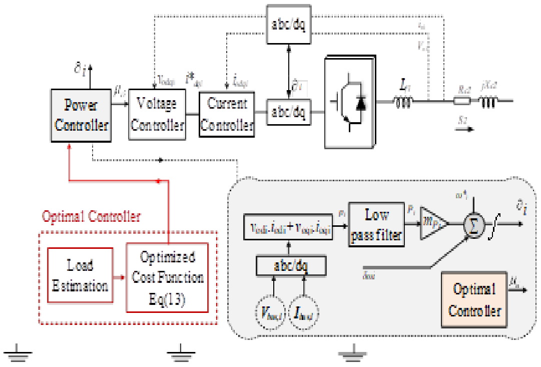


Fig. 5. Proposed Optimal Control Strategy.

D. Frequncy Regulation

The secondary controllers for MGs are based upon frequency restoration. Since, the real power highly influences the frequency of generator-dominated grids. This feature is an edge, since frequency is controllable variable that gives the information regarding to generation or consumption balance of the grid. The frequency regulation strategy which is implemented in order to restore the frequency of system is illustrated in Fig. 4(b) which regulates the frequency deviations of i^{th} DGs units to its nominal value. Frequency restoration strategy can be expressed by ω* and ω_{avg} are the nominal reference frequency and measure system frequency that is being sensed by each node of DGs unit's interface inverters in the neighborhood of the node i being considered. Frequency correction is send to frequency reference of the i^{th} inverters node, while K_{pf} and K_i are the proportional and integral gains, respectively, for controllers.

$$\left.\begin{aligned}
\overline{\omega_{avg}} &= \frac{\sum_{k=1}^{N}\omega_k}{N}\\
\omega_i &= (\omega^* - \overline{\omega_{avg}})\\
\delta\omega_i &= k_{pf}\omega_i + k_{if}\int \omega_i dt
\end{aligned}\right\} \qquad (19)$$

E. Mechanism of Reactive Power Sharing

The output of the optimizer is set of voltage phasors and its implementations require the information of voltage phasers of all inverters participating the MG. A direct reconstruction of such information needs very fast and reliable communication and computation infrastructure. However, in presented approach the phases have been modified through frequency/real power droop control and only voltage magnitudes are updated in accordance with the optimizer output. This combination will render accelerates active power sharing which is also considered as constraint in cost function. Consequently, the correct phase angle will automatically be adjusted by system.

III. Result and Discussion

In order to verify the effectiveness of proposed control approach, the simulations have been carried out on MATLAB/Simulink for a three phase 50-Hz islanded MG. As illustrated in Fig. 6 the simulated MG is composed on two DG1 and DG2 unites connected in parallel with three linear loads via feeder impedances. The circuit and control parameters are shown in Fig. 6 and Table 1, respectively. Simulation verifications are composed on two cases. The case 1 and case 2 investigates the effectiveness of proposed control strategy on different disturbance locations in inductive and resistive MG, respectively.

TABLE I. Typical Line Impedances

Type of Lines	R(Ω /km)	X(Ω/km)	R/X
Low voltage line	0.642	0.083	7.7
Medium voltage line	0.161	0.190	0.85
High voltage line	0.06	0.191	0.31

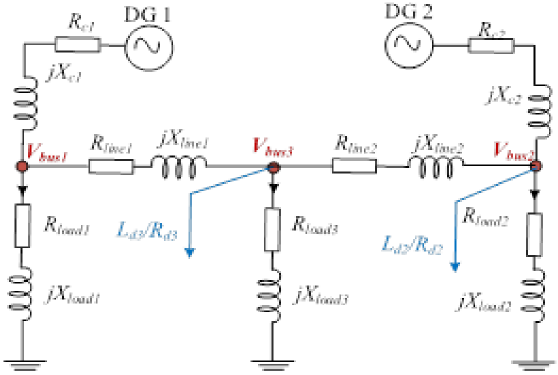


Fig. 6. Proposed Circuit Configuration.

A. Case 1: Optimal Control Strategy for Inductive MG

In this section, results obtained from proposed strategy and without proposed strategy for inductive are discussed. The key parameters and configuration for inductive MG are given in table 2 and Fig. 6, respectively

TABLE II. System Parameters for Inductive Microgrid

Parameters	**Symbol**	**Value**
Frequency	*fs*	50Hz
Inverter Rating	*VA*	10kVA
Filter 1 impedances	$R_{c1}+jX_{c1}$	0.25+*j*0.785
Filter 2 impedances	$R_{c2}+jX_{c2}$	0.25+*j*0.785
Line 1 impedances	$R_{line1}+jX_{line1}$	0.2+*j*0.628
Line 2 impedances	$R_{line2}+jX_{line2}$	0.2+*j*0.628
Load 1 impedances	$R_{load1}+jX_{load1}$	20+*j*3.140
Load 2 impedances	$R_{load2}+jX_{load2}$	20.4+*j*3.15
Load 3 impedances	$R_{load3}+jX_{load3}$	20+*j*3.140
Disturbance load at Bus 2	L_{d2}/R_{d2}	10mH/19
Disturbance load at Bus 2, 3	$L_{d2}/R_{d2,}$ L_{d3}/R_{d3}	7+*j*0.314, 10+*j*0.628

1) Power Sharing with load variation at bus 2: In this case, the all buses voltage error weight ωa-c and reactive power error weight ωQ are not changed and set at 1 which does not have any effect on cost function. In order to realize proportional power sharing and verify the optimal control strategy a heavy disturbance load Ld2/Rd2 with value of 19+j3.140 exerted at bus 2 on 0.2 seconds. Fig. 7 and 8 depicts the performance of voltage regulations and power sharing, respectively, with and without proposed control strategy. In conventional control strategy the bus voltages drop (dotted curves) can be seen in Fig. 7(a-b). since droop controllers decrease voltages in order to track the aggravated reactive power. More than 3 volts' deviation at bus 2 and bus 3 has been compensated in proposed strategy as illustrated in Fig. 7(a-b) and stabled at 297.2 V and 297.56 V, respectively. In addition, active power sharing error 4.3KW and reactive power sharing error 710VARcan be noticed in Fig. 7(c) and Fig. 8(b). Once the proposed optimal control strategy is activated at 0.2 seconds, it can be observed in Fig. 8(a and c) that active and reactive power sharing error are compensated in to almost zero with a smaller startup divergent behavior.

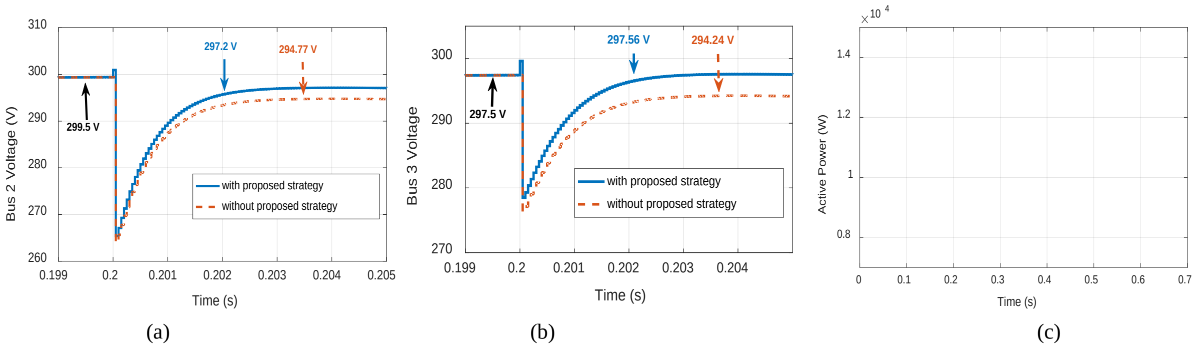


Fig. 7. Simulations Results under Heavily Load Conditions (A)Voltage Response at Bus 2 (B) Voltage Response at Bus 3 (C) Active Power Sharing without Proposed Strategy.

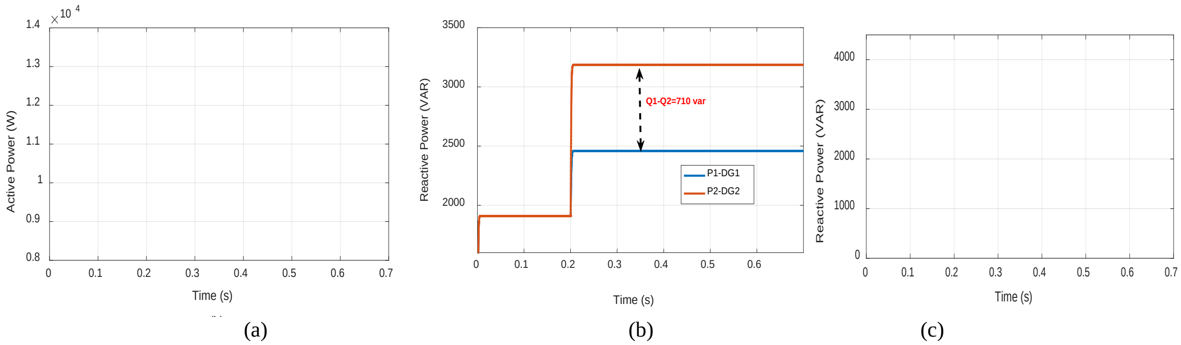


Fig. 8. (a)Active Power Sharing in Proposed Strategy, Reactive Power Sharing in with and without Proposed Strategy is Illustrated in Figure (B) and (C), Respectively.

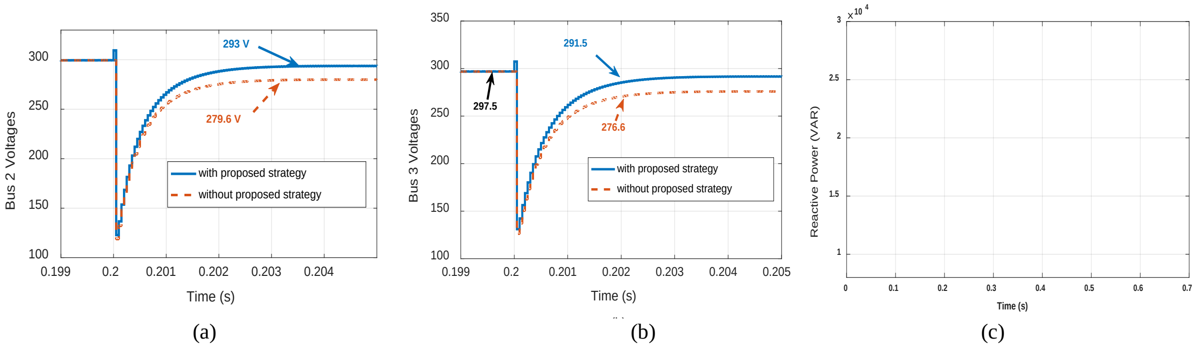


Fig. 9. Simulations Results under Heavily Load Conditions with (Green Cure) and without (Yellow Dotted Line) Proposed Control Strategy (a)Voltage Response at Bus 2 (b) Voltage Response at Bus 3 (c)active Power Sharing without Proposed Strategy.

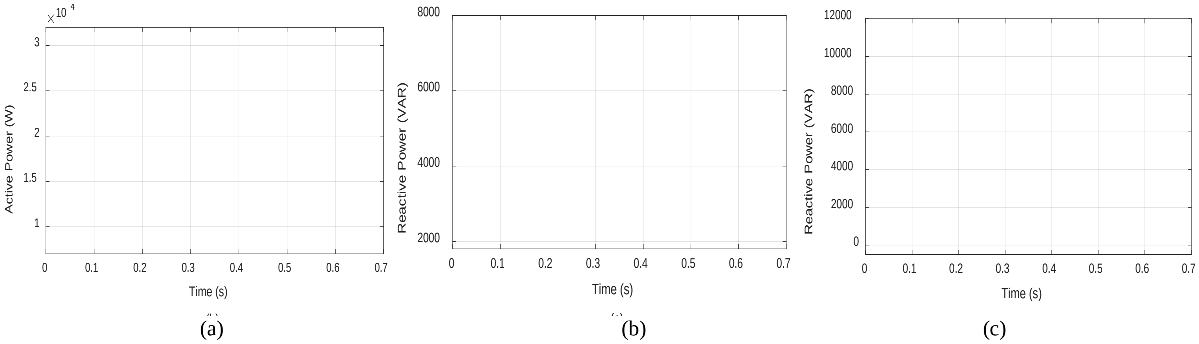


Fig. 10. Simulation Results with Proposed Strategy for active Power is Shown in Figure (a) , While Results Obtained with and without Proposed Strategy for Real and Reactive Power are Illustrated in (b) and (c), Respectively.

2) Power Sharing for different disturbance locations: This section further investigates the effectiveness of system voltage regulations and power sharing behavior to unknown multiple disturbances. In this case, a serious disturbance load (Ld2/Rd2) and (Ld3/Rd3) are exerted on same time on 0.2 seconds at bus 2 and 3, respectively, as shown in Fig. 5 which led up to20.4v and 23.4Volts deviations at bus 2 and bus 3, respectively. Proposed strategy is activated at 0.2 seconds, which reduces deviations of 13.4 V at bus 2 and stabled voltage curve at 293 V as illustrated in Fig. 9 (a), while 14.9V deviation has been compensated at bus 3 as shown in Fig. 9(b), and voltage curve is stabled within acceptable +-0.5V range. Further, multiple disturbance load at different locations effects power sharing, as active power error 8.01Kw and reactive power error 1.08KVAR is noticed in Fig. 10(c) and 10(b). When the proposed control strategy is activated at 0.2 seconds power sharing error is compensated to almost zero as shown in Fig. 9 and 10. Optimal Control Strategy for Resistive MG.

TABLE III. SYSTEM PARAMETERS FOR RESISTIVE MICROGRID

Parameters	**Symbol**	**Value**
Filter 1 impedances	Rc1+jXc1	1.5+j0.314
Filter 2 impedances	Rc2+jXc2	1.5+j0.314
Line 1 impedances	Rline1+jXline1	0.75+j0.157
Line 2 impedances	Rline2+jXline2	0.75+j0.157
Load 1 impedances	Rload1+jXload1	30+j3.140
Load 2 impedances	Rload2+jXload2	30.4+j3.15
Load 3 impedances	Rload3+jXload3	35+j3.140
Disturbance load at Bus 2	Ld4/Rd4	1mH/20

To further investigates the effectiveness of proposed optimal control strategy, the results are obtained for resistive MG as shown in Fig. 11, 12 and 13. System configuration and parameters for resistive MG are illustrated in Fig. 6 and table 3 respectively. In this case II, the objective of optimal control strategy is to realize proportional power sharing and hold bus 2 voltages at its nominal voltage V_{ref} valued 300 V, in the presence of load disturbance Ld4/Rd4.

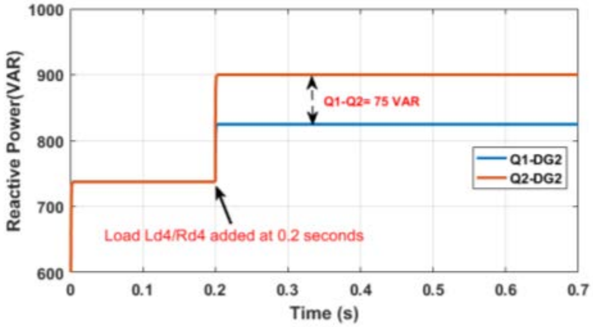


Fig. 11. Reactive Power Sharing without Proposed Control Strategy.

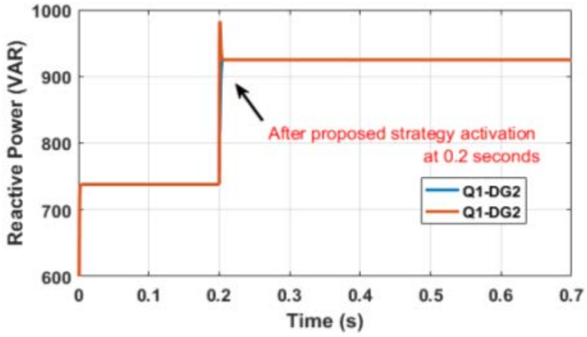


Fig. 12. Reactive Power Sharing with Proposed Control Strategy.

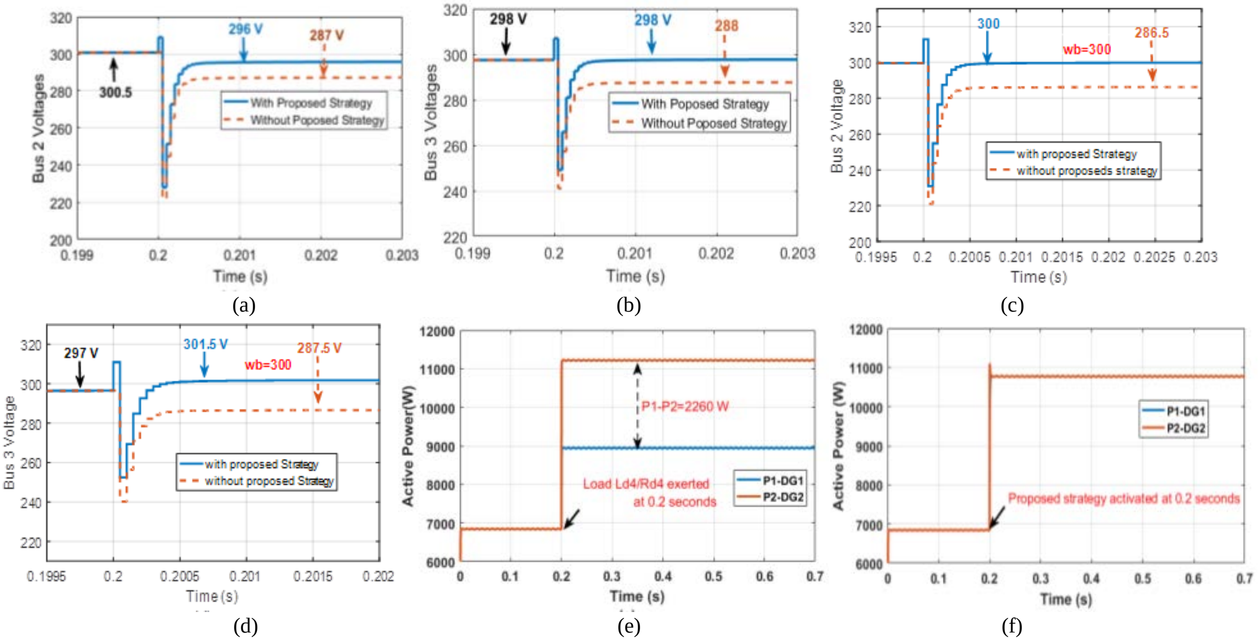


Fig. 13. Bus 2 and 3 Voltages with and without Proposed Strategy Illustrated in Figure (a) and (b), Respectively, Keeping Weight Ω_b=1, While at Ω_b=300 is Illustrated in Figure (c) and (d). Figure (e) and (f) Illustrates active Power Sharing without and with Proposed Control Strategy.

3) Power Sharing and Bus 2 Voltage Control: To validate the optimal control strategy, a disturbance load Ld4/Rd4 with valued 1mH/20ohm is exerted at bus 2 on 0.2 seconds. Voltage deviation of valued 13 V and 12 V is observed in conventional control strategy at bus 2 and 3. This voltage deviation has been compensated with help proposed control strategy, which stabled bus 2 voltage at 296 V and bus 3 voltage at 298 V, while keeping weight values wb=1 and wc=1 as illustrated in Fig. 13(a- b). Still 4V volts deviation occurred at Bus 2 in proposed control strategy as shown in Fig. 13(a). To hold bus 2 voltage at its nominal value Vref in presence of disturbance load, the voltage error weight Wb is set 300 while all other buses and reactive power weights are set at 1. The results obtained for Bus 2 after changing of its weight, are shown in Fig. 13(c-d) where it stabled to its nominal value 300 V voltage. Further, active power error 2260W and reactive power error 75 VAr has been observed in conventional control strategy as illustrated in Fig. 11 and 13(e), respectively. This power sharing error has been compensated to zero in proposed control strategy as depicted in Fig. 12 and Fig. 13(f), respectively.

The strategy adopted for frequency regulation is illustrated in Fig. 4(b). Frequency deviation is eliminated at 0.2 seconds as shown in below Fig. 14, which is within acceptable range ±0.5 Hz.

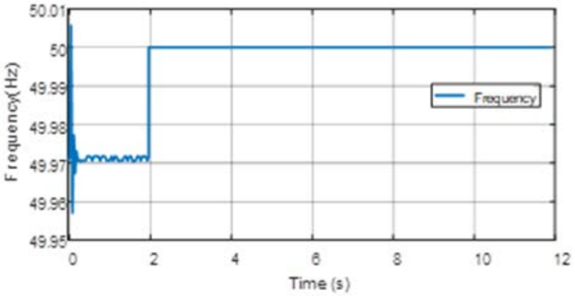


Fig. 14. Frequency Regulation.

IV. Conclusion

In this paper an optimal control strategy was proposed which performs the twofold objectives in order to realize proportional power sharing and system voltage regulation for multiple feeders in islanded AC MGs. The strategy firstly, estimates the load impedances of specified buses by using slow communication channel. Secondly, an optimal controller based optimized cost function with immunity to parameters perturbations has been developed which sends control command to inner loop in order to realize proportional power sharing and voltage control for the specified bus. Finlay, the effectiveness of proposed optimal control strategy was investigated under load parameters uncertainties in both inductive and resistive MGs. The obtained simulation results show that the proposed optimal control strategy is not sensitive to MG's configurations and able to realize proportional power sharing and controls the specified multiple feeder's voltages in ac islanded MG which, thus enhances the reliability and flexibility of islanded MG.

References

[1] S. Member, J. A. Mueller, S. Member, J. W. Kimball, and S. Member, "An Accurate Small-Signal Model of Inverter- Dominated Islanded Microgrids Using dq Reference Frame," vol. 2, no. 4, pp. 1070–1080, 2014.

[2] R. H. Lasseter and P. Paigi, "Microgrid : A Conceptual Solution," no. June, pp. 4285–4290, 2004.

[3] K. Hashmi et al., "A Virtual Micro-Islanding-Based Control Paradigm for Renewable Microgrids," pp. 1–23, 2018.

[4] M. Shahid, Muhammad Umair Mansoor Khan, K. Hashmi, S. Habib, J. Huawei, and H. Tang, "A Control Methodology for Load Sharing System Restoration in Islanded DC Micro Grid with Faulty Communication Links," Electronics, vol. 7, no. 90, pp. 1–15, 2018.

[5] J. He and Y. W. Li, "An Enhanced Microgrid Load Demand Sharing Strategy," vol. 27, no. 9, pp. 3984–3995, 2012.

[6] F. Blaabjerg, R. Teodorescu, M. Liserre, and A. V. Timbus, "Overview of control and grid synchronization for distributed power generation systems," IEEE Trans. Ind. Electron., vol. 53, no. 5, pp. 1398–1409, 2006.

[7] M. Umair and S. Id, "An Improved Control Strategy for Three-Phase Power Inverters in Islanded AC Microgrids," 2018.

[8] Han et al., "Aalborg Universitet Review of Power Sharing Control Strategies for Islanding Operation of AC Microgrids," I E E E Trans. Smart Grid, vol. 7, no. 1, pp. 200–215, 2016.

[9] W. Yao, M. Chen, J. Matas, J. M. Guerrero, and Z. M. Qian, "Design and analysis of the droop control method for parallel inverters considering the impact of the complex impedance on the power sharing," IEEE Trans. Ind. Electron., vol. 58, no. 2, pp. 576–588, 2011.

[10] C. K. Sao and P. W. Lehn, "Autonomous Load Sharing of Voltage Source Converters," IEEE Trans. Power Deliv., vol. 20, no. 2, pp. 1009–1016, 2005.

[11] K. De Brabandere et al., "A voltage and frequency droop control method for parallel inverters," Power Electron. Spec. Conf. 2004. PESC 04. 2004 IEEE 35th Annu., vol. 4, no. 4, p. 2501–2507 Vol.4, 2004.

[12] J. M. Guerrero, J. C. Vasquez, J. Matas, L. G. De Vicuña, and M. Castilla, "Hierarchical control of droop-controlled AC and DC microgrids - A general approach toward standardization," IEEE Trans. Ind. Electron., vol. 58, no. 1, pp. 158–172, 2011.

[13] C. T. Lee, C. C. Chu, and P. T. Cheng, "A new droop control method for the autonomous operation of distributed energy resource interface converters," IEEE Trans. Power Electron., vol. 28, no. 4, pp. 1980–1993, 2013.

[14] G. Modes, A. Mehrizi-sani, G. S. Member, and R. Iravani, "Potential-Function Based Control of a Microgrid in," vol. 25, no. 4, pp. 1883–1891, 2010.

[15] J. He and Y. W. Li, "An accurate reactive power sharing control strategy for DG units in a microgrid," 8th Int. Conf. Power Electron. - ECCE Asia "Green World with Power Electron. ICPE 2011-ECCE Asia, pp. 551–556, 2011.

[16] M. N. Marwali, J. Jung, S. Member, and A. Keyhani, "Control of Distributed Generation Systems — Part II : Load Sharing Control," vol. 19, no. 6, pp. 1551–1561, 2004.

[17] P. Cheng, C. Chen, T. Lee, and S. Kuo, "A Cooperative Imbalance Compensation Method for Distributed-Generation Interface Converters," vol. 45, no. 2, pp. 805–815, 2009.

[18] J. M. Guerrero, L. G. De Vicuña, J. Matas, J. Miret, and M. Castilla, "Output impedance design of parallel-connected UPS inverters," IEEE Int. Symp. Ind. Electron., vol. 2, no. 4, pp. 1123–1128, 2004.

[19] D. N. Zmood and D. G. Holmes, "Stationary frame current regulation of PWM inverters with zero steady-state error," IEEE Trans. Power Electron., vol. 18, no. 3, pp. 814–822, 2003.

[20] Y. Wang, S. Member, X. Wang, Z. Chen, and S. Member, "Distributed Optimal Control of Reactive Power and Voltage in Islanded Microgrids," vol. 53, no. 1, pp. 340–349, 2017.

Modeling of the Consensus in the Allocation of Resources in Distributed Systems

Federico Agostini[1], David L. La Red Martínez[2], Julio C. Acosta[3]
Faculty of Exact and Natural Sciences and Surveying, North-eastern National University
Corrientes, Argentina

***Abstract*—When it comes to processes distributed in process nodes that access critical resources shared in the modality of distributed mutual exclusion, it is important to know how these are managed and the order in which the demand for resources is resolved by the processes. Being in a shared environment, it is necessary to comply with certain rules, for instance, access to resources must be achieved through mutual exclusion. In this work, through an aggregation operator, a consensus mechanism is proposed to establish the order of allocation of resources to the processes. The consensus is understood as the agreement that must be achieved for the allocation of all the resources requested by each process. To model this consensus, it must be taken into account that the processes can form group of processes or be independent, the state of the nodes where each of them is located, the computational load, the number of processes, the priorities of the processes, CPU usage, use of main memory, virtual memory, etc. These characteristics allow the evaluation of the conditions to agree on the order in which allocations of resources to processes will be made.**

Keywords—Aggregation operators; communication between groups of processes; mutual exclusion; operating systems; processor scheduling

I. Introduction

The proliferation of computer systems, many of them distributed in different nodes with multiple processes that cooperate for the achievement of a particular function, requires decision models that allows groups of processes to use shared resources that can only be accessed in the modality of mutual exclusion.

The traditional solutions for this problem are found in [1] and [2], both papers describe the main synchronization algorithms in distributed systems. The author in [3] presents an efficient and fault tolerant solution for the problem of distributed mutual exclusion. The authors in [4], [5] and [6] present algorithms to manage the mutual exclusion in computer networks. In [7] are detailed the main algorithms for distributed processes management, distributed global states and distributed mutual exclusion.

The allocation of resources in processes should be performed taking into account the priorities of the processes and also the state in terms of workload of the computational nodes in which the processes are executed.

Also, solutions (which may be considered traditional) have been proposed for different types of distributed systems in [8], [9], [10], [11] and [12]. Other works that focused on ensuring mutual exclusion have been presented in [13] and [14]. An interesting distributed solution based on permissions is presented in [15] and a solution based on process priorities can be found in [16].

In this paper, a new aggregation operator will be presented specifically for solving the aforementioned problem. This falls under the category of OWA (Ordered Weighted Averaging) operators, more specifically Neat OWA. The use of aggregation operators in decision models has been widely studied. For example, [17], develops methodologies that solve problems in the presence of multiple attributes and criteria and in [18] the way to obtain a priority vector is collectively studied, which is created from different formats of expression of the preferences of decision makers. The model can reduce the complexity of decision-making and avoid the loss of information when the different formats are transformed into a single format of expression of preferences. In addition, [19] presents the main mathematical properties and behavioural measures related to the aggregation operators. A review of aggregation operators, especially those of the OWA family, is presented in [20], [21] and [22]. OWA operators applied to multicriteria decision making are presented and analysed in [23], and [24] analyse the OWA operators and their applications in the decision making process. In turn, in [25] a complex and dynamic problem of group decision making with multiple attributes is defined and a resolution method is proposed, which uses a consensus process for groups of attributes, alternatives and preferences, resulting in a decision model for problems of the real world.

This study will present a variant of an innovative method for the management of shared resources in distributed systems, based on [26] and [27], in which an aggregation operator is developed to assign resources in distributed systems. Here, we establish a consensus model that favours the sequential access of the processes to all the requested resources. The premises, data structures and the operator mentioned in [26] and [27], are used as a starting point to create a new operator in the scenario described next.

This paper, which presents an innovative method for the management of shared resources in distributed systems is structured as follows: Section 2 explains the data structures that the proposed operator will use, Section 3 describes the aggregation operator, in Section 4 a detailed example of this is shown, then the Conclusions and the Future lines of work are presented, and then the Acknowledgments, the References and the appendix are shown.

II. Data Structures to be Used

The proposed scenario considers the following conditions: In first place, the processes must have access to shared resources in the mutual exclusion modality. In second place, they must be able to form groups of processes (independent processes are considered as unitary groups). In third place, the processes must not require synchronization (that is, to be active in their respective processors at the same time) and they must have strict consensus requirements in order to gain access to the resources (an agreement is required in order to consecutively allocate the resources requested by a process, that is, once the resources allocation sequence is started, it cannot be interrupted to allocate resources to other processes, until the active process releases the resources).

These are groups of processes that are distributed in process nodes that access critical resources. These resources are shared in the form of distributed mutual exclusion and it must be decided, according to the demand for resources by the processes, what the priorities to allocate the resources to the processes that require them will be (only the resources that are available to be assigned in the processes will be taken into account, that is, those not yet allocated in certain processes).

- The access permission to the shared resources of a node will not only depend on whether the nodes are using them or not, but on the aggregation value of the preferences (priorities) of the different nodes regarding granting access to shared resources (alternatives) as well.
- The opinions (priorities) of the different nodes regarding granting access to shared resources (alternatives) will depend on the consideration of the value of variables that represent the state of each one of the different nodes. Each node must express its priorities for assigning the different shared resources according to the resource requirements of each process (which may be part of a group of processes).

These available shared resources hosted on different nodes of the distributed system may be required by the processes (clustered or independent) running on the nodes.

Possible states of each process:

- Independent process.
- Process belonging to a group of processes.
- Possible state of each one of the nodes:
- Number of processes.
- Priorities of the processes.
- CPU usage.
- Main memory usage.
- Use of virtual memory.
- Additional memory required for each resource requested by each process (depending on the availability of the data).
- Additional estimated processor load required for each resource requested by each process (depending on data availability).
- Additional estimated input / output load required for each resource requested by each process (depending on data availability).
- Status of each one of the shared resources in the distributed mutual exclusion mode in the node:
- Assigned to a local or remote process.
- Available.
- Predisposition (nodal priority) to grant access to each of the r shared resources in the mode of distributed mutual exclusion (will result from the consideration of the variables representing the node status, the priority of the processes and the additional computational load, which would mean allocating the resource to the requesting process).
- Current load of the node, which can be calculated as the average CPU, memory and input / output usage percentages at any given time (these load indicators may vary depending on the case, some may be added or changed); the current load categories, for example, High, Medium and Low, should also be defined, with value ranges for each category being indicated.

The scenario proposed in this study considers resources and processes in distributed operating systems, applied to the telecommunications environment, but without being limited to any specific communications protocol, meaning that it is a generic scheme. It is considered that the application of the proposed method would result in an increase in the traffic of control information, but the overall performance of the system would improve by allocating resources to the processes according to a holistic and cognitive decision-making scheme that also guarantees mutual exclusion in access to shared resources.

Fig. 1 shows the resources requests by the processes, the resources already assigned and the nodes in which they are located.

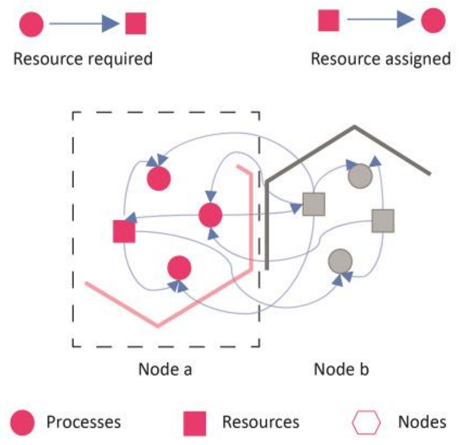


Fig. 1. Resources and Processes at Nodes in Distributed Systems.

III. Description of the Aggregation Operator

The proposed operator consists of the following steps:

1) Calculation of the current computational load of the nodes.

2) Establishment of the categories of computational load and the vectors of weights associated with them.

3) Calculation of the priorities or preferences of the processes considering the state of the node (in each node for each process).

4) Calculation of the priorities or preferences of the processes to access the available shared resources. (calculated in the centralized manager of shared resources) and determination of the allocation order and to which process the resources will be allocated.

Each one of the steps of Fig. 2 is described in [26] and [27].

In Fig. 2, there is a list of the necessary steps to obtain the final global priorities to assign the resources (*DSAF*, Distributed Systems Assignment Function).

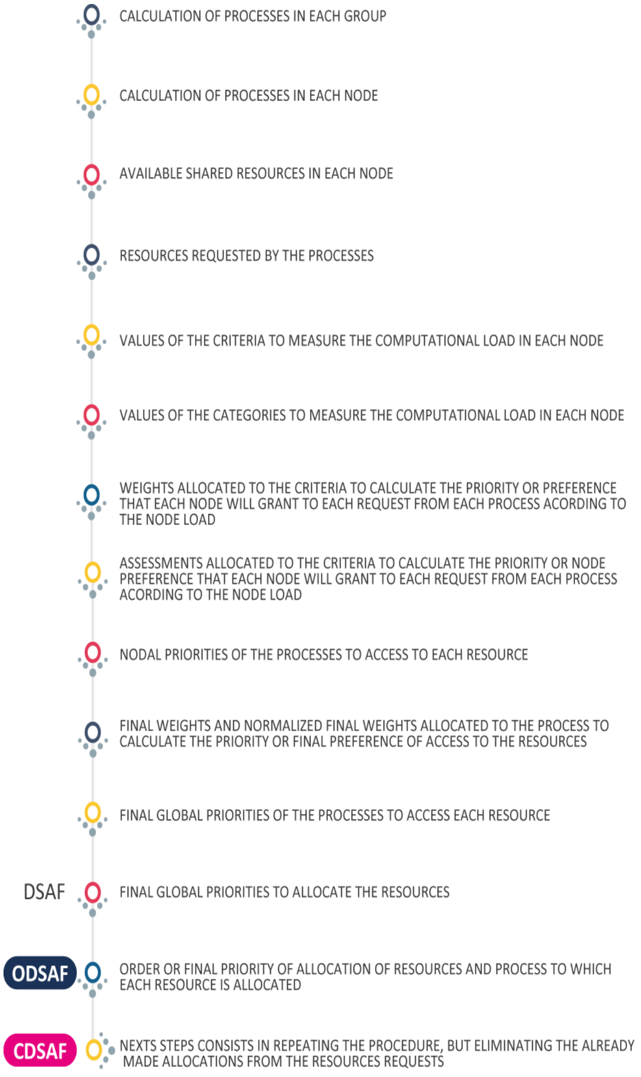

CALCULATION OF PROCESSES IN EACH GROUP
CALCULATION OF PROCESSES IN EACH NODE
AVAILABLE SHARED RESOURCES IN EACH NODE
RESOURCES REQUESTED BY THE PROCESSES
VALUES OF THE CRITERIA TO MEASURE THE COMPUTATIONAL LOAD IN EACH NODE
VALUES OF THE CATEGORIES TO MEASURE THE COMPUTATIONAL LOAD IN EACH NODE
WEIGHTS ALLOCATED TO THE CRITERIA TO CALCULATE THE PRIORITY OR PREFERENCE THAT EACH NODE WILL GRANT TO EACH REQUEST FROM EACH PROCESS ACORDING TO THE NODE LOAD
ASSESSMENTS ALLOCATED TO THE CRITERIA TO CALCULATE THE PRIORITY OR NODE PREFERENCE THAT EACH NODE WILL GRANT TO EACH REQUEST FROM EACH PROCESS ACORDING TO THE NODE LOAD
NODAL PRIORITIES OF THE PROCESSES TO ACCESS TO EACH RESOURCE
FINAL WEIGHTS AND NORMALIZED FINAL WEIGHTS ALLOCATED TO THE PROCESS TO CALCULATE THE PRIORITY OR FINAL PREFERENCE OF ACCESS TO THE RESOURCES
FINAL GLOBAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE
DSAF
FINAL GLOBAL PRIORITIES TO ALLOCATE THE RESOURCES
ODSAF
ORDER OR FINAL PRIORITY OF ALLOCATION OF RESOURCES AND PROCESS TO WHICH EACH RESOURCE IS ALLOCATED
CDSAF
NEXTS STEPS CONSISTS IN REPEATING THE PROCEDURE, BUT ELIMINATING THE ALREADY MADE ALLOCATIONS FROM THE RESOURCES REQUESTS

Fig. 2. Steps to Obtain the *DSAF*, *ODSAF* and *CDSAF* Functions.

TABLE I. Concatenation of the Ordered Assignment Tables (***ODSAF***) of Each One of the Iterations Corresponding to the General Method

ODSAF	***Iterations***
1st iteration	Rows from 1 to *n* *n*= number of rows of the ***ODSAF*** first iteration
2nd iteration	Rows from *n*+1 to *m* *m* = number of rows of the ***ODSAF*** second iteration
last iteration	*m* = number of rows of the ***ODSAF*** second iteration

The order or priority of allocation of the resources and the process to which each resource is assigned (*ODSAF*, Ordered Distributed System Assignment Function) can be seen in Table 1.

The last step is to repeat the procedure but removing the already made allocations from the resources requests (*CDSAF*, Concatenated Distribution Systems Assignment Function), as shown in Fig. 3.

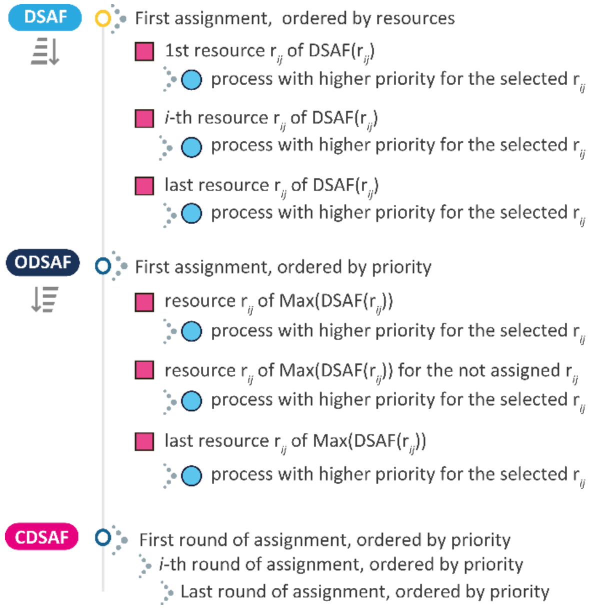

DSAF
First assignment, ordered by resources
1st resource r_ij of DSAF(r_ij)
process with higher priority for the selected r_ij
i-th resource r_ij of DSAF(r_ij)
process with higher priority for the selected r_ij
last resource r_ij of DSAF(r_ij)
process with higher priority for the selected r_ij
ODSAF
First assignment, ordered by priority
resource r_ij of Max(DSAF(r_ij))
process with higher priority for the selected r_ij
resource r_ij of Max(DSAF(r_ij)) for the not assigned r_ij
process with higher priority for the selected r_ij
last resource r_ij of Max(DSAF(r_ij))
process with higher priority for the selected r_ij
CDSAF
First round of assignment, ordered by priority
i-th round of assignment, ordered by priority
Last round of assignment, ordered by priority

Fig. 3. Steps to Obtain the *DSAF*, *ODSAF* and *CDSAF* with their Corresponding Iterations.

The *CDSAF* table is obtained from the concatenation of the *ODSAF* tables of each iteration, as shown in Table 1.

a) Final global priority of the process

Once the *CDSAF* table is completed (Table 1), the final global priorities of the processes will be calculated in order to access all of its resources, and the order in which each one will be allocated will be established, receiving all the requested resources. For this, the *CDSAF* table will be considered, the priorities of all the resource/process assignments will be added for each process, and they will be divided by the number of assignments of that process. The process with the higher final global priority will be the first one to get the requested resources. This constitutes what will be called the Final Global Priority of the Process (*FGPP*), as shown in Fig. 4.

$$FGPP_i = i = 1,\ldots,h = \frac{\sum \boldsymbol{CDSAF\,j}}{\text{Number of global assignments of the } i \text{ process}}$$

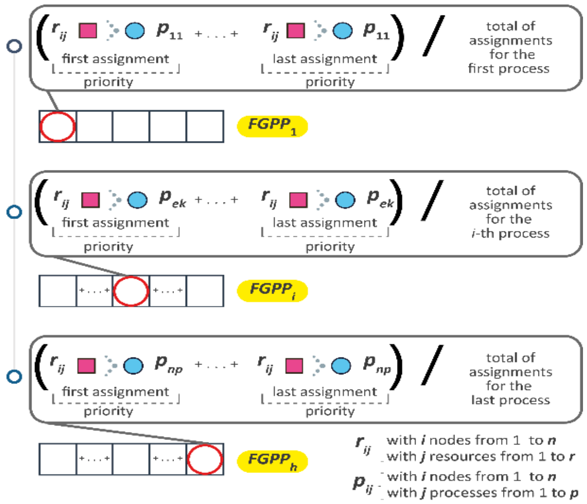

r_{ij} p_{11} + . . . + r_{ij} p_{11}
first assignment priority
last assignment priority
total of assignments for the first process
$FGPP_1$
r_{ij} p_{ek} + . . . + r_{ij} p_{ek}
total of assignments for the i-th process
$FGPP_i$
r_{ij} p_{np} + . . . + r_{ij} p_{np}
total of assignments for the last process
$FGPP_h$
r_{ij} with i nodes from 1 to n with j resources from 1 to r
p_{ij} with i nodes from 1 to n with j processes from 1 to p

Fig. 4. Calculation of the *FGPP* of each process.

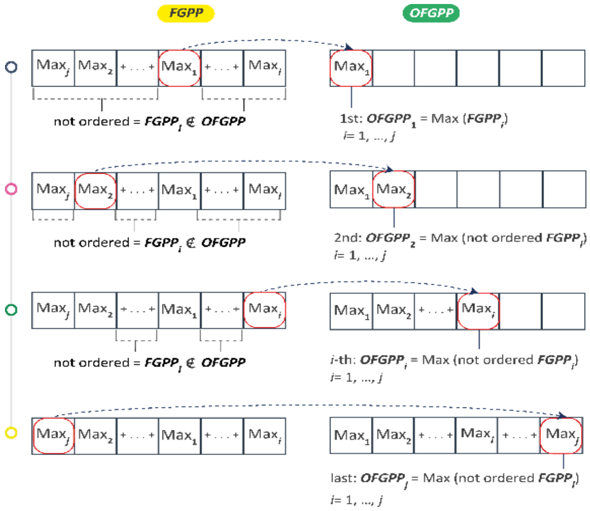

FGPP
OFGPP
Max_j Max_2 + . . . + Max_1 + . . . + Max_i
not ordered = $FGPP_i \notin OFGPP$
1st: $OFGPP_1$ = Max ($FGPP_i$) i= 1, ..., j
2nd: $OFGPP_2$ = Max (not ordered $FGPP_i$) i= 1, ..., j
i-th: $OFGPP_i$ = Max (not ordered $FGPP_i$) i= 1, ..., j
last: $OFGPP_j$ = Max (not ordered $FGPP_i$) i= 1, ..., j

Fig. 5. An example of the calculation of the OFGPP of each process.

h= total number of processes in the system (summation of processes of the nodes); *j*=number of resources allocated to the *i* process.

The elements of the *FGPP* vector must be ordered from highest to lowest to obtain the global priority order of allocation of resources to processes, as shown in Fig. 5.

Ordered Final Global Priority of the Process (*OFGPP*)

j= cardinality of *FGPP* (number of processes in the system)

OFGPPi= Max (not ordered *FGPPi*) *i*= 1, …, *j*

not ordered = *FGPPi* ∉ *OFGPP*

1st: $OFGPP_1$ = Max (*FGPPi*) *i*= 1, …, *j*

2nd: $OFGPP_2$ = Max (not ordered *FGPPi*) *i*= 1, …, *j*

last: OFGPPj = Max (not ordered FGPPi) i= 1, …, j

b) Ordered concatenated distributed system assignment function (ocdsaf)

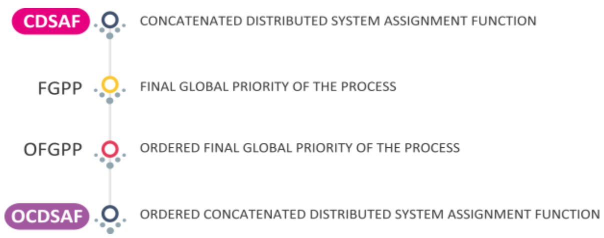

CDSAF
CONCATENATED DISTRIBUTED SYSTEM ASSIGNMENT FUNCTION
FGPP
FINAL GLOBAL PRIORITY OF THE PROCESS
OFGPP
ORDERED FINAL GLOBAL PRIORITY OF THE PROCESS
OCDSAF
ORDERED CONCATENATED DISTRIBUTED SYSTEM ASSIGNMENT FUNCTION

Fig. 6. Steps to go from the *CDSAF* to the *OCDSAF*.

The *OCDSAF* will establish the order of the final global priority allocation of processes to access its resources, and the order in which each one will be allocated, getting all the requested resources. For this, the *CDSAF* and *OFGPP* tables will be considered, as shown in Fig. 6.

The cardinalities (number of allocation of resources to each process) obtained from each one of the processes of the *OFGPP* vector in the *CDASF* table will be calculated.

CP_i = process cardinality ($OFGPP_i$) in *CDSAF*.

Then, each one of the allocations of resources to processes in the *CDSAF* table of each one of the *OFGPP* vector processes will be obtained. The total number of allocations for each process will be determined by the cardinality calculated in the previous step, as shown in Fig. 7.

In Fig 7, the first step is to calculate the priority of the process $\boldsymbol{p_{ek}}$, considering all rounds at *CDSAF*. The second step is to obtain the position in the *OFGPP* vector according to the calculated priority. The third step is to find all the assignments of the $\boldsymbol{p_{ek}}$ process in the *CDSAF* and place them in the *OCDSAF* in the order in which the p_{ek} process appears in the *OFGPP*. The representation of resources r_{ij} indicate the resources (whose first sub-index represents the node where it is and the second sub-index represents the resource number itself) that are assigned to the $\boldsymbol{p_{ek}}$ process (whose first sub-index represents the node where it is and the second sub-index represents the process number itself) in each round. Although the resources have the same sub-indexes, they are not necessarily the same resources, but they can represent different resources that are assigned several times in the different rounds, but always to the same $\boldsymbol{p_{ek}}$ process. The location in the *FASDCO* table will depend on the location in the *PGFPO* vector.

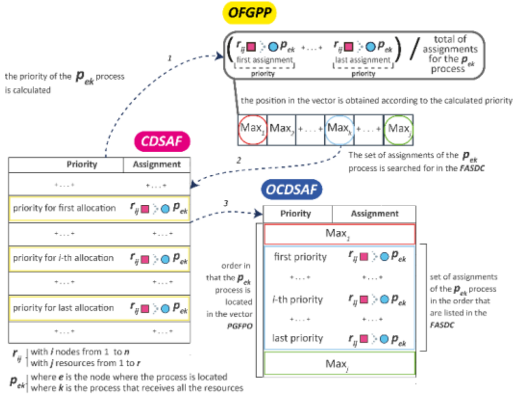

OFGPP
the priority of the p_{ek} process is calculated
1
first assignment priority
last assignment priority
total of assignments for the p_{ek} process
the position in the vector is obtained according to the calculated priority
Max_j Max_2 + . . . + Max_1 + . . . + Max_i
2
The set of assignments of the p_{ek} process is searched for in the FASDC
CDSAF
Priority
Assignment
priority for first allocation
priority for i-th allocation
priority for last allocation
3
OCDSAF
order in that the p_{ek} process is located in the vector PGFPO
first priority
i-th priority
last priority
Max_1
Max_j
set of assignments of the p_{ek} process in the order that are listed in the FASDC
r_{ij} with i nodes from 1 to n with j resources from 1 to r
p_{ek} where e is the node where the process is located where k is the process that receives all the resources

Fig. 7. Calculation of Priorities for the $\boldsymbol{P_{ek}}$ Process with the Highest Priority in *PGFPO*.

$OCDSAF_1$ = first allocation of the *CDSAF* for the (*$OFGPP_1$*) process

$OCDSAF_{cp1}$ = last allocation of the *CDSAF* for the (*$OFGPP_1$*) process

$OCDSAF_{cp1+1}$ = first allocation of the *CDSAF* for the (*$OFGPP_2$*) process

$OCDSAF_{cp1+cp2}$ = last allocation of the *CDSAF* for the (*$OFGPP_2$*) process

$OCDSAF_{cp1+cp2+\ldots+cpk-1+1}$ = first allocation of the *CDSAF* for the (*$OFGPP_k$*) process

$OCDSAF_{cp1+cp2+\ldots+cpk}$ = last allocation of the *CDSAF* for the (*$OFGPP_k$*) process

$OCDSAF_{cp1+cp2+\ldots+cpn-1+1}$ = first allocation of the *CDSAF* for the (*$OFGPP_n$*) process

$OCDSAF_{cp1+cp2+\ldots+cpn}$ = last allocation of the *CDSAF* for the (*$OFGPP_n$*) *process*

c) Considerations for aggregation operations

The characteristics of the aggregation operations described allow to consider that the proposed method belongs to the family of aggregation operators Neat-OWA, which are characterized as follows:

The definition of OWA operators indicates

$$f\left(a_1, a_2, \ldots, a_n\right) = \sum_{j=1}^{n} w_j \cdot b_j \tag{1}$$

Where b_j is the j_{th} highest value of the a_n, with the restriction for weights to satisfy

$$w_i \in [0{,}1] \tag{2}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{3}$$

For the Neat OWA operator family, the weights will be calculated according to the elements that are added, or more exactly to the values to be orderly added, the b_j, maintaining conditions (2) and (3). In this case the weights are $w_i=f_i$ $(b_1,\ldots,b_n)$, defining the operator:

$$F(a_1,\ldots a_n) = \sum_i f_i(b_1,\ldots,b_n) \cdot b_i \tag{4}$$

This family, in which the weights depend on the aggregation, do not require to meet all properties of OWA operators.

In addition, in order to be able to assert that an aggregation operator is *neat*, the final aggregation value needs to be independent of the order of the values. $\boldsymbol{A}=(a_1,\ldots,a_n)$ being the entries to add, $\boldsymbol{B}=(b_1,\ldots,b_n)$ being the ordered entries and $\boldsymbol{C}=(c_1,\ldots,c_n)=$ *Perm*$(a_1,\ldots,a_n)$ being a permutation of the entries. An OWA operator is defined as *neat* if

It produces the same result for any assignment $C = B$

$$F\left(a_1, a_2, \ldots, a_n\right) = \sum_{i=1}^{n} w_i \cdot b_i \tag{5}$$

One of the characteristics to be pointed out by Neat OWA operators is that the values to be added do not need to be sorted out for their process. This implies that the formulation of a neat operator can be defined by the arguments instead of the orderly elements.

In the proposed aggregation operator, the weights are calculated according to context values. From this context, arise the values to be aggregated.

IV. Example and Discussion of Results

This section will explain in detail an example of application of the proposed aggregation operator. This example takes as a starting point the ordered *DSAF* vector from [26] and [27], and these steps are shown in Fig. 2.

The example seen in [26] shows the following calculations:

- The priorities or preferences of the processes to access the available shared resources.
- The vector of final weights that will be used in the final aggregation process to determine the order or priority of access to the resources.

The greatest of these products made for the different processes in relation to the same resource, will indicate which one of the processes will get access to the resource.

The summation of all these products in relation to the same resource will indicate the priority that said resource will have in order to be assigned. This constitutes the Distributed System Assignment Function (*DSAF*) that can be seen in Table 2.

The final order of allocation of the resources and the recipient processes is obtained by ordering Table 2, as shown in Table 3.

TABLE II. Final Global Priorities for Allocating the Resources (*DSAF*) in the First Iteration

Resources	***Priority***	***Assignment***
r_{11}	0.35120968	r_{11} to p_{37}
r_{12}	0.47306452	r_{12} to p_{37}
r_{13}	0.32862903	r_{13} to p_{13}
r_{21}	0.33000000	r_{21} to p_{37}
r_{22}	0.34403226	r_{22} to p_{34}
r_{23}	0.24919355	r_{23} to p_{11}
r_{24}	0.18951613	r_{24} to p_{34}
r_{31}	0.37048387	r_{31} to p_{34}
r_{32}	0.30322581	r_{32} to p_{34}
r_{33}	0.46798387	r_{33} to p_{23}

TABLE III. ORDER OR FINAL PRIORITY OF ASSIGNMENT OF RESOURCES AND PROCESS TO WHICH IS ALLOCATED EACH RESOURCE (***ODSAF***) IN THE FIRST ITERATION

Ordered Final Global Priority	***Assignment***
0.47306452	r_{12} to p_{37}
0.46798387	r_{33} to p_{23}
0.37048387	r_{31} to p_{34}
0.35120968	r_{11} to p_{37}
0.34403226	r_{22} to p_{34}
0.33000000	r_{21} to p_{37}
0.32862903	r_{13} to p_{13}
0.30322581	r_{32} to p_{34}
0.24919355	r_{23} to p_{11}
0.18951613	r_{24} to p_{34}

The next step is to repeat the procedure, but removing the requests of already made allocations; it must be noted that the assigned resources will be available once they are released by the processes, and can therefore be allocated to other processes.

In this way, all the resources requests from all the processes will be answered, considering mutual exclusion and priorities of the processes, nodal priorities and final priorities, according to the scenario presented in [26] and [27].

The scenario presented next, starts from the concatenation of the ordered assignment of each one of the iterations corresponding to the above mentioned scenario.

The *CDSAF* table will be obtained from the concatenation of the *ODSAF* table of each iteration, as shown in Table 4.

TABLE IV. ORDER OR FINAL PRIORITY OF ASSIGNMENT OF RESOURCES AND PROCESS TO WHICH IS ALLOCATED EACH RESOURCE IN ALL ITERATIONS (*CDSAF*)

Ordered Final Priority	Assignment	Round
0.47306452	r_{12} al p_{37}	1
0.46798387	r_{33} al p_{23}	1
0.37048387	r_{31} al p_{34}	1
0.35120968	r_{11} al p_{37}	1
0.34403226	r_{22} al p_{34}	1
0.33000000	r_{21} al p_{37}	1
0.32862903	r_{13} al p_{13}	1
0.30322581	r_{32} al p_{34}	1
0.24919355	r_{23} al p_{11}	1
0.18951613	r_{24} al p_{34}	1
0.40653226	r_{33} al p_{34}	2
0.39951613	r_{12} al p_{34}	2
0.30346774	r_{31} al p_{13}	2
0.28153226	r_{11} al p_{11}	2
0.27024194	r_{22} al p_{11}	2
0.26274194	r_{21} al p_{25}	2
0.25701613	r_{13} al p_{34}	2
0.23790323	r_{32} al p_{37}	2
0.17322581	r_{23} al p_{34}	2
0.13435484	r_{24} al p_{11}	2
0.34677419	r_{33} al p_{13}	3
0.33443548	r_{12} al p_{23}	3
0.24250000	r_{31} al p_{21}	3
0.22330645	r_{22} al p_{13}	3
0.21233871	r_{11} al p_{13}	3
0.19983871	r_{21} al p_{13}	3
0.18612903	r_{13} al p_{31}	3
0.17524194	r_{32} al p_{13}	3
0.10790323	r_{23} al p_{21}	3
0.09516129	r_{24} al p_{23}	3
0.28725806	r_{33} al p_{37}	4
0.27637097	r_{12} al p_{13}	4
0.19637097	r_{31} al p_{23}	4
0.17975806	r_{22} al p_{12}	4
0.15725806	r_{21} al p_{12}	4
0.14314516	r_{11} al p_{12}	4
0.13629032	r_{13} al p_{21}	4
0.11717742	r_{32} al p_{23}	4
0.07096774	r_{23} al p_{32}	4
0.06298387	r_{24} al p_{35}	4
0.22798387	r_{33} al p_{12}	5
0.22459677	r_{12} al p_{11}	5
0.15185484	r_{31} al p_{31}	5
0.13846774	r_{22} al p_{21}	5
0.11596774	r_{21} al p_{22}	5
0.09709677	r_{13} al p_{32}	5
0.08991935	r_{11} al p_{32}	5
0.06685484	r_{32} al p_{36}	5
0.04403226	r_{23} al p_{33}	5
0.04112903	r_{24} al p_{36}	5
0.18282258	r_{33} al p_{31}	6
0.17669355	r_{12} al p_{12}	6
0.11411290	r_{31} al p_{12}	6
0.09943548	r_{22} al p_{22}	6
0.07741935	r_{21} al p_{11}	6
0.06983871	r_{13} al p_{36}	6
0.06604839	r_{11} al p_{36}	6
0.04322581	r_{32} al p_{35}	6
0.02056452	r_{23} al p_{24}	6
0.02024194	r_{24} al p_{24}	6
0.14056452	r_{33} al p_{21}	7
0.13669355	r_{12} al p_{21}	7
0.07669355	r_{31} al p_{22}	7
0.05443548	r_{22} al p_{35}	7
0.04354839	r_{13} al p_{35}	7
0.04306452	r_{21} al p_{33}	7
0.04266129	r_{11} al p_{33}	7
0.02104839	r_{32} al p_{33}	7
0.10975806	r_{12} al p_{33}	8
0.09862903	r_{33} al p_{22}	8
0.04782258	r_{31} al p_{36}	8
0.03306452	r_{22} al p_{33}	8
0.02145161	r_{21} al p_{36}	8
0.02104839	r_{13} al p_{33}	8
0.02032258	r_{11} al p_{24}	8
0.08443548	r_{12} al p_{36}	9
0.06588710	r_{33} al p_{33}	9
0.02217742	r_{31} al p_{35}	9
0.01217742	r_{22} al p_{36}	9
0.06032258	r_{12} al p_{24}	10
0.04250000	r_{33} al p_{35}	10
0.03798387	r_{12} al p_{32}	10
0.01959677	r_{33} al p_{36}	10
0.01693548	r_{12} al p_{35}	11

TABLE V. FINAL GLOBAL PRIORITY ORDERED BY PROCESS

Final Global Priority	Resources	Process	Round
0.24919355	r_{23}	p_{11}	1
0.28153226	r_{11}	p_{11}	2
0.27024194	r_{22}	p_{11}	2
0.13435484	r_{24}	p_{11}	2
0.22459677	r_{12}	p_{11}	5
0.07741935	r_{21}	p_{11}	6
0.17975806	r_{22}	p_{12}	4
0.15725806	r_{21}	p_{12}	4
0.14314516	r_{11}	p_{12}	4
0.22798387	r_{33}	p_{12}	5
0.17669355	r_{12}	p_{12}	6
0.11411290	r_{31}	p_{12}	6
0.32862903	r_{13}	p_{13}	1
0.30346774	r_{31}	p_{13}	2
0.34677419	r_{33}	p_{13}	3
0.22330645	r_{22}	p_{13}	3
0.21233871	r_{11}	p_{13}	3
0.19983871	r_{21}	p_{13}	3
0.17524194	r_{32}	p_{13}	3
0.27637097	r_{12}	p_{13}	4
0.24250000	r_{31}	p_{21}	3
0.10790323	r_{23}	p_{21}	3
0.13629032	r_{13}	p_{21}	4
0.13846774	r_{22}	p_{21}	5
0.14056452	r_{33}	p_{21}	7
0.13669355	r_{12}	p_{21}	7
0.11596774	r_{21}	p_{22}	5
0.09943548	r_{22}	p_{22}	6
0.07669355	r_{31}	p_{22}	7
0.09862903	r_{33}	p_{22}	8
0.46798387	r_{33}	p_{23}	1
0.33443548	r_{12}	p_{23}	3
0.09516129	r_{24}	p_{23}	3
0.19637097	r_{31}	p_{23}	4
0.11717742	r_{32}	p_{23}	4
0.02056452	r_{23}	p_{24}	6
0.02024194	r_{24}	p_{24}	6
0.02032258	r_{11}	p_{24}	8
0.06032258	r_{12}	p_{24}	10
0.26274194	r_{21}	p_{25}	2
0.18612903	r_{13}	p_{31}	3
0.15185484	r_{31}	p_{31}	5
0.18282258	r_{33}	p_{31}	6
0.07096774	r_{23}	p_{32}	4
0.09709677	r_{13}	p_{32}	5
0.08991935	r_{11}	p_{32}	5
0.03798387	r_{12}	p_{32}	10
0.04403226	r_{23}	p_{33}	5
0.04306452	r_{21}	p_{33}	7
0.04266129	r_{11}	p_{33}	7
0.02104839	r_{32}	p_{33}	7
0.10975806	r_{12}	p_{33}	8
0.03306452	r_{22}	p_{33}	8
0.02104839	r_{13}	p_{33}	8
0.06588710	r_{33}	p_{33}	9
0.37048387	r_{31}	p_{34}	1
0.34403226	r_{22}	p_{34}	1
0.30322581	r_{32}	p_{34}	1
0.18951613	r_{24}	p_{34}	1
0.40653226	r_{33}	p_{34}	2
0.39951613	r_{12}	p_{34}	2
0.25701613	r_{13}	p_{34}	2
0.17322581	r_{23}	p_{34}	2
0.06298387	r_{24}	p_{35}	4
0.04322581	r_{32}	p_{35}	6
0.05443548	r_{22}	p_{35}	7
0.04354839	r_{13}	p_{35}	7
0.02217742	r_{31}	p_{35}	9
0.04250000	r_{33}	p_{35}	10
0.01693548	r_{12}	p_{35}	11
0.06685484	r_{32}	p_{36}	5
0.04112903	r_{24}	p_{36}	5
0.06983871	r_{13}	p_{36}	6
0.06604839	r_{11}	p_{36}	6
0.04782258	r_{31}	p_{36}	8
0.02145161	r_{21}	p_{36}	8
0.08443548	r_{12}	p_{36}	9
0.01217742	r_{22}	p_{36}	9
0.01959677	r_{33}	p_{36}	10
0.47306452	r_{12}	p_{37}	1
0.35120968	r_{11}	p_{37}	1
0.33000000	r_{21}	p_{37}	1
0.23790323	r_{32}	p_{37}	2
0.28725806	r_{33}	p_{37}	4

Once the ***CDSAF*** table is completed, the Final Global Priorities of the Process (***FGPP***) will be calculated:

FGPP$_1$ = (0.24919355 + 0.28153226 + 0.27024194 + 0.13435484 + 0.22459677 + 0.07741935) / 6

FGPP$_2$ = (0.17975806 + 0.15725806 + 0.14314516 + 0.22798387 + 0.17669355 + 0.11411290) / 6

FGPP$_3$ = (0.32862903 + 0.30346774 + 0.34677419 + 0.22330645 + 0.21233871 + 0.19983871 + 0.17524194 + 0.27637097) / 7

FGPP$_4$ = (0.24250000 + 0.10790323 + 0.13629032 + 0.13846774 + 0.14056452 + 0.13669355) / 6

FGPP$_5$ = (0.11596774 + 0.09943548 + 0.07669355 + 0.09862903) / 4

FGPP$_6$ = (0.46798387 + 0.33443548 + 0.09516129 + 0.19637097 + 0.11717742) / 5

FGPP$_7$ = (0.02056452 + 0.02024194 + 0.02032258 + 0.06032258) / 4

FGPP$_8$ = 0.26274194 / 1

FGPP$_9$ = (0.18612903 + 0.15185484 + 0.18282258) / 3

FGPP$_{10}$ = (0.07096774 + 0.09709677 + 0.08991935 + 0.03798387) / 4

FGPP$_{11}$ = (0.04403226 + 0.04306452 + 0.04266129 + 0.02104839 + 0.10975806 + 0.03306452 + 0.02104839 + 0.06588710) / 8

FGPP$_{12}$ = (0.37048387 + 0.34403226 + 0.30322581 + 0.18951613 + 0.40653226 + 0.39951613 + 0.25701613 + 0.17322581) / 8

FGPP$_{13}$ = (0.06298387 + 0.04322581 + 0.05443548 + 0.04354839 + 0.02217742 + 0.04250000 + 0.01693548) / 7

FGPP$_{14}$ = (0.06685484 + 0.04112903 + 0.06983871 + 0.06604839 + 0.04782258 + 0.02145161 + 0.08443548 + 0.01217742 + 0.01959677) / 9

FGPP$_{15}$ = (0.47306452 + 0.35120968 + 0.33000000 + 0.23790323 + 0.28725806) / 5

The ***CDSAF*** table ordered by process, as shown in **Table 5**.

By calculating the ***FGPP*** for all the processes, as shown in **Table 5**, a vector will be obtained, as shown in **Table 6**.

The elements of the ***FGPP*** vector must be ordered from highest to lowest, in order to obtain the global priority order of allocation of resources to processes, as can be seen in **Table 7**.

TABLE VI. FINAL GLOBAL PRIORITY OF THE PROCESS (***FGPP***)

Ordered Final Global Priority	***Assignment***
0.20622312	p_{11}
0.16649193	p_{12}
0.25824597	p_{13}
0.15040323	p_{21}
0.09768145	p_{22}
0.24222581	p_{23}
0.03036291	p_{24}
0.26274194	p_{25}
0.17360215	p_{31}
0.07399193	p_{32}
0.04757057	p_{33}
0.30544355	p_{34}
0.04082949	p_{35}
0.04770609	p_{36}
0.33588710	p_{37}

TABLE VII. ORDERED FINAL GLOBAL PRIORITY OF THE PROCESS (***OFGPP***)

Ordered Final Global Priority	***Process***
0.33588710	p_{37}
0.30544355	p_{34}
0.26274194	p_{25}
0.25824597	p_{13}
0.24222581	p_{23}
0.20622312	p_{11}
0.17360215	p_{31}
0.16649193	p_{12}
0.15040323	p_{21}
0.09768145	p_{22}
0.07399193	p_{32}
0.04770609	p_{36}
0.04757057	p_{33}
0.04082949	p_{35}
0.03036291	p_{24}

The cardinalities (number of allocation of resources to each process) obtained from each one of the ***OFGPP*** vector processes in the ***CDSAF*** table will be calculated.

CP$_{37}$= process cardinality(***OFGPP***$_{1}$) in ***CDSAF*** = 5

CP$_{34}$= process cardinality(***OFGPP***$_{2}$) in ***CDSAF*** = 8

CP$_{25}$= process cardinality(***OFGPP***$_{3}$) in ***CDSAF*** = 1

CP$_{13}$= process cardinality(***OFGPP***$_{4}$) in ***CDSAF*** = 8

CP$_{23}$= process cardinality(***OFGPP***$_{5}$) in ***CDSAF*** = 5

CP$_{11}$= process cardinality(***OFGPP***$_{6}$) in ***CDSAF*** = 6

CP$_{31}$= process cardinality(***OFGPP***$_{7}$) in ***CDSAF*** = 3

CP$_{12}$= process cardinality(***OFGPP***$_{8}$) in ***CDSAF*** = 6

CP$_{21}$= process cardinality(***OFGPP***$_{9}$) in ***CDSAF*** = 6

CP$_{22}$= process cardinality(***OFGPP***$_{10}$) in ***CDSAF*** = 4

CP$_{32}$= process cardinality(***OFGPP***$_{11}$) in ***CDSAF*** = 4

CP$_{36}$= process cardinality(***OFGPP***$_{12}$) in ***CDSAF*** = 9

CP$_{33}$= process cardinality(***OFGPP***$_{13}$) in ***CDSAF*** = 8

CP$_{35}$= process cardinality(***OFGPP***$_{14}$) in ***CDSAF*** = 7

CP$_{24}$= process cardinality(***OFGPP***$_{15}$) in ***CDSAF*** = 4

Then, each one of the allocation of resources to processes in the ***CDSAF*** table of each process of ***OFGPP*** vector must be obtained. The total number of elements for each process will be determined by the cardinality calculated in the previous step.

OCDSAF$_{1}$ = first element of the ***CDSAF*** for the process (***OFGPP***$_{1}$)

OCDSAF$_{5}$ = last element of the ***CDSAF*** for the process (***OFGPP***$_{1}$)

OCDSAF$_{5+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{2}$)

OCDSAF$_{5+8}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{2}$) process

OCDSAF$_{5+8+1}$ = the element of the ***CDSAF*** for the (***OFGPP***$_{3}$)

OCDSAF$_{5+8+1+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{4}$) process

OCDSAF$_{5+8+1+8}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{4}$) process

OCDSAF$_{5+8+1+8+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{5}$) process

OCDSAF$_{5+8+1+8+5}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{5}$) process

OCDSAF$_{5+8+1+8+5+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{6}$) process

OCDSAF$_{5+8+1+8+5+6}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{6}$) process

OCDSAF$_{5+8+1+8+5+6+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{7}$) process

OCDSAF$_{5+8+1+8+5+6+3}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{7}$) process

OCDSAF$_{5+8+1+8+5+6+3+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{8}$) process

OCDSAF$_{5+8+1+8+5+6+3+6}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{8}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{9}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{9}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{10}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{10}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{11}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{11}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{12}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{12}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{13}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+8}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{13}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+8+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{14}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+8+7}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{14}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+8+7+1}$ = first element of the ***CDSAF*** for the (***OFGPP***$_{15}$) process

OCDSAF$_{5+8+1+8+5+6+3+6+6+4+4+9+8+7+4}$ = last element of the ***CDSAF*** for the (***OFGPP***$_{15}$) process.

Table 8 shows the order of all the resource allocations for each process, which one is the first process with greater global priority, and is the one to which the resources are assigned first. The complete table continues for each one of the requests for each process (***OCDSAF***).

TABLE VIII. FINAL ORDER OF ALLOCATION OF EACH ONE OF THE RESOURCES TO EACH OF ONE PROCESSES OF THE ***(OCDSAF)***

Priority	**Resource**	**Process**	**Round**
0.4730645	r_{12}	p_{37}	1
0.3512097	r_{11}	p_{37}	1
0.3300000	r_{21}	p_{37}	1
0.2379032	r_{32}	p_{37}	2
0.2872581	r_{33}	p_{37}	4
0.3704839	r_{31}	p_{34}	1
0.3440323	r_{22}	p_{34}	1
0.3032258	r_{32}	p_{34}	1
0.1895161	r_{24}	p_{34}	1
0.4065323	r_{33}	p_{34}	2
0.3995161	r_{12}	p_{34}	2
0.2570161	r_{13}	p_{34}	2
0.1732258	r_{23}	p_{34}	2
0.2627419	r_{21}	p_{25}	2
0.3286290	r_{13}	p_{13}	1
0.3034677	r_{31}	p_{13}	2
0.3467742	r_{33}	p_{13}	3
0.2233065	r_{22}	p_{13}	3
0.2123387	r_{11}	p_{13}	3
0.1998387	r_{21}	p_{13}	3
0.1752419	r_{32}	p_{13}	3
0.2763710	r_{12}	p_{13}	4
0.4679839	r_{33}	p_{23}	1
0.3344355	r_{12}	p_{23}	3
0.0951613	r_{24}	p_{23}	3
0.1963710	r_{31}	p_{23}	4
0.1171774	r_{32}	p_{23}	4
0.2491936	r_{23}	p_{11}	1
0.2815323	r_{11}	p_{11}	2
0.2702419	r_{22}	p_{11}	2
0.1343548	r_{24}	p_{11}	2
0.2245968	r_{12}	p_{11}	5
0.0774194	r_{21}	p_{11}	6
0.1861290	r_{13}	p_{31}	3
0.1518548	r_{31}	p_{31}	5
0.1828226	r_{33}	p_{31}	6
0.1797581	r_{22}	p_{12}	4
0.1572581	r_{21}	p_{12}	4
0.1431452	r_{11}	p_{12}	4
0.2279839	r_{33}	p_{12}	5
0.1766936	r_{12}	p_{12}	6
0.1141129	r_{31}	p_{12}	6
0.2425000	r_{31}	p_{21}	3
0.1079032	r_{23}	p_{21}	3
0.1362903	r_{13}	p_{21}	4
0.1384677	r_{22}	p_{21}	5
0.1405645	r_{33}	p_{21}	7
0.1366936	r_{12}	p_{21}	7
0.1159677	r_{21}	p_{22}	5
0.0994355	r_{22}	p_{22}	6
0.0766936	r_{31}	p_{22}	7
0.0986290	r_{33}	p_{22}	8
0.0709677	r_{23}	p_{32}	4
0.0970968	r_{13}	p_{32}	5
0.0899194	r_{11}	p_{32}	5
0.0379839	r_{12}	p_{32}	10
0.0668548	r_{32}	p_{36}	5
0.0411290	r_{24}	p_{36}	5
0.0698387	r_{13}	p_{36}	6
0.0660484	r_{11}	p_{36}	6
0.0478226	r_{31}	p_{36}	8
0.0214516	r_{21}	p_{36}	8
0.0844355	r_{12}	p_{36}	9
0.0121774	r_{22}	p_{36}	9
0.0195968	r_{33}	p_{36}	10
0.0440323	r_{23}	p_{33}	5
0.0430645	r_{21}	p_{33}	7
0.0426613	r_{11}	p_{33}	7
0.0210484	r_{32}	p_{33}	7
0.1097581	r_{12}	p_{33}	8
0.0330645	r_{22}	p_{33}	8
0.0210484	r_{13}	p_{33}	8

0.0658871	r_{33}	p_{33}	9
0.0629839	r_{24}	p_{35}	4
0.0432258	r_{32}	p_{35}	6
0.0544355	r_{22}	p_{35}	7
0.0435484	r_{13}	p_{35}	7
0.0221774	r_{31}	p_{35}	9
0.0425000	r_{33}	p_{35}	10
0.0169355	r_{12}	p_{35}	11
0.0205645	r_{23}	p_{24}	6
0.0202419	r_{24}	p_{24}	6
0.0203226	r_{11}	p_{24}	8
0.0603226	r_{12}	p_{24}	10

In this way, all the requests of resources from all the processes were answered, considering the mutual exclusion and the priorities of the processes, the nodal priorities and the final priorities, taking into account the strict consensus requirements established for this scenario.

V. Evaluation

The data structure mentioned above and the aggregation method used are not fully covered by traditional methods.

This work considers the global average of priorities that each process has over all the resources of all its assignments in the different rounds, but for the final global allocation, it respects the same order of allocation of each resource in the different rounds in which they were assigned in the general scenario. That is, the choice of which process will be granted resources, is established with the global average of priorities in all assignments, but the order in which those assignments are to be made, respects the one in the table ***FASD***, for each process.

The proposed model manages to establish a consensus that allows processes to access all their resources sequentially and that these cannot be removed until the process that holds them releases them. The order of assignment will be determined by the overall average priority of all the assignments. The distributed system regulates and constantly updates the local state of each node, the decisions of access to resources modify these states so it must be readjusted repeatedly, guaranteeing mutual exclusion and reordering new priorities. The method must be repeated whenever there are processes that require shared resources.

VI. Conclusions

The proposed model includes, as a particular case, a method that consists in considering the global priority of the processes, instead of a group of state variables of each node. As the processes are executed in different processors using all their resources, there is no conflict in running several processes in the same processor. In this scenario, no account is taken of the amount of time each process will use in a processor of a particular node. Nor is the amount of time in which each resource will be assigned to a particular process Another notable feature of the proposal is its ease of implementation in the environment of a centralized administrator of shared resources of a distributed system.

VII. Future Lines of Research

It is considered to develop decision models from the cognitive point of view for decision making in groups of processes, contemplating the principles of cybernetics of second order, in the context of complex systems of self-regulation, which transcend the traditional approach of computer science considering the possibility of imputation of missing data, for example, as a consequence of problems in communications between processes, and fuzzyfication of variables to support situations where it is not possible or convenient to express exact values.

In addition, the aim is to investigate the impact on data traffic of applying the proposed method and comparing it with other classical methods. To this end, a simulator will be developed in which the different possible scenarios will be considered to allow the system to predict, compare and optimise the behaviour of its simulated processes in a very short time without the cost or risk of carrying them out, making it possible to represent the processes, resources and nodes in a dynamic model.

Another possible line of research considers aspects related to security in the execution of processes, access to resources and communication between nodes.

Acknowledgment

This work has been supported by the Project: "Decision models and aggregation operators for process management in distributed systems", code 16F001 of the Northeastern National University (Argentina).

References

[1] S. Tanenbaum, Sistemas Operativos Distribuidos. Prentice - Hall Hispanoamericana S.A., México, 1996.

[2] A. S. Tanenbaum, **Sistemas Operativos Modernos**. 3ra. Edición. Pearson Educación S. A., México, 2009.

[3] D. Agrawal, A. El Abbadi, "An Efficient and Fault-Tolerant Solution of Distributed Mutual Exclusion". **ACM Trans. on Computer Systems**. Vol. 9, pp. 1-20, USA, 1991.

[4] G. Ricart, A. K. Agrawala, "An Optimal Algorithm for Mutual Exclusion in Computer Networks". **Commun. of the ACM**. Vol. 24, pp. 9-17, 1981.

[5] G. Cao, M. Singhal, "A Delay-Optimal Quorum-Based Mutual Exclusion Algorithm for Distributed Systems". **IEEE Transactions on Parallel and Distributed Systems.** Vol. 12, no. 12, pp. 1256-1268. USA, 2001.

[6] S. Lodha, A. Kshemkalyani, "A Fair Distributed Mutual Exclusion Algorithm". **IEEE Trans. Parallel and Distributed Systems.** Vol. 11, no. 6, pp. 537-549, USA, 2000.

[7] W. Stallings, **Sistemas Operativos**. 5ta. Edición. Pearson Educación S.A., España, 2005.

[8] G. Andrews, Foundation of Multithreaded, Parallel, and Distributed Programming. Reading, MA: Addison-Wesley. USA, 2000.

[9] R. Guerraoui, L. Rodrigues, **Introduction to Reliable Distributed Programming.** Berlin, Springer-Verlag, 2006.

[10] N. Lynch, **Distributed Algorithms**. Morgan Kauffman, San Mateo, CA, USA, 1996.

[11] G. Tel, **Introduction to Distributed Algorithms**. Cambridge University Press, 2nd ed. Cambridge, UK, 2000.

[12] H. Attiya, J. Welch, **Distributed Computing Fundamentals, Simulations, and Advanced Topics**, John Wiley, 2nd ed., New York, USA, 2004.

[13] P. Saxena, J. Rai, "A Survey of Permission-based Distributed Mutual Exclusion Algorithms". **Computer Standards and Interfaces**, vol. (25)2, pp 159-181, 2003.

[14] M. Velazquez, "A Survey of Distributed Mutual Exclusion Algorithms". **Technical Report** CS-93-116, University of Colorado at Boulder, 1993.

[15] S.-D. Lin, Q. Lian, M. Chen, Z. Zhang, "A Practical Distributed Mutual Exclusion Protocol in Dynamic Peer-to-Peer Systems". **Proc. Third International Workshop on Peer-to-Peer Systems**, vol. 3279 of Lect. Notes Compo Sc., (La Jolla, CA). Springer-Verlag, Berlin, 2004.

[16] L. Sha, R. Rajkumar, J. P. Lehoczky, "Priority inheritance protocols: An approach to real-time synchronization". **Computers, IEEE Transactions on**, vol. 39(9), pp1175–1185, 1990.

[17] S. Greco, B. Matarazzo, R. Slowinski, "Rough sets methodology for sorting problems in presence of multiple attributes and criteria", **European Journal of Operational Research**, 2002, 138, pp. 247-259.

[18] X. Chao, G. Kou, Y. Peng, "An optimization model integrating different preference formats", **6th International Conference on Computers Communications and Control (ICCCC)**, 2016, pp. 228 - 231.

[19] D. L. La Red Martínez, J. C. Acosta, "Aggregation Operators Review - Mathematical Properties and Behavioral Measures", **International Journal of Intelligent Systems and Applications (IJISA)**, Hong Kong, 2015, 7, (10), pp. 63-76.

[20] D. L. La Red Martínez, N. Pinto, "Brief Review of Aggregation Operators", **Wulfenia Journal**, Austria, 2015, 22, (4), pp. 114-137.

[21] R. Yager, "On Ordered Weighted Averaging Aggregation Operators in Multi-Criteria Decision Making", **IEEE Trans. On Systems, Man and Cybernetics**, 1988, 18, (1), pp. 183-190.

[22] R. Yager, "Families of OWA Operators. Fuzzy Sets and Systems", 1993, 59, (2), pp. 125-148.

[23] R. Yager, Kacprzyk J.: "The Ordered Weighted Averaging Operators", **Theory and Applications, Kluwer Academic Publishers**, USA, 1997.

[24] R. Yager, Pasi, G.: "Modelling Majority Opinion in Multi-Agent Decision Making", **International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems**, 2002.

[25] Y. Dong, Zhang, H., Herrera-Viedma, E., "E. Consensus reaching model in the complex and dynamic MAGDM problem", **Knowledge Based Systems, Elsevier**, 2016, 106, pp. 206-219.

[26] D. L. La Red Martínez, "Aggregation Operator for Assignment of Resources in Distributed Systems." **International Journal of Advanced Computer Science and Applications (IJACSA)**. The Science and Information (SAI) Organization, England, U.K, 2017, 8, (10), pp. 406-419, ISSN N° 2156-5570.

[27] D. L. La Red Martínez, J. C. Acosta, F. Agostini, "Assignment of Resources in Distributed Systems". **9th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2018)**, Orlando, USA, 2018.

3D Printing of Personalized Archwire Groove Model for Orthodontics: Design and Implementation

Gang Liu[1]
State Key Laboratory of Solid Lubrication, Institute of Chemical Physics
Chinese Academy of Sciences, School of Information Science and Engineering Lanzhou University
Lanzhou,China

He Qin[2]
School of Information Science and Engineering
Lanzhou University, Lanzhou, China

Haiyan Zhen[3]
First Hospital of Lanzhou University
Lanzhou, China

Bin Liu[4]
School/Hospital of Stomatology
Lanzhou University
Lanzhou, China

Xiaolong Wang[*5]
State Key Laboratory of Solid Lubrication
Institute of Chemical Physics, Chinese Academy of Sciences
Lanzhou, China

Xinyao Tao[6]
Neoglory Holdings Group
Yiwu, China

***Abstract*—In traditional dental treatment, archwires are bent by orthodontists using standard methods. However, the standard models cater to patients with common oral problems, and are unsuitable for personalized orthodontic treatment, which is highly desired in many cases. A method to prepare a personalized archwire groove model is, undoubtedly, useful for orthodontic treatment in clinical diagnosis. In this study, a three-dimensional (3D) printing technology is demonstrated to achieve the personalized archwire groove model in a rapid, computed tomography image compatible manner, to assist orthodontists. This method is expected to improve the efficiency and accuracy of archwire bending and the resultant product can distribute the uniform dentofacial stress, improve the wearing comfort of the patient and further shorten the period of treatment and repair of the tooth.**

Keywords—3D Printing; personalized; archwire groove model; orthodontic treatment

I. Introduction

With the development of 3D printing technology, the clinical orthodontic effect can be better realized by constructing different individualized models. In orthodontic treatment, archwires are the vital and motivating parts of an orthodontic appliance. Wires are bent and attached to teeth to align them via elastic recovery [1], and they store and deliver power through the brackets and bands to the teeth and surrounding tissues. A good archwire forming technique is an essential part of quality orthodontic treatment [2]. Clinicians normally adopt standard procedures and bending methods causing problems such as poor correction effect and increasing consultation hours because of the mismatch between individual needs and standardized clinical techniques. Therefore, in terms of diagnosis [3], treatment planning and mechanical therapy, strictly following standardized procedures is not suitable nowadays for personalized precise treatment, which therefore, needs to be improved.

According to historical records [4,5], standard dental archwire has been studied for many years and several types have been formed which are commonly used in clinics, such as Vari-Simplex, Tru-arch, Bonwill-Hawley and Brader. The vital elements of an archwire are both its size and shape. It is simple and easy to classify the standard archwire into large, medium and small on the basis of its size [6], but this classification ignores the shape factor. Some drawbacks are found, on analyzing several cases, in the types of archwire mentioned: the Vari-Simplex arch, designed by Alexander, is not adaptive to special patients because of its inaccurate classification; the Tru-arch, designed by Roth, has an obvious danger of anterior dental arch expansion and posterior dental arch shrink in clinical applications, because it widens the front arch radian (especially in the premolar regions) and shrinks the posterior arch radian, which leads to a 3cm wide difference on each side; the Bonwill-Hawley arch, designed by Hawley, has two characteristics viz. one is a proper front radian for incisor and canine teeth and the other is by using line segments to represent premolars and molars. But, it also has very apparent disadvantages such as low coincidence rate with normal dental arch, as gleaned from a large amount of literature, lacks aesthetic appearance. Using the mathematical model of a triangular ellipse, Brader designed the Brader arch which is effective in narrowing the canine area during clinical application, according to available literature, because Brader only considered the size. It can be seen obviously, that there are still some problems in the existing standard archwires.

Meanwhile, it could be found, by considering orthodontic clinical cases over the years, that applying the same kind of archwire to all patients did not get satisfactory results, hence a

*Corresponding author

personalized approach should be adopted [7]. Some dentists found that it was important to choose individualized archwire according to the dental shapes of the patient, after straight-wire appliances had been used in the clinic for 20 years [8], failing which, it was easy to cause recidivism and loosening of the tooth. The size of the personalized archwire is not the only important concern; in addition, the chair-side time would be largely reduced during each return visit [9] and the time span between two return visits would be extended. Moreover, the holistic shift of the teeth is more effective and the occluding relation is more normal so that the alignment period would be reduced. In this way, the clinicians could quickly control overbite, reduce overjet and close the space[10,11], which would shorten the therapy time and spare dentists more time to make intense adjustment of the teeth and enhance the curative effect.

In recent years, with the rapid development of 3D printing, it is possible to convert a digital model into a solid one in a short time [12-14]. Using this kind of information engineering technology, clinicians can perform the study of real-time virtual correction effects on patients with malocclusion deformity and carry out orthodontic clinical practice [15]. Aided by 3D printing technology [16-20], clinicians can design and truly achieve the personalized archwire model with minimum difference.

The purpose of this study is to construct and print a personalized archwire groove model, using 3D printing technology, which can assist orthodontists in shaping personalized archwire rapidly thus improving the efficiency and accuracy of archwire bending. In addition, this method can distribute the uniform dentofacial stress and improve the wearing comfort of the patient and further shorten the period of treatment and repair of the tooth.

II. Materials and Methods

A. Materials

A male patient, who had voluntarily joined the study and given prior consent was selected as the experimental subject. A CBCT (cone-beam computerized tomography) scanner (KaVo 3D eXam 5, KaVo Corporation, Germany, layer thickness 0.25 mm, exposure time 14.7 s) was used to obtain several DICOM (Digital Imaging and Communications in Medicine) formats of the lower jaw teeth data of the patient. The “Import Images” tool of Mimics software was then used to import these DICOM format files and from the original images, graphs of coronal plane and sagittal plane were automatically calculated and generated through Mimics. The location of these pictures is shown by Mimics using three views, which are related to each other and can be rapidly located through the mouse and location toolbar. The corresponding tissue pixels were extracted through the threshold, and put in a mask (Mask) and the Draw and Erase tools were utilized to edit and modify the mask, so as to extract the required image of the teeth and jaw bone tissue. Subsequently, an STL (Stereography) file format of 3D model was generated by using Export Binary STL command, i.e. 2D scan images were converted into 3D entities successfully.

The constructed STL format 3D mandible model was loaded into Geomagic where the repair of the model was carried out. The mandible model after filling the hole approximately presented a rough model, on which the fairing operation had to be performed by fast fairing command. After completion of the editing processing of the polygon, the shape had to be refined with operations like detection of curvature, construction and editing patches, construction of the grid and fitting surface, which was used for reconstruction of Non-Uniform Rational B-Splines (NURBS) surface of the mandible. The building patches of the mandible were completed by performing the “upgrade / constraint” command based on the Andrews six keys theory of maxillofacial coordination [21, 22], landmarks were selected on the anlagen bone surface, and personalized archwire was then drawn to guide the orthodontic clinical work. The steps for the selection of a range of personalized archwire grooves are as follows:

1) Selection of the corresponding landmarks: In this study, the landmarks were selected mainly from one-third dental root, i.e., reference points of forming the archwire from the outside and inner part of the tooth jaw, in order to provide a reasonable selection range for orthodontists who traditionally selected landmarks on the crown surface of the molar and canine. The reference point from one-third dental root of the sixth teeth was selected and the horizontal plane parallel to it was constructed before adjusting the grid size of the plane from the angle of view for selecting the key points. The selected point data are as shown in Table 1.

TABLE I. Coordinate Values of Point Data

Coordinate values of point data on external maxillofacial (mm)		**Coordinate values of point data on internal maxillofacial (mm)**	
x	**y**	**x**	**y**
5.4850	46.3128	16.0828	48.7771
9.4565	34.7752	19.6157	39.0663
12.5626	26.3565	22.0526	30.6267
15.8220	19.3156	23.4347	23.3133
20.4598	13.1289	25.9774	17.6465
26.3904	8.7769	28.7833	13.7254
33.3653	7.0508	33.3608	11.4007
38.4456	6.6948	38.7544	12.1141
45.5535	8.0638	43.5308	13.1859
51.3331	11.3389	47.5262	17.3475
57.0771	16.7109	50.3400	22.4378
60.7823	23.6831	52.4883	30.0677
63.9181	31.4431	54.9923	38.0855
67.7258	43.3829	57.5778	46.4019

2) Determination of the range of personalized archwire groove

Research findings from relevant literature [23] have been referred to. The fitting curve of dental arch was drawn by the following mandibular Equation (1).

$$y = -3.0555 \times 10-8 \times x6 - 7.3778 \times 10-7 \times x5 + 4.0995 \times 10-5 \times x4 + 7.4784 \times 10-4 \times x3 + 2.162555 \times 10-2 \times x2 - 3.5621 \times 10-1 \times x - 6.65453 + 9.0892 \times 10-2 \times l \times x + 2.36991 - 1.9409 \times 10-1 \times l2 \quad (1)$$

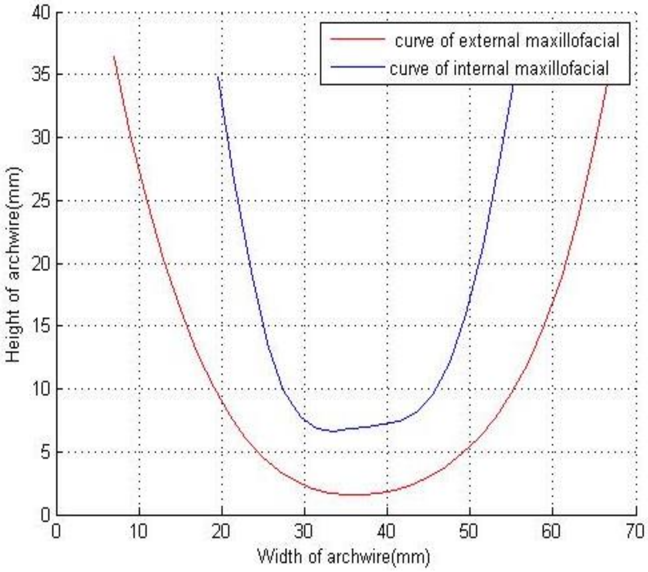


Fig 1. Selected Range Of Personalized Archwire Groove.

Where, l is the average width value of left and right central incisor crown.

The mesiodistal maximum diameter of mandibular central incisor crown from the patient's model was obtained by using dividers. If the values of the width of the left and right central incisor crown were obviously inconsistent, the mean value was taken. When the coordinate values of point data, as shown in Table 1, were imported into the computer, the output of the dental arch i.e. the range of personalized archwire groove was acquired immediately, as shown in Fig. 1.

B. Methods

After the designed 3D solid model of the teeth of the patient was loaded into ANSYS, as shown in Fig. 2(a), the working plane with grid size was constructed, and key points were selected around the teeth in the range of the working plane shown in Fig. 2(b), after which they were connected by using the Spline through KPs command under Extrude tools, and the bow- shaped curve was obtained, as shown in Fig. 2(c).

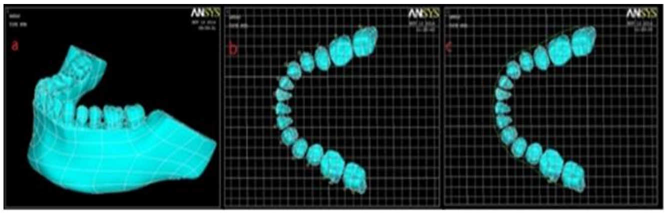


Fig 2. Construction of Personalized Archwire Groove (a) 3D Solid Model; (b) Key Points; (c) Formation of Archwire Line.

In the design of the archwire slot model to meet personalized needs, a key factor to be considered is the physical properties of wire materials. Orthodontic wires made of metal usually have good elasticity and rebound while being bent. The concrete shape of the archwire groove needs to be calculated according to the bow parameters, which would enable the archwire to gain a preset accurate shape after springback.

TABLE II. SPRINGBACK CURVE OF NICKEL TITANIUM ALLOY ARCHWIRE

Bending angle (°)	**Spingback angle (°)**
0	0
20	6
40	7.4
60	9
80	10.5
100	12.7
120	15

Nickel titanium (NiTi) alloy was used to implement the subsequent archwire bending experiment, because this material has excellent mechanical properties including super elasticity, shape memory effect and damping characteristics. Its elastic limit is far greater than ordinary materials and it does not obey Hooke's law during increasing stress range of deformation with the increase of strain, therefore, it is commonly used in making orthodontic wires.

In order to measure the springback angle of archwires made of this material at different bending angles, an experiment was specially designed. The archwire was arranged and fixed on a base plate and vertical pressure was applied on it at a distance of 2 mm on the right side of the base plate where θ, the bending angle, is the angle between the tangent line of bending arc and horizontal line before bending and α, the rebound angle, is the angle between the tangent line of bending arc before and after the archwire rebound.

The experimental results (as shown in Table 2) indicate that the springback angle increases with the increase of bending angle when the bending angle is greater than 10 degrees, which is close to the law of linear growth.

Combined with the data of dental arch of patients already obtained, some discrete key points at intervals of 2 mm were selected from the arch curve and the angle composed of three adjacent discrete points, which was the plastic deformation angle, was calculated. The bending angle based on the rebound curve of archwire and the plastic deformation angle which is the difference between bending angle and the springback angle, were also worked out. Through coordinate transformation of all angles, the key points of the personalized archwire groove were finally determined and the archwire thread based on these key points was drawn.

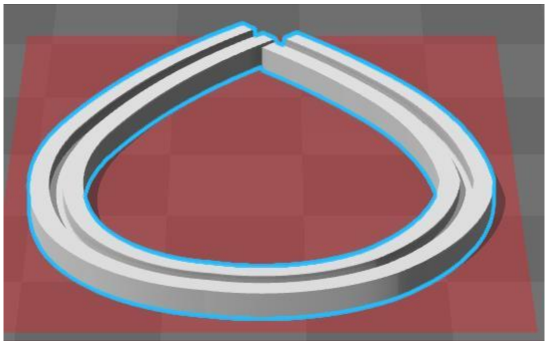


Fig 3. 3D Solid Model of Personalized Archwire Groove.

In this way, a 3D solid model of an archwire groove was shaped (as shown in Fig. 3) after construction of the working plane based on key points at the end and drawing the archwire groove on it.

III. Results and Discussion

The personalized archwire groove model was printed out by the 3D printer (MakerBot Company, American) and the archwire was bent by using this groove model (as shown in Fig. 4).

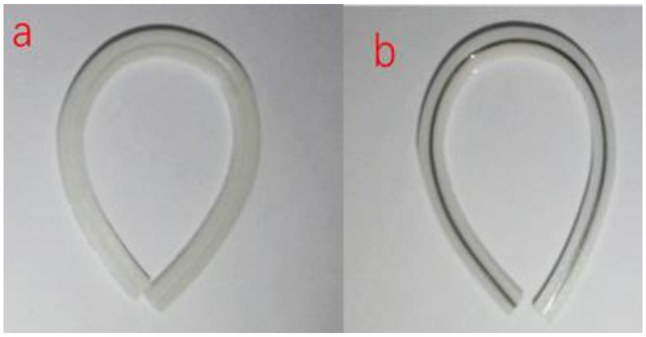


Fig 4. (a) 3D Printed Archwire Groove Model and (b) Bent Archwire.

Meanwhile, a wire bent by the archwire groove model was compared with that bent by the standard drawing method, by using a wire of 0.016” diameter. Fig. 5(a) represents the archwire of the test group using the method illustrated in this study, and Fig. 5(b) represents that of the control group using the standard method. It is obvious that the opening degree of the wire in Fig. 5(a) is smaller than that in Fig. 5(b), and is closer to the actual oral situation in patients.

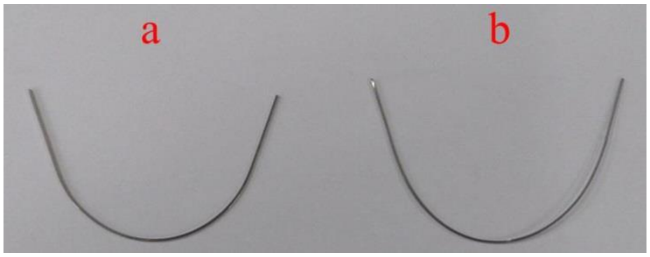


Fig 5. Archwire Bent with (a) 3D Printed Model and (b) Standard Method.

In order to further verify the advantages of personalized archwire groove model, the author specially displayed the individual dental arch model of the patient and the standardized dental arch model in the same coordinate system and carried out the test of goodness of fit by means of Equation (2). The goodness of fit values are shown in Table 3.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

Where, yi is the value of dental arch in the patient; y is the average value of dental arch in the patient; ŷi is value of the different kinds of archwires.

It can be seen from Table 3 that the dental arch form in the test group was similar to that of patients in the control group.

TABLE III. Goodness of fit Values

Group	Goodness of fit values between the actual archwire and group
a group	0.999446
b group	0.994317

To summarize, the results showed that the idea of designing a personalized archwire groove was viable. In theory, the performance and feasibility of the archwire groove as an application of 3D printing technology in orthodontics was tested.

The method based on Mimics and ANSYS software was developed specifically for designing the archwires groove. There were several drawbacks in the software that need to be addressed. Firstly, the complete archwires groove design process depended on commercial software (Geomagic and Mimics) and required human interaction, which reduced its efficiency. In future, special software should be developed to enable an automated design process. Secondly, the threshold for preprocessing DICOM format teeth data required to be manually set and adjusted; in addition, although the collection of key points to form the archwire could be selected manually in the visual condition, the accuracy needs to be improved.

However, the key point of orthodontic digitization research was its application to clinical orthodontics after obtaining the complete digital information. In previous research, the arrangement of the tooth crown was a more developed three-dimensional way of arrangement of teeth, and almost every commercial orthodontic virtual treatment software and the arrangement adopted in the present study were based on this, but the ideal position of the root of the tooth was not considered in this method, in other words, the key factor that influenced the stability and functionality was ignored. The concept of the treatment was limited to the tooth crown, but the relationships between the root of the tooth and supporting bone, the tooth and lip, as well as other soft tissues were ignored. The application of the integrated digital model data which includes the information obtained from cone beam computed tomograms (CBCT) of the tooth root and jawbones can deal effectively with many problems brought on by the previous virtual arrangement of teeth, avoiding many problems such as periodontal fenestration defects and fracture, which result in the process of aligning the dentition to be more applicative to the anatomical physiology characteristics of humans.

IV. Conclusions

In conclusion, the methods of personalized archwire groove model design were analyzed and explained, which examined its substantial role in further promoting 3D printing technology in the application of orthodontic treatment. However, the design result needs to be precisely modified using the tools available in the two software packages to obtain a satisfactory shape. In addition, the entire 3D printing process was completed through human-computer interaction. Therefore, further research is necessary in automated design software to make archwire groove model design both simpler and faster.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Gansu Province, China (145RJZA126, 1606RJZA051 and 17JR5RA318), the Fundamental Research Funds for the Central Universities (grant number lzujbky-2015-102), and the Western Project of CAS.

REFERENCES

[1] R.J. Nikolai, Orthodontic wire: a continuing evolution. Semin. Orthod. Vol. 3. pp. 157-165,1997 .

[2] J. Ferčec, I. Anžel, R. Rudolf, Stress dependent electrical resistivity of orthodontic wire from the shape memory alloy NiTi, Mater. Des. Vol.55,pp.699-706,2014 .

[3] S.M. Castro, M.J. Ponces, J.D. Lopes, M. Vascomcelos, M.C.F. Pollmann, Orthodontic wires and its corrosion-The specific case of stainless steel and beta-titanium, J. Dent. Sci. Vol.10, pp.1-7, 2015.

[4] E.A. Begole, Computer based methodology for construction of orthodontic arch wire templates, Comput. Prog. Bio. Vol.19, pp.61-68, 1984.

[5] X. Li, J. Wang, E.H. Han, W. Ke, Influence of fluoride and chloride on corrosion behavior of NiTi orthodontic wires, Acta Biomater. Vol. 3, pp. 807-815, 2007 .

[6] A.A. Nasef, A.R. El-Beialy, Y.A. Mostafa, Virtual techniques for designing and fabricating a retainer, Am. J. Orthod. Dentofacial. Orthop. Vol.146 , pp.394-398, 2014 .

[7] F. Yuan, Y. Sun, Y. Wang, P. Lv, Computer-aided design of tooth preparations for automated development of fixed prosthodontics, Comput. Biol. Med. Vol.44, pp.10-14 ,2014 .

[8] G. N. Boone, Archwires designed for individual patients, Angle Orthod. Vol.33,pp.178-185, 1963 .

[9] M. Salmi, K.S. Paloheimo, J. Tuomi, J. Wol, A. Makitie, Accuracy of medical models made by additive manufacturing (rapid manufacturing) J. Cranio. Maxill. Surg. Vol.41,pp.603-609, 2013 .

[10] S. C. Ligon, R. Liska, J. Stampfl, M. Gurr, and R, Mülhaupt, cy of medical models made by additive manufacturing (rapid manufacturing) J. Cranio. Maxill. Surg. he Cent

[11] T.M. Rankin, N.A. Giovinco, D.J. Cucher, G. Watts, B. Hurwitz, D.G. Armstrong, Three-dimensional printing surgical instruments: are we there yet? J. Surg. Res. Vol.189, pp.193-197, 2014 .

[12] B. Berman, 3-D printing: The new industrial revolution, Bus. Horiz. Vol.55, pp. 155-162, 2012.

[13] Y. Zhang, S. Zuo, J. Jiang, Y. Liu, Y. Liu, Interactive adjustment of individual orthodontic archwire curve, Chin. J. Sci. Instrum. Vol.38, pp.1616-1624, 2017 .

[14] D.S.C. Soon, M.P. Chae, C.H.C. Pilgrim, W.M. Rozen, R.T. Spychal, D.J. Hunter-Smith. 3D haptic modelling for preoperative planning of hepatic resection: A systematic review, Ann. Med. Surg. Vol.10, pp. 1-7, 2016 .

[15] Birtchnell T, Urry J. 3D, SF and the future, Futures, 50 (2013) 25-34.C. Growth, N. D. Kravitz, and J. M. Shirck, "Incorporating three-dimensional printing in orthodontics", J Clin Orthod. 2018, pp. 28-33, 2018.

[16] A. Sutradhar, J. Park, D. Carrau, M.J. Miller, Experimental validation of 3D printed patient-specific implants using digital image correlation and finite element analysis, Comput, Biol. Med. Vol.52, pp. 8-17, 2014 .

[17] H.T. Yau, T.J. Yang, Y.C. Chen, Tooth model reconstruction based upon data fusion for orthodontic treatment simulation, Comput. Biol. Med. Vol.48, pp.8-16, 2014 .

[18] M. Gebler, A.J.M.S. Uiterkamp, C. Visser, A global sustainability perspective on 3D printing technologies, Energy Policy . Vol.74, pp.158-167, 2014.

[19] S. Bose, S. Vahabzadeh, A. Bandyopadhyay, Bone tissue engineering using 3D printing, Mater. Today. Vol.16, pp. 496-504, 2013.

[20] G. Chen, W. Fan, S. Mishra, M.J. Miller, Tooth fracture risk analysis based on a new finite element dental structure models using micro-CT data, Comput. Biol. Med. Vol.42, pp. 957-963, 2012 .

[21] P. Shah, and B. S. Chong, , M.J. Miller, Tooth fracture risk anal planning in Endodontics", Clinical oral investigations. pp. 1-14, 2018.

[22] D. Rejeski, F. Zhao, Y. Huang, . Miller, Tooth fracture risk anal planning in Endodontics", Clinical oral investigations. pp. 1-14, 2018.T data, Comput. Biol. Med. a

[23] J. Qiu, L. Han, J. Li, H. Li, The study of mathematics model of customized dental arch in Han race of young adult in Tianjin, Chin. J. Prosthodont. Vol.15, pp. 6-10, 2014.

Software vs Hardware Implementations for Real-Time Operating Systems

Nicoleta Cristina GAITAN[1], Ioan Ungurean[2]
Faculty of Electrical Engineering and Computer Science, Stefan cel Mare University of Suceava
Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies,
and Distributed Systems for Fabrication and Control (MANSiD)
Suceava, Romania

***Abstract*—In the development of the embedded systems a very important role is played by the real-time operating system (RTOS). They provide basic services for multitasking on small microcontrollers and the support to implement the deadlines imposed by critical systems. The RTOS used can have important consequences in the performance of the embedded system. In order to eliminate the overhead generated by RTOS, the RTOS primitives have begun to be implemented in hardware. Such a solution is the nMPRA architecture (Multi Pipeline Register Architecture - n degree of multiplication) that implements in hardware of all primitives of an RTOS. This article makes a comparison between software RTOS and nMPRA systems in terms of response time to an external event. For comparison, we use three of the most commonly used RTOS in developing embedded systems: FreeRTOS, uC/OS-III and Keil RTX. These RTOSs are executed on a microcontroller that works at the same frequency as the implementations of the nMPRA architecture on a FPGA system.**

Keywords*—*Embedded system; real time operating systems; microcontrollers; FPGA

I. Introduction

Real-time operating systems (RTOSs) [1] are very important at the development of software applications for embedded systems. They allow the modular design and development of the software application and ensure the deadlines imposed by the critical systems. Two very important features of these operating systems are the predictability and the reliability. The RTOS systems are very important in developing software applications for embedded systems because they provide basic services such as multitasking services for synchronization and communication between tasks [2]. Usually, these RTOSs are used for microcontrollers (MCUs) or processors that do not use virtual memory and that have limited code and data memory [3].

In the selecting of an RTOS for embedded application development, several parameters are considered. For critical applications, there are very important the reliability, predictability and the responsive time to the internal or external events (Worst Case Execution Time – WCET). It should be specified that the response time depends on the implementation mode in RTOS but is closely related to the frequency of operation of the MCU and the way of the assignment of the priorities to these events. In addition to these parameters, licensing costs, code and data memory requirements, RTOS overhead and the expertise/experience in using an RTOS used in previous projects are considered [4].

In an embedded market survey published in 2017, 67% of the embedded projects in progress in 2017 use a form of operating system (RTOS, kernel, software executive). Those who do not use an operating system have specified that they do not need it because the applications being very simple and there are not real time application. This study shows a growing trend in the utilization of the open source operating systems and a downward trend in the utilization of the commercial operating systems from 2012 to 2017. The main reason for choosing a commercial operating system is the real-time capabilities [5].

The most used operating systems in the ongoing projects in 2017 are Embedded Linux (22%), FreeRTOS (20%), OS developed in house (19%), Android (13%), Debian (13%), Ubuntu (11%), Window Embedded 7 (8%), Texas Instruments RTOS (5%), Texas Instruments (DSP/BIOS), Micrium (uC/OS-III) Windows 7 Compact or earlier (5%), Keil (RTX) (4%), Micrium (uC/OS-II) (4%), Wind River (VxWorks) (4%). It can be noticed that for small microcontrollers, with low code and data memory and without virtual memory that can be used to develop hard real-time applications (such as those based on ARM Cortex Mx or ARM Cortex Rx), the most used RTOS are FreeRTOS, Texas Instruments RTOS, Micrium uC/OS-III and uC/OS-II. On the systems that use microprocessors virtual memory systems (such as those based on ARM Cortex Ax), it is possible to develop only the soft real-time applications, which are of the best effort type [5].

Typically, embedded applications are a combination of hard real time and soft real time tasks. In order to help the embedded software developers, MCU manufacturers have begun to provide solutions with one or more MCUs for hard real time tasks (for example ARM Cortex M0) and one or more MCUs or microprocessors (such as ARM Cortex A9) for soft real time application. Examples of such solutions are Sitara System on Chip from Texas Instruments or i.MX 7 Series Application Processors from NXP [6][7].

Since the development of FPGA systems, solutions for the implementation of RTOS in hardware have begun to be designed and developed in order to eliminate the overhead generated by the software RTOSs. These systems are experimental, not being widely adopted by the embedded developers. An example of such system is the nMPRA architecture presented in [8].

This paper presents a comparison of the performances obtained by the real time operating systems implemented in the software (FreeRTOS, Keil RTX and uC/ OS-III) and the nMPRA architecture that implements in hardware the primitives of a real-time operating system [8]. The comparison is performed in terms of the response time of the task with the highest priority at the trigger of an event expected by this task.

Further, this paper is structured as follows: Section II presents a brief description about software RTOSs; Section III presents some solutions for RTOS primitives implemented in hardware, and in Section IV are presented the experimental results achieved. The conclusions are drawn in Section V.

II. Software Implementation of the RTOSs

RTOSs are operating systems specially designed for embedded applications with real time requirements. For this reason, the overhead generated by the kernel execution is very important because it can interfere with response time to external events. Within these systems, the interrupts handling play an important role. In almost all RTOS interrupts are handled outside the kernel, the task being able to synchronize or receive messages from interrupt service routines [9].

According to the market study presented in, the most used RTOS for small microcontrollers is FreeRTOS. This is open source, providing basic services for multitasking, synchronization, and inter-task communication based on a preemptive kernel with static priorities. An analysis of the performances of this RTOS is presented in [12], where the evolution of FreeRTOS is followed over the course of 10 years. From this study, it can be concluded that FreeRTOS has improved its performance in terms real time facilities [5][10] [11].

Another RTOS widely used is Micrium uC/OS-III. This is a commercial RTOS designed to be used in embedded systems with hard real time requirements based on a preemptive kernel with static priorities. It provides all the services of a multitasking system in terms of synchronization and inter-task communication. It is provided as sources in ANSI-C being ported to a wide range of microcontrollers. His predecessor was Micrium uC/OS-II, which is still used in many projects [5] [9].

A royalty-free real time operating system based on CMSIS-RTOS API is Keil RTX. This was designed for applications based on microcontrollers. In the CMSIS-PACK package for Keil MDK is included the RTX kernel, along with source files and libraries. Keil RTX delivers benefits like task scheduling, multitasking, inter task communication, and shorter ISR system management [13].

A comparison of these RTOS in terms of time for task switching is presented in [3]. These comparisons are made by using three types of synchronization objects: events, semaphores, and mutex. The best performances are obtained for Keil RTX, followed by uC/OS-II and rt-thread. The lowest performance is obtained for FreeRTOS (time is approximately twice as high as Keil RTX). From these tests, it can be seen that the most widely used open source RTOS in embedded projects has much lower performance than commercial solutions represented by Keil RTX and uC/OS-II. This is because licensing costs can be an important criterion in selecting an RTOS.

III. Hardware Implementation of the RTOSs

With the development of FPGAs, a new trend has emerged through the hardware implementation of the primitives of an RTOS. Software RTOS treats interruptions outside the executive, interrupting any task running, which means that they can lead to missed deadlines when there is more than one interruption at the same time. To increase the predictability of the RTOS behavior, they can be implemented in hardware. To this end, the nMPRA architecture (Multi Pipeline Register Architecture - n degree of multiplication) was suggested which, together with nHSE (Hardware Scheduler Engine for n tasks), are a solution of a microcontroller with RTOS implemented in hardware. The nMPRA architecture, together with the nHSE module, is an innovative solution with a response time to external events of 1-3 processor cycles, which means a significant improvement over the software solutions of real-time operating systems or over those of software / hardware hybrids. The nMPRA architecture, defined in [8], uses a 5-stage implementation MIPS pipeline. This is a very strong architecture due to its proprieties, namely: switching between tasks is usually carried out in a single machine cycle or in a maximum of three machine cycles when working with global memory. The system’s reaction to an external event will not exceed 1.5 clock cycles if the event is attached to a higher priority task than the current task. The pipeline is not the reset; as a result, there is no need to restore/save the context, due to the multiplication of resources (PC, pipeline registers and registry file). It uses a powerful instruction through which a task can wait for various types of events (time, mutex, event, interrupt, timers for deadlines, etc.). The nMPRA is provided with distributed interrupt controller from which the interrupt inherits the task priority; it supports a static scheduling and has support for the dynamic scheduling of tasks. This architecture has been updated continuously [14][15][16].

An example of such operating systems is the ReconOS that is used for reconfigurable computing. This embedded operating system offers OS services for hardware and software execution. It provides a standardized interface that permit to include hardware accelerators. This solution is hardware-software co-design of a RTOS [17].

Another solution for OS hardware implementation is mosartMCU. It is implemented around a 32-bit RISC-V microcontroller and implements most of the OS directives in hardware. This solution achieves a lower response time for the interrupt handling [18].

μC /OS-III HW-RTOS is a hardware implementation of the μC/OS-III operating system. It implements in hardware the primitives of the μC /OS-III with a significant improvement in performance in terms of response time to internal and external events. It is implemented in R-IN32M3 from Renesas around an ARM Cortex M3 MCU [19].

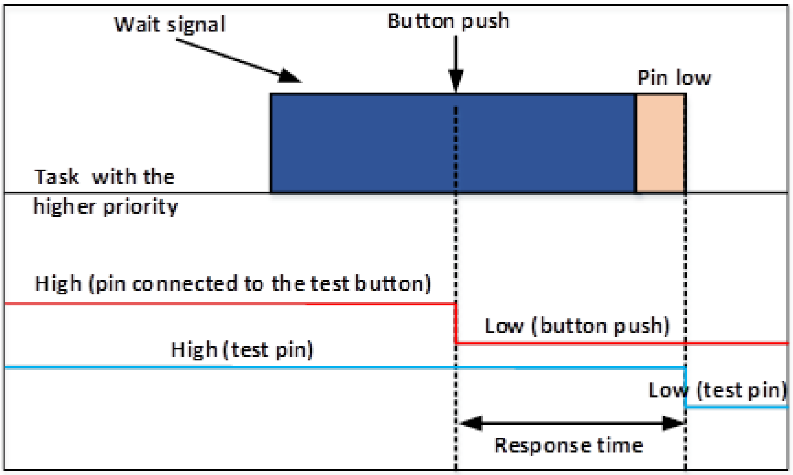


Fig. 1. Time Diagram of Test Applications.

In [20] the authors present the Real-Time Task Manager (RTM) that is an extension for the processor that aims to eliminate the overhead of RTOS primitives. This solution explore the execution in parallel and in hardware the RTOS primitives, function such as scheduling, synchronization primitives, and time management. This solution achieved a decrease of the response time by an order of magnitude.

IV. Performance Comparison

For the experimental tests, it used an implementation of the nMPRA with four task (sCPU) based on an MIPS processor at 33MHz on the Xilinx Zynq-7000 SoC ZC706 Evaluation Kit. For this reason, a microcontroller with the same operating frequency will be used for software operating systems. In this case, it has been used the STM32 NUCLEO-L053R system that is based on the STM32L053R8 ARM Cortex ™ -M0+ MCU, which can be programmed to operate at a frequency of 32MHz.

In this case, it is desirable to measure the response time to an external event. A button connected to a pin port of the microcontroller will generate the external event. This port will generate an external interrupt that must be handled by the task with the higher priority from the system. The higher priority task will remain blocked on the event, and when it goes into run state in response to the event, it will pass a port pin to low. Fig. 1 presents this mode of operation. It can be noted that the task with the highest priority is waiting an event. The event is triggered by pressing the test button (the port at which it is connected will go from high to low). After a time that depends on the operating system and the scheduler operation, the task with the highest priority will pass a second test pins port to the low. Thus, using a two-probe oscilloscope connected to the two ports, the response time to the external event can be measured.

In order to implement these performance tests, we will use three software RTOSs: uC-OS/III, Keil RTX, and FreeRTOS. These are the most used RTOSs for small MCU systems according to the study presented in [5]. The chosen RTOSs systems do not allow the direct attachment of an interrupt to a task, the interrupt service routines being executed outside the kernel. In the present case, for each RTOS, a multi-task application has been developed in which, after the initialization part, the task with the higher priority enters in the waiting state for an event. In the external interrupt service routine for the port that is connected to the test button, the expected event is triggered and the scheduler is executed. The scheduler will select the task with the higher priority for execution. This task will pass to low the second test pin port as soon as it enter in execution state.

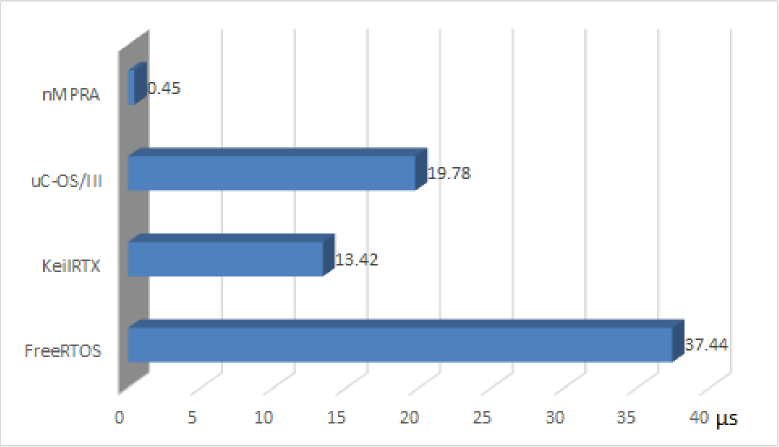


Fig. 2. Response Time to a External Event.

In order to implement this performance test on nMPRA, the external interrupt associated with the pin connected to the test button is attached to the task with the higher priority. After the initialization part, the task with the higher priority will wait the occurrence of an attached event (in this case the external interrupt). When the button is pressed, the external interrupt is triggered and the hardware scheduler will select the task with the higher priority for the execution. This task will pass to low the second test pin port as soon as it enters in execution state.

The results for the four tests (three software RTOSs and one hardware RTOS) are shown in Fig. 2. It can be noticed that the response time for the hardware RTOS is very small compared to software RTOS because the overhead generated by a software RTOS has been eliminated. Although hardware RTOSs have very good performances, they are not used in the industry because they are specialized RTOSs where programming is different and the users are very reluctant, preferring to use software RTOSs that have demonstrated their reliability in previous projects.

V. Conclusions

In this paper, it was presented a comparison between three software RTOSs and one hardware RTOS hardware in terms of response time to an external event. From this comparison, it was observed that the shortest response time is obtained by the RTOS implemented in hardware. This happens because the hardware implementation eliminates the overhead generated by software RTOS. From the tests it can be noticed that the weakest performances are obtained by the open source RTOS. However, FreeRTOS is the most widely used RTOS in embedded projects on small microcontrollers in 2017. Although much better performance are obtained with hardware implementations, the software RTOSs are still used because they are more customized and they were tested in very many projects. They can also be executed on a wide range of microcontrollers.

Acknowledgment

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI-UEFISCDI, Contract no. 220PED⁄2017, PN-III-P2-2.1-PED-2016-1473, within PNCDI III.

References

[1] Zelenova, S.A. and Zelenov, S.V., 2018. Schedulability Analysis for Strictly Periodic Tasks in RTOS. Programming and Computer Software, 44(3), pp.159-169. https://doi.org/10.1134/S0361768818030076

[2] Wang, K. C. "Models of Embedded Systems." Embedded and Real-Time Operating Systems. Springer, Cham, 2017. 95-111.

[3] I. Ungurean and N. C. Gaitan, "Performance analysis of tasks synchronization for real time operating systems," 2018 International Conference on Development and Application Systems (DAS), Suceava, 2018, pp. 63-66. doi: 10.1109/DAAS.2018.8396072

[4] G. C. Buttazzo. “Hard Real-Time Computing Systems:Predictable Scheduling Algorithms and Applications”, Springer Science & Business Media, 2011.

[5] 2017 Embedded Markets Study, Integrating IoT and Advanced Technology Designs, Application Development & Processing Environments, https://m.eet.com/media/1246048/2017-embedded-market-study.pdf

[6] Sitara™ Processors, http://www.ti.com/processors/sitara-arm/overview.html

[7] i.MX 7 Series Applications Processors: Multicore Arm® Cortex®-A7, Cortex-M4, https://www.nxp.com/products/processors-and-microcontrollers/arm-based-processors-and-mcus/i.mx-applications-processors/i.mx-7-processors:IMX7-SERIES

[8] V. G. Gaitan, N. C. Gaitan and I. Ungurean, "CPU Architecture Based on a Hardware Scheduler and Independent Pipeline Registers," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 9, pp. 1661-1674, Sept. 2015. doi: 10.1109/TVLSI.2014.2346542

[9] LABROSSE, Jean J. uC/OS-III, The Real-Time Kernel, or a High Performance, Scalable, ROMable, Preemptive, Multitasking Kernel for Microprocessors, Microcontrollers & DSPs. Micrium Press, 2009, ISBN:0982337531 9780982337530.

[10] The FreeRTOS™ project website. http://www.freertos.org. Accessed 29 Oct 2018.

[11] FERREIRA, Joao F., et al. Automated verification of the FreeRTOS scheduler in Hip/Sleek. International Journal on Software Tools for Technology Transfer, 2014, 16.4: 381-397

[12] GUAN, Fei, et al. Open source FreeRTOS as a case study in real-time operating system evolution. Journal of Systems and Software, 2016, 118: 19-35.

[13] VIRUTHAMBAL, K., et al. RTOS Based Dynamic Scheduler in Power Quality Applications. International Journal of Scientific Engineering and Technology, 2013, 2.6: 554-559.

[14] I. Zagan and V. G. Găitan, "Implementation of nMPRA CPU architecture based on preemptive hardware scheduler engine and different scheduling algorithms," in IET Computers & Digital Techniques, vol. 11, no. 6, pp. 221-230, 11 2017.

[15] I. Zagan, V. G. Gaitan, "Improving the Performances of the nMPRA Processor using a Custom Interrupt Management Scheduling Policy," Advances in Electrical and Computer Engineering, vol.16, no.4, pp.45-50, 2016, doi:10.4316/AECE.2016.04007

[16] N. C. Gaitan, "Enhanced Interrupt Response Time in the nMPRA based on Embedded Real Time Microcontrollers," Advances in Electrical and Computer Engineering, vol.17, no.3, pp.77-84, 2017, doi:10.4316/AECE.2017.03010

[17] A. Agne et al., "ReconOS: An Operating System Approach for Reconfigurable Computing," in IEEE Micro, vol. 34, no. 1, pp. 60-71, Jan.-Feb. 2014. doi: 10.1109/MM.2013.110

[18] F. Mauroner and M. Baunach, "mosartMCU: Multi-core operating-system-aware real-time microcontroller," 2018 7th Mediterranean Conference on Embedded Computing (MECO), Budva, 2018, pp. 1-4. doi: 10.1109/MECO.2018.8406007

[19] Jean J. Labrosse, “Hardware-Accelerated RTOS: µC/OS-III HW-RTOS and the R-IN32M3”, https://www.micrium.com/hardware-accelerated-rtos-%C2%B5cos-iii-hw-rtos-and-the-r-in32m3/

[20] P. Kohout, B. Ganesh and B. Jacob, "Hardware support for real-time operating systems," First IEEE/ACM/IFIP International Conference on Hardware/ Software Codesign and Systems Synthesis (IEEE Cat. No.03TH8721), Newport Beach, CA, USA, 2003, pp. 45-51., doi: 10.1109/CODESS.2003.127525

A Two-Level Fault-Tolerance Technique for High Performance Computing Applications

Aishah M. Aseeri[1], Mai A. Fadel[2]
Faculty of Computing and Information Technology
King Abdulaziz University, KSA

***Abstract*—Reliability is the biggest concern facing future extreme-scale, high performance computing (HPC) systems. Within the current generation of HPC systems, projections suggest that errors will occur with very high rates in future systems. Thus, it is fundamental that we detect errors that can cause the failure of important applications, such as scientific ones. In this paper, we have presented a two-level fault-tolerance approach for the detection and classification of errors for Compute United Device Architecture (CUDA)-based Graphics Processing Units (GPUs). In the first level, it detects the existence of errors by using software redundancy that applies design diversity. In the second level, it investigates the problematic software version and re-executes it on a different hardware component to classify whether the error is a permanent hardware error or a software error. We implemented our approach to run on GPUs and conducted proof of concept experiments by running three versions of matrix multiplications with different error scenarios and results show the feasibility of the proposed approach.**

***Keywords*—High performance computing; fault tolerance; graphics processing units (GPUS); error detection; n-version programming (NVP); multi-GPU; reliability**

I. Introduction

High performance computing (HPC) is a term used in reference to integrated computing environments that rely on parallel processing in the running of applications. This boosts efficiency, speed, and reliability, while ensuring that complex scientific problems can be solver faster than if they were performed serially.

HPC systems are used to resolve complex scientific problems that, because of memory or computer performance limitations, either cannot be solved or are impractical to solve using traditional computing systems.

These systems promise to push the boundaries for scientists by augmenting their research across a range of disciplines, including: chemistry, nuclear physics, high energy, astrophysics, nanotechnology, biology, medicine, and material sciences [1]. However, to realize the full potential and reach the breakthroughs of this technology, software development tools are of great importance, such as compilers and debuggers; to be more specific test frameworks are among tools that should be part of the HPC infrastructure [2]. Test frameworks are becoming increasingly important as resilience is one of the major challenges to the growth of the complex systems mentioned above. System resilience is substantially reduced due to the increase in the number of components, regardless of the reliability and efficiency of individual components. Besides the addition of more components, many other factors increase the rate of failure for future HPC applications, including: number of components both memory and processors, smaller circuit sizes, heterogeneous systems, the number of operations, and increasing system and algorithm complexity [3]. This leads to the fact that hardware faults are becoming inevitable [4, 5] and the way is to be aware of and handle its effects [6]. From another point of view, as HPC power is targeting applications beyond the graphics domain, such as scientific applications and stock markets, it faces the challenge of addressing the need to generate accurate results that should be free of errors, as these applications cannot tolerate the existence of errors as graphical applications [7]. Hard errors are not the only concern of the HPC community, soft errors are a concern as well [8]. In [9] a study done on the data of two large-scale sites of a set of systems showed that hardware and software errors covering a considerable large proportion of root causes of failures. Hence, it is imperative to provide effective fault-tolerance capabilities, both at hardware and software levels as part of the test framework. HPC community has developed various solutions to generally tolerate faults, and more specifically to mitigate faults caused by hardware defects [10] and to detect and recover from errors [5, 11]. We will elaborate more on some of the relevant approaches in Section 2. Some of the used approaches depend on using checkpoints/reset [12], redundancy and Algorithm-Based Fault Tolerance ABFT [13, 14]. In our research, we have applied redundancy-based fault tolerance, as checkpointing has high communication overhead and ABFT is customized to fit the algorithm under analysis, thus, it is very difficult to generalize the solution to other applications without addressing the specifics of the new algorithm. In particular, we use software-based redundancy with design diversity; that is, we provide several versions of the same application that differ in their design to check for errors during execution time. Design diversity lessens the likelihood of having all versions fail exactly the same way in the same time. We use this technique in a broader view, as we aim to support the need to detect not only software errors but through these errors we can detect if the actual cause is a hardware error. This two-level approach starts by applying software-based redundancy with design diversity to identify the existence of a problematic copy of the software, then re-execute this copy on a different hardware to determine if the original hardware was the cause of the error or the software itself has an error.

In the following section, we present some of the research related to our work. In Section 3, we briefly describe some of the basics of CUDA-based GPU Architecture and Open MP programming, as they are the main tools of the infrastructure that we used to implement our system. In Section 4, we present our proposed methodology. In Section 5, we present the experimental results, and finally we conclude our paper in Section 6 and highlight some of future research directions.

II. Related Work

This section presents existing error detection techniques based on redundancy that are considered one of the protective techniques that provide resilient computing in the HPC domain. Many approaches based on hardware redundancy have been used successfully in mission-critical systems such as triple-modular redundancy (TMR) and dual-modular redundancy (DMR) [15]. The latter approach is achieved by supplying two similar physical components that can execute the same task. When an error occurs, the extra component transparently recovers from the peer one [16]. A TMR approach is based on three fully redundant components which perform the same process. The result is processed by a voting system to ensure the results are the same. If one component fails, the other two can correct and mask the fault. This approach causes performance overhead because of the need to synchronize original hardware and its replica and also doubles the hardware cost. In addition, running the same copy of the software on all components will not reveal an actual error, as all copies will generate the same incorrect result.

Design diversity among the software replicas is implemented as a solution for this problem. Thus, lessens the likelihood of having all copies failing on the same set of input data. In [17], this approach is categorized as software redundancy. This method has been widely exploited in targeting software errors, i.e., design faults or software bugs [18].

There are several approaches to software redundancy techniques, such as N-version programming [19], recovery blocks [20] and N self-checking [21]. Faults can be detected in these approaches by consistency checking/self-checking or time redundancy. Time redundancy is defined as running the same program several times and compare the results. All the above approaches target sequential applications, as for redundancy-based fault detection approaches that target concurrent applications running on GPUs can be found in [11, 22, 23, 24]. These approaches detect software errors whether they use software or hardware redundancy. In [25], the proposed system detects hardware errors using different types of redundancy.

From a different perspective, part or our method is to execute the problematic version of the software on different hardware to classify whether the error is caused by hardware or software. This idea has been applied in the SWAT tool [26] which will be discussed further in Section VI.

It is noticed that benefiting from software redundancy with design diversity is applied in several researches, to either detect software or hardware errors. However, its power has not been integrated with the step of classifying the error. Up to our knowledge no one applied design diversity in HPC for detecting errors and also no one used software redundancy for detecting hardware errors on GPUs.

III. Basic Concepts of Dependability

We now briefly present basic ideas and terminology used in the field of fault tolerance. A detailed background and taxonomy of the related terms can be found in [27, 28].

A. Fault-Error-Failure

A system is an entity that interacts with other entities. A system can be hardware based, for example a processor, or software based, such as a running application. A system consists of components which can be systems themselves. A system failure is defined as the deviation of the system behavior that is inconsistent with the system's specification. When the observed behavior differs from the specified behavior, we call it a failure. A failure occurs because of an error that is caused by a fault. An error is the part of the system state which results from the activation of a fault and causes the system to be in an illegal state. Errors are liable to lead to a failure. Fault propagation chain from faults to failures in a system is illustrated in Fig 1.

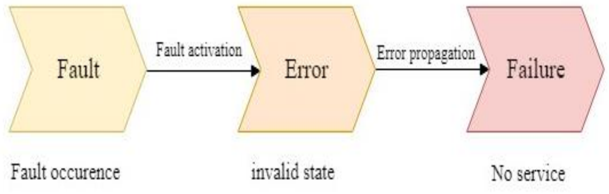


Fig. 1. Relationship between Fault, Error and Failure.

There are numerous sources of a fault that can be either software or hardware [29] as shown in Fig. 2.

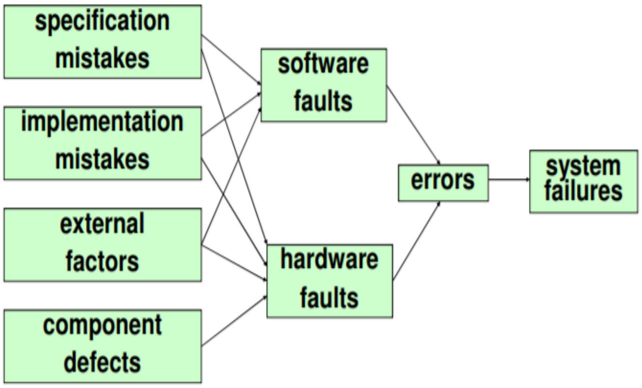


Fig. 2. Classification of the Sources of Faults [29].

Software faults are most often caused by design faults and operational faults [29]. Design faults occur when a designer, either misunderstands a specification or simply makes a mistake. Hardware faults are most often caused by incorrect specification, incorrect implementation, manufacturing imperfections or external factors.

System errors that impact the application's and supercomputer's reliability can be classified as "soft" or "hard": soft errors are usually caused by a transient fault and temporary environmental factors. Soft errors, unlike manufacturing or design faults, do not occur consistently. Some of the factors that can cause this type of faults are radiation-induced upsets in electronic circuits [27, 30], leakage from adjacent circuits, timing violations, and improper signal routing or power design [31]. These events do not cause permanent physical damage to the processor but can alter signal transfers or stored values and thus cause incorrect program execution.

By contrast, hard errors are caused by a permanent fault in the system and are usually caused by design faults or inherent manufacturing defects, thermal stress, wear out, and process variation. Permanent hard errors are easier to detect, because hardware deterioration is often irreversible, and their symptoms tend to be predictable and persistent over time. However, they must be detected because they present a threat to the application stability in a well-maintained environment [32]. Permanent faults usually require that the faulty component be avoided until it is repaired or replaced to avoid errors in system behavior. On the other hand, transient faults do not require repair/replacement of the component, but the impact of the resulting soft error needs to be masked.

B. Fault Tolerance

Fault-tolerance means the ability of a system to continue correct performance of its intended tasks and the ability to avoid failure after the occurrence of hardware and software errors. When a system is said to be fault-tolerant this means that the behavior of the external system is not affected by faults. A fault- tolerant system must be able to detect errors and recover from them.

IV. Overview the CUDA-based GPU Architecture and Openmp Programming

In this section, we give a brief description of the GPU architecture and CUDA, as the target applications that our tool analyzes are implemented using CUDA and Open MP and run on GPUs.

A. GPU Architecture

Fig. 3 illustrates a simplified overview of the GPU architecture. Modern GPU architecture is composed of an array of Streaming Multiprocessors (SM) [33]. The SMs are the main building blocks of a GPU. SMs consist of a set of Stream Processors (SPs) or CUDA cores, in which each core executes several threads in parallel at a specific time. SPs share control logic and an instruction cache, while SMs allow access to the global memory. In modern GPU devices, there are thousands of such SPs; this indicates that each GPU has the potential of executing thousands of threads at any moment. Moreover, each SM has shared memory and the L1 cache that is designed to improve the computational performance by storing the data common to the threads running on the SM.

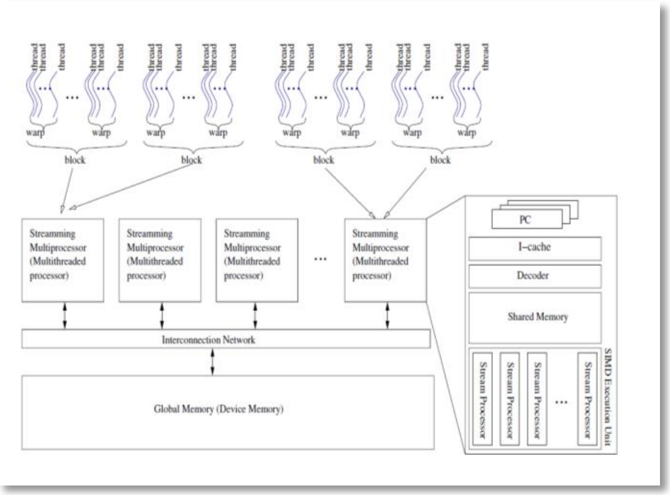


Fig. 3. An Overview of the GPU Architecture [33].

GPUs use this architecture in SIMT (Single Instruction Multiple Threads) [34], in which a group of (currently 32) threads known as a warp performs the same instruction. All the threads in one block are performed on one SM, or they can be implemented as multiple concurrently running blocks. The number of blocks that can be processed concurrently on one SM depends on the resource requirements of each block like shared memory usage and the number of registers.

There are many GPU programming languages that aim to provide an environment in which GPU and CPU programs can exist with each other. The main goal of these programming languages is to offload the GPU friendly portion of the program into the GPU memory. In this work, we use the CUDA programming language that is specifically employed for NVidia GPUs.

B. CUDA

CUDA (Compute United Device Architecture) is a parallel computing architecture developed by NVidia for massively parallel high-performance computing [35]. It can be accessible through CUDA-accelerated libraries, compiler directives, application programming interfaces, and standard programming languages including C, C++, Fortran, and Python. There are several programming models accessible to create program for GPU but CUDA by NVidia is the best option to accomplish parallelism through GPU processing.

Recently, this platform has proven successful in parallel computing architecture at programming multi-threaded on many-core GPUs .The GPU acts as a coprocessor that performs data-parallel kernel functions. CUDA has a hierarchy of thread groups. Threads are composed of a three level hierarchy. A grid consists of set of thread blocks that are responsible for executing a kernel function. Each block is composed of hundreds of threads. Threads inside one block have shared memory that allows sharing data. All threads within a block are executed concurrently on a multithreaded model.

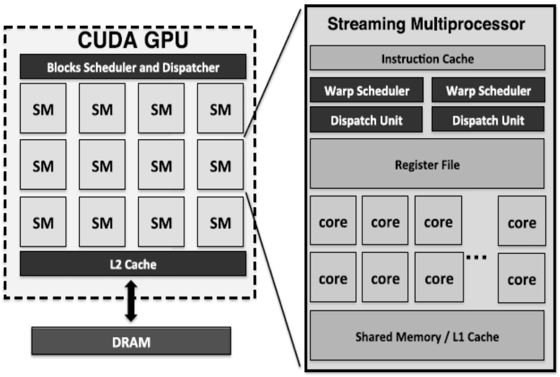


Fig. 4. A Representative CUDA-based GPU Architecture [36].

A CUDA-based system is a type of heterogeneous programming, since a program is usually running on two different platforms: a host and a device. The host system usually consists primarily of the CPU, main memory and its supporting architecture. The device generally includes the video card consisting of a CUDA-enabled GPU and its supporting architecture. The CPU begins to execute a CUDA program in order to provide inputs for the kernel and to start its implantation; this means providing a kernel grid to the GPU. The CUDA GPU begins the implantation of the kernel. Upon completion of a kernel implantation, the CPU can acquire the output data by accessing the contents of the GPU memory. The software organization of a kernel is related to the GPU architecture, since the threads hierarchy assigns immediately into the GPU internal components.

Fig. 4 illustrates CUDA GPU internal architecture. When the CPU starts to invoke a kernel grid, each thread block is assigned a Thread Block ID and dispatched to a SM that ensures enough available resources. Each thread of a thread block is executed on a CUDA core.

The programmer can specify the number of threads per block, and the number of blocks per grid. A thread in the CUDA programming language is characterized as much lighter weight than in traditional operating systems.

C. *OpenMP Programming*

Open multi-processing (OpenMP) is a programming model that has the ability to handle multithreading by computing in parallel modules. The basic idea of this programming model is that data are processed in parallel. It consists of a number of directives and libraries that are called runtime libraries [37]. The code inserted in these directives executes in parallel on multi-cores in the form of a basic OpenMP unit called "Thread" [38]. It also has the ability to process the looping region in a parallel way by adding compiler directives in the starting region of the OpenMP module that improve the efficiency of the program and overall application performance [39].

V. Proposed Method

In this research, we aim to detect hardware and software errors in CUDA applications that run on GPUs. Our proposed method consists of two levels of detection. The first level detects the presence of error in the results generated by the running software. The second level classifies the source of the error whether it caused by a hardware error or software error. We used multi-version programming at the first level of detection, where several versions having diverse designs of the same application are used. All versions of the software are executing in parallel. The correctness of the results is determined by running a voter in which common answers by the majority are considered the rightful result. In case the voter indicated the presence of an error in one of the versions for example, then the second level of detection is conducted. In this level, we reinvestigate the version that produced the incorrect result, by running it on a different set of cores or a different GPU with the same input data. The cause of the error is classified as hardware error in case the result is correct as stated by the majority and software error otherwise. The steps of the methods are illustrated in Fig. 5.

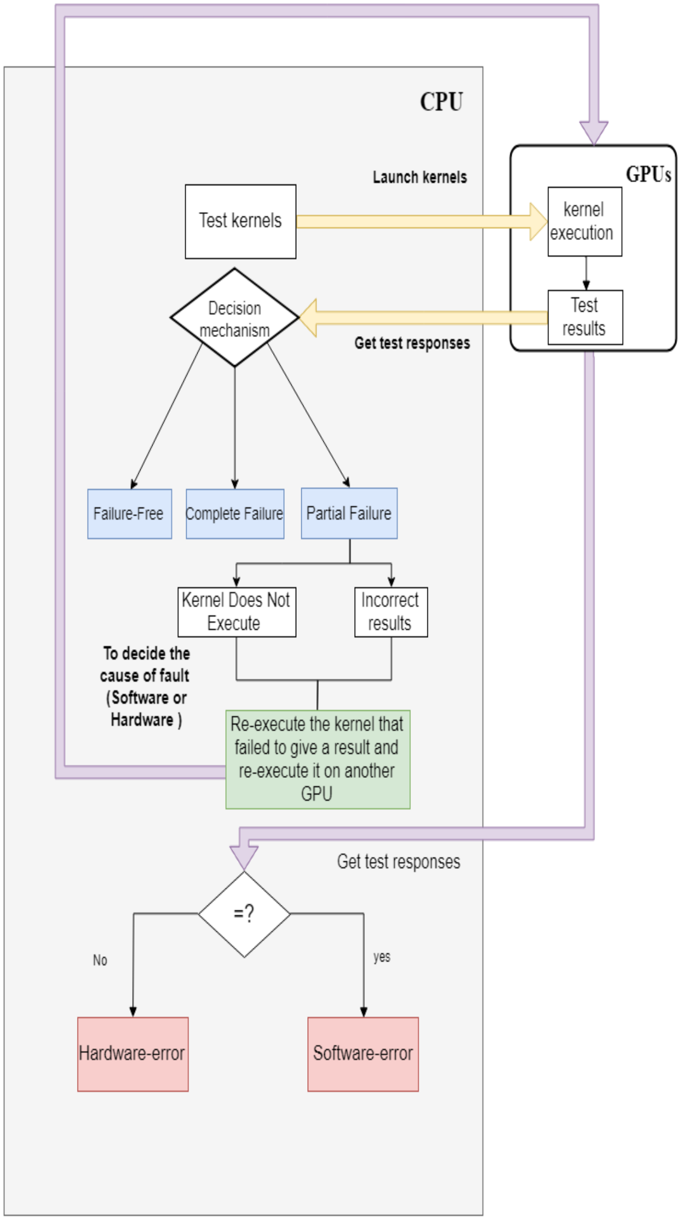


Fig. 5. Proposed Approach.

As can be seen, the different versions of the application, referred to as kernels, are executed on the GPU, whereas the control of the method steps are mainly done in the CPU, which are: launching the kernel, running the decision mechanism – the voter – and starting level two of the detection by re-executing the problematic kernel in case the voter's output indicates there is partial failure. Other possible output of the voter is that the software is error-free; i.e. all versions generated the same results, and complete failure; i.e. there is no agreement on the results among any of the versions. The figure also shows that partial failure can have the form of problematic interrupt of execution; i.e. a version of the application hangs, or the form of generating incorrect results by one of the versions.

VI. Experimental Results

In this section, we describe the experiments conducted to test the applicability of our method. First, we describe the system specifications on which the experiments are conducted and then we describe the application chosen to conduct the experiments. After that, we present the techniques used to inject faults in systems. In the following section, we describe the design of the experiments and show the results. Finally, we discuss the findings derived from our experiments.

A. System Specification

First, the hardware specifications of the system on which the experiments are conducted are listed. The machine contains a single Intel (R) Core (TM) i7-7700K CPU @4.20 GHz , equipped with three Nvidia GeForce GPUs: two of them are of the model GTX 1070 and the third GPU is of the model GTX1060. More details of the different GPUs are shown in Table I.

TABLE I. An Overview of the GPUs used. SM Denotes Streaming Multi-Processor

GPU Name	GTX 1060	2x GTX 1070
Architecture	Pascal	Pascal
# SMs	10	15
# cores/SM	1280	1920

Next, the software specifications are described. The machine runs Windows 10 as an operating system, and the development environment used is Microsoft Visual Studio community 2015 which as it is compatible with CUDA Toolkit 8.0 that we used to develop the different versions of the application – that is described in the next section. Simple visual studio C++, and OpenMP are also needed to develop the application and our tool.

B. The Application used for Testing

In this section, we describe the application we used to conduct our experiments on. We chose Matrix Multiplication as it is a computational mathematical operation that is widely used in the computational sciences in general, and scientific modeling in high performance computing domain as well [14, 22].

Several algorithms and mathematical formulas have been proposed to solve matrix multiplication, one of the proposed approaches exploits the massive parallelism of GPUs to speed up computations. Our method detects errors in parallel applications, thus, we have chosen three different parallel algorithms for solving matrix multiplication to conduct our experiments. The chosen algorithms show the design diversity required by the multi-version programming system. Next, we present the mathematical formula of matrix multiplication and then describe the three different algorithms used to solve this mathematical problem.

The mathematical formula for matrix multiplication is given in equation 1.

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k} * B_{k,j} \qquad (1)$$

Equation 1: general formula of Matrix Multiplication

Where A and B are matrixes of the sizes n×m and m×w respectively. C is a matrix of the size n×w that stores the product of matrix A and matrix B. For simplicity, we only created square matrices during our experiments.

The different algorithms chosen for matrix multiplications differ in the kind of memory being used, thus causing adequate changes in the design of each algorithm that are enough to introduce the design diversity required by our method – using different set of steps in each algorithm. The first algorithm used one thread to compute the result of an element of the matrix C. It depends on using global memory causing the performance to become relatively slow. The second algorithm uses shared memory to avoid unnecessarily accessing global memory multiple of times. The third algorithm is different from the second algorithm in that it transposes the second matrix, which is referred to as matrix B in (1).

C. Fault Injection

In our experiments, we need a procedure to inject faults and monitor the effect of these faults on system's behavior [40]. Fault injection is a widely used method for improving the reliability of applications. Reviews of fault injection techniques and methodologies in electronic and computer systems can be found in [18, 41]. Research has also been done to provide a framework to allow fault injection in HPC applications with the focus of facilitating designing complex experiments by defining workloads [42]. This framework, called FINJI, allows the integration of existing fault injection tools for heterogenous types of errors. Testing the detection of hardware and software errors requires fault injection of both types. Hardware errors will result in Silent Data Corruption (SDC) which is a kind of soft error that can simply be described as the flip of a bit or two in both kind of storage volatile and non-volatile [43]. Some of the approaches used to inject hardware errors are FPGA-based fault injection [44] and simulations to conduct microarchitecture-level fault injection [26]. The latter has been applied in a multicore environment called mSWAT for detecting hardware errors [45]. The idea is to detect hardware errors via software anomalies such as fatal-traps and system hangs, these detected errors are then diagnosed to identify which part pf the micro-architectural of the system is the source of the error as described in [46].

FPGA-based fault injections are performed on gate-level models and accelerated by FPGA-based hardware emulation. This approach injects errors at gates based on a user-provided model of hardware design. Another tool that injects errors at gates is Argus [47]. In our tool, there is no need to follow any of the above-mentioned hardware fault injection approaches, as our tool satisfies the objective of the research by reporting back the faulty core without specifying the kind of hardware error. Whereas the other approaches aim to identify the source of the error, for example, Argus, based on running certain instruction, it identifies the source of the error in a simple core. The research applying FPGA-based fault injection is measuring the accuracy of detecting the SDC that results from the specific types of hardware errors.

As for software fault injection, we have performed fault injection at the source code level, by conducting mutation of the source code of the application being analyzed and observing the outcomes. This kind of injection has been applied in other research such as [40, 48, 49].

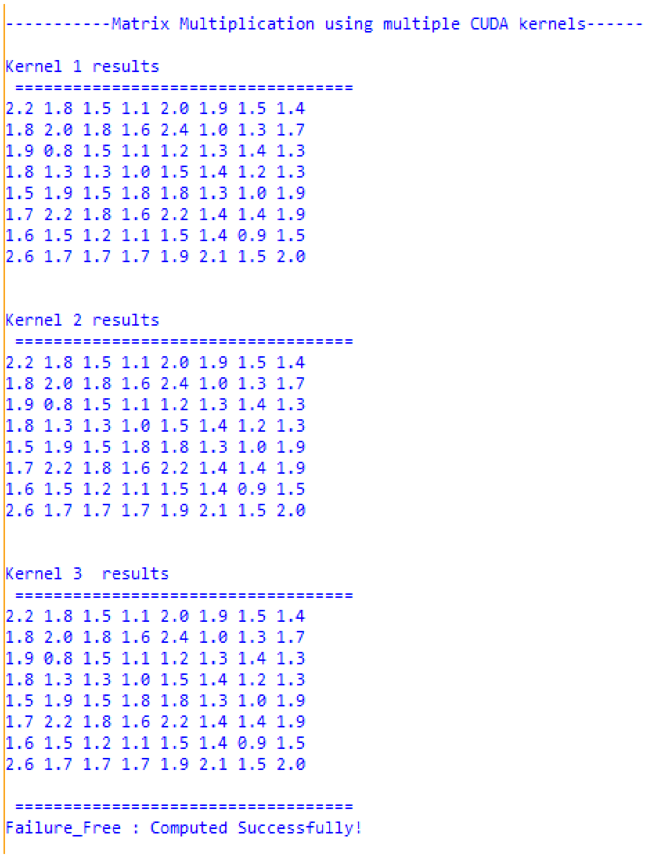
D. Experiments Design and Results

In order to test the applicability of our method, we need to ensure that the method is able to report back error-free, partial failure and complete failure cases. In this section, we focus on the error-free and partial failure cases as the complete failure case can be tested in the same way we test partial failure. In addition, to summarize the results, we report back the partial failure case in which the problematic version of the application generates in correct results. In addition, for the partial failure case, we conduct an experiment to detect hardware errors and another experiment to detect software errors. Soft errors are injected by changing the one or more of the code instructions to generate incorrect result. Hard errors are assumed they exist in one of the GPUs and we designed a method that returns an incorrect result in both levels of our detection method. In the following, we present the result of each experiment, and then present a table showing relevant measurements:

Fig. 6 shows the result of executing the three algorithms in which all are error-free, consequently depicting the error-free case:

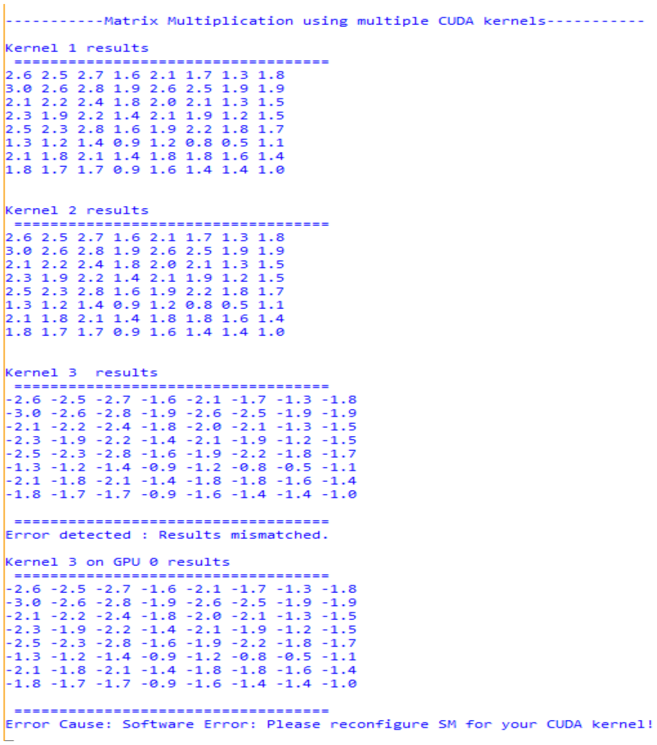
Fig.7 shows the result of executing the three algorithms, where one of the algorithms has an injected soft error, thus generating incorrect results. This depicts the case of partial failure caused by a soft error. In this case, the second level of the detection method is used to determine that it type of error is a soft error, since re-executing the algorithm on a different GPU generated an incorrect result as well.

Fig.8 shows the result of executing the three algorithms, where one of the algorithms generates in correct results (as returned by our designed method that mimics hard errors as described in the fault injection section). This depicts the case of partial failure caused by a hard error. In this case, the second level of the detection method is used to determine that the type of error is a hard error, since the algorithm generated a correct result when run on a different GPU.



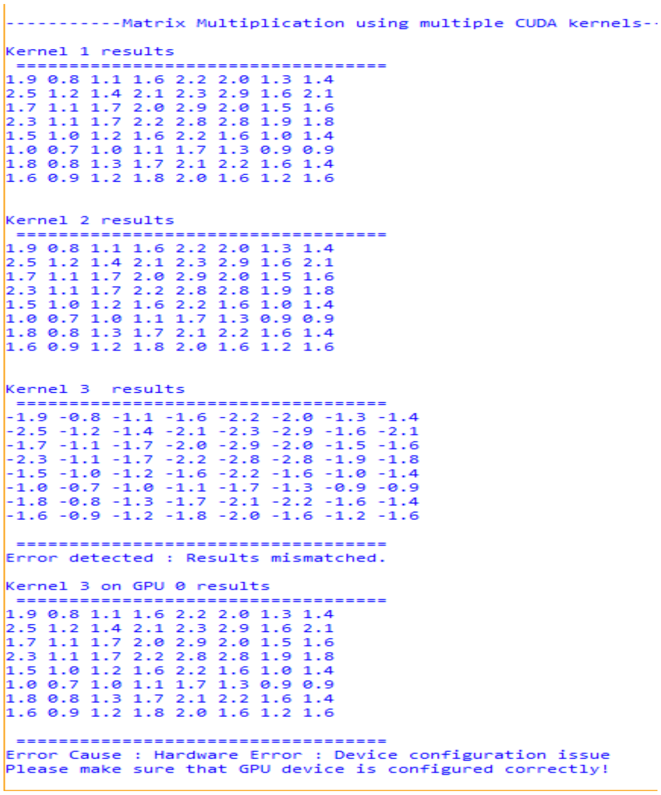
```
----------Matrix Multiplication using multiple CUDA kernels------

Kernel 1 results
 ==================================
2.2 1.8 1.5 1.1 2.0 1.9 1.5 1.4
1.8 2.0 1.8 1.6 2.4 1.0 1.3 1.7
1.9 0.8 1.5 1.1 1.2 1.3 1.4 1.3
1.8 1.3 1.3 1.0 1.5 1.4 1.2 1.3
1.5 1.9 1.5 1.8 1.8 1.3 1.0 1.9
1.7 2.2 1.8 1.6 2.2 1.4 1.4 1.9
1.6 1.5 1.2 1.1 1.5 1.4 0.9 1.5
2.6 1.7 1.7 1.7 1.9 2.1 1.5 2.0


Kernel 2 results
 ==================================
2.2 1.8 1.5 1.1 2.0 1.9 1.5 1.4
1.8 2.0 1.8 1.6 2.4 1.0 1.3 1.7
1.9 0.8 1.5 1.1 1.2 1.3 1.4 1.3
1.8 1.3 1.3 1.0 1.5 1.4 1.2 1.3
1.5 1.9 1.5 1.8 1.8 1.3 1.0 1.9
1.7 2.2 1.8 1.6 2.2 1.4 1.4 1.9
1.6 1.5 1.2 1.1 1.5 1.4 0.9 1.5
2.6 1.7 1.7 1.7 1.9 2.1 1.5 2.0


Kernel 3  results
 ==================================
2.2 1.8 1.5 1.1 2.0 1.9 1.5 1.4
1.8 2.0 1.8 1.6 2.4 1.0 1.3 1.7
1.9 0.8 1.5 1.1 1.2 1.3 1.4 1.3
1.8 1.3 1.3 1.0 1.5 1.4 1.2 1.3
1.5 1.9 1.5 1.8 1.8 1.3 1.0 1.9
1.7 2.2 1.8 1.6 2.2 1.4 1.4 1.9
1.6 1.5 1.2 1.1 1.5 1.4 0.9 1.5
2.6 1.7 1.7 1.7 1.9 2.1 1.5 2.0

 ==================================
Failure_Free : Computed Successfully!
```

Fig. 6. A Screenshot of the Result in the Case of Failure-Free.



```
----------Matrix Multiplication using multiple CUDA kernels----------

Kernel 1 results
 ==================================
2.6 2.5 2.7 1.6 2.1 1.7 1.3 1.8
3.0 2.6 2.8 1.9 2.6 2.5 1.9 1.9
2.1 2.2 2.4 1.8 2.0 2.1 1.3 1.5
2.3 1.9 2.2 1.4 2.1 1.9 1.2 1.5
2.5 2.3 2.8 1.6 1.9 2.2 1.8 1.7
1.3 1.2 1.4 0.9 1.2 0.8 0.5 1.1
2.1 1.8 2.1 1.4 1.8 1.8 1.6 1.4
1.8 1.7 1.7 0.9 1.6 1.4 1.4 1.0


Kernel 2 results
 ==================================
2.6 2.5 2.7 1.6 2.1 1.7 1.3 1.8
3.0 2.6 2.8 1.9 2.6 2.5 1.9 1.9
2.1 2.2 2.4 1.8 2.0 2.1 1.3 1.5
2.3 1.9 2.2 1.4 2.1 1.9 1.2 1.5
2.5 2.3 2.8 1.6 1.9 2.2 1.8 1.7
1.3 1.2 1.4 0.9 1.2 0.8 0.5 1.1
2.1 1.8 2.1 1.4 1.8 1.8 1.6 1.4
1.8 1.7 1.7 0.9 1.6 1.4 1.4 1.0


Kernel 3  results
 ==================================
-2.6 -2.5 -2.7 -1.6 -2.1 -1.7 -1.3 -1.8
-3.0 -2.6 -2.8 -1.9 -2.6 -2.5 -1.9 -1.9
-2.1 -2.2 -2.4 -1.8 -2.0 -2.1 -1.3 -1.5
-2.3 -1.9 -2.2 -1.4 -2.1 -1.9 -1.2 -1.5
-2.5 -2.3 -2.8 -1.6 -1.9 -2.2 -1.8 -1.7
-1.3 -1.2 -1.4 -0.9 -1.2 -0.8 -0.5 -1.1
-2.1 -1.8 -2.1 -1.4 -1.8 -1.8 -1.6 -1.4
-1.8 -1.7 -1.7 -0.9 -1.6 -1.4 -1.4 -1.0

 ==================================
Error detected : Results mismatched.

Kernel 3 on GPU 0 results
 ==================================
-2.6 -2.5 -2.7 -1.6 -2.1 -1.7 -1.3 -1.8
-3.0 -2.6 -2.8 -1.9 -2.6 -2.5 -1.9 -1.9
-2.1 -2.2 -2.4 -1.8 -2.0 -2.1 -1.3 -1.5
-2.3 -1.9 -2.2 -1.4 -2.1 -1.9 -1.2 -1.5
-2.5 -2.3 -2.8 -1.6 -1.9 -2.2 -1.8 -1.7
-1.3 -1.2 -1.4 -0.9 -1.2 -0.8 -0.5 -1.1
-2.1 -1.8 -2.1 -1.4 -1.8 -1.8 -1.6 -1.4
-1.8 -1.7 -1.7 -0.9 -1.6 -1.4 -1.4 -1.0

 ==================================
Error Cause: Software Error: Please reconfigure SM for your CUDA kernel!
```

Fig. 7. A Screenshot of the Result in the Case of Partial Failure (Software Error).



```
-----------Matrix Multiplication using multiple CUDA kernels--

Kernel 1 results
 ===================================
1.9 0.8 1.1 1.6 2.2 2.0 1.3 1.4
2.5 1.2 1.4 2.1 2.3 2.9 1.6 2.1
1.7 1.1 1.7 2.0 2.9 2.0 1.5 1.6
2.3 1.1 1.7 2.2 2.8 2.8 1.9 1.8
1.5 1.0 1.2 1.6 2.2 1.6 1.0 1.4
1.0 0.7 1.0 1.1 1.7 1.3 0.9 0.9
1.8 0.8 1.3 1.7 2.1 2.2 1.6 1.4
1.6 0.9 1.2 1.8 2.0 1.6 1.2 1.6


Kernel 2 results
 ===================================
1.9 0.8 1.1 1.6 2.2 2.0 1.3 1.4
2.5 1.2 1.4 2.1 2.3 2.9 1.6 2.1
1.7 1.1 1.7 2.0 2.9 2.0 1.5 1.6
2.3 1.1 1.7 2.2 2.8 2.8 1.9 1.8
1.5 1.0 1.2 1.6 2.2 1.6 1.0 1.4
1.0 0.7 1.0 1.1 1.7 1.3 0.9 0.9
1.8 0.8 1.3 1.7 2.1 2.2 1.6 1.4
1.6 0.9 1.2 1.8 2.0 1.6 1.2 1.6


Kernel 3  results
 ===================================
-1.9 -0.8 -1.1 -1.6 -2.2 -2.0 -1.3 -1.4
-2.5 -1.2 -1.4 -2.1 -2.3 -2.9 -1.6 -2.1
-1.7 -1.1 -1.7 -2.0 -2.9 -2.0 -1.5 -1.6
-2.3 -1.1 -1.7 -2.2 -2.8 -2.8 -1.9 -1.8
-1.5 -1.0 -1.2 -1.6 -2.2 -1.6 -1.0 -1.4
-1.0 -0.7 -1.0 -1.1 -1.7 -1.3 -0.9 -0.9
-1.8 -0.8 -1.3 -1.7 -2.1 -2.2 -1.6 -1.4
-1.6 -0.9 -1.2 -1.8 -2.0 -1.6 -1.2 -1.6

 ===================================
Error detected : Results mismatched.

Kernel 3 on GPU 0 results
 ===================================
1.9 0.8 1.1 1.6 2.2 2.0 1.3 1.4
2.5 1.2 1.4 2.1 2.3 2.9 1.6 2.1
1.7 1.1 1.7 2.0 2.9 2.0 1.5 1.6
2.3 1.1 1.7 2.2 2.8 2.8 1.9 1.8
1.5 1.0 1.2 1.6 2.2 1.6 1.0 1.4
1.0 0.7 1.0 1.1 1.7 1.3 0.9 0.9
1.8 0.8 1.3 1.7 2.1 2.2 1.6 1.4
1.6 0.9 1.2 1.8 2.0 1.6 1.2 1.6

 ===================================
Error Cause : Hardware Error : Device configuration issue
Please make sure that GPU device is configured correctly!
```

Fig. 8. A Screenshot of the Result in the Case of Partial Failure (Hardware Error).

E. Discussion

This section aims to give insights into our proposed method and to compare it with other approaches in terms of detecting faults and ability to distinguish them. As explained in the previous section, our proposed method was designed as a two-level technique, in which the first level based on design diversity that applies N-Version Programming technique. Whereas the second level is designed to distinguish the type of error that is detected.

In [30, 50], NVP is used for detecting hardware and software faults, however, they do not address concurrent applications. It is noteworthy to mention that in [30] the system detects transient faults either software or hardware and permanent hardware faults. More specifically, it can detect errors that cause one of the components become disabled or cause the generation of incorrect results.

The most relevant tool to our work is mSWAT [45]. It applies the two-level approach for detecting permanent hardware and software errors, in a similar manner to our work. However, they use TMR approach in the first level, whereas we use NVP. In the second level, mSWAT stores traces of execution for each core, then checks if there is divergence in the execution of one of the cores then it will be considered as a faulty core. In addition, they conduct further analysis to identify the faulty micro-architectural component for repair. In our work, we only report back that there is a permanent hardware error or a software error.

mSWAT also addresses transient errors at the beginning by re-executing the process on all cores in a similar manner of rollback/replay. If the error is not repeated, then it is considered a transient software bug. In our work, we have not included the detection of transient errors.

We detect the existence of errors in the first phase by identifying that one of the software versions are producing in correct results or experiencing application hang. In mSWAT, they have addressed four kinds of software anomalies, including hangs, fatal traps, panic, etc.

It is also worth mentioning that using NVP in our work has the difficulty of designing and implementing three versions of the software, however, it needs no tracing of execution and it only re-executes the problematic version once unlike mSWAT. We also do not need to store the re-execution and do comparisons for divergence checking, we only compare the results in case the application do not hang.

mSWAT, addresses more error types and faulty micro-architectural components identification. However, in our work we investigated the possibility of benefiting from NVP that up-to-our knowledge has not been previously investigated for HPC applications.

VII. Conclusion

Faults are becoming more frequent in large supercomputers, and their impact is higher in the case of long-duration applications. This research seeks to address resilience challenges by presenting an innovative method to detect software and hardware errors that can be become a concern for the performance of scientific applications running on these future systems.

We have investigated an approach to detect and classify faults for CUDA applications using multiple GPUs. Our approach benefits from NVP for detecting errors then carrying another analysis by running the problematic software version on a different GPU to classify the type of error. Our proposed approach is flexible in the sense that it can be applied to different applications not just matrix multiplication. Experimental results indicate the capability of the proposed method to detect errors and classify whether they are permanent hardware errors or software errors. Hence, assisting in improving reliability. We plan to integrate this detection algorithm in a more comprehensive framework that includes error recovery and sophisticated fault injection techniques and test our approach on other types of applications to collect further measurements of the coverage and overhead of our approach.

Acknowledgment

We would like to express our gratitude and appreciation to Prof. Fathy Eassa the head of our research group in the department of Computer Science at KAU. He always provides us with promising research directions and continuous guidance.

REFERENCES

[1] Ashby, Steve, et al. "The opportunities and challenges of exascale computing–Summary report of the advanced scientific computing advisory committee (ASCAC) subcommittee." US Department of Energy Office of Science ,2010.

[2] Van De Vanter, Michael L., D. E. Post, and Mary E. Zosel. "HPC needs a tool strategy." Proceedings of the second international workshop on Software engineering for high performance computing system applications. ACM, 2005.

[3] G.Rinku, et al. "Introspective fault tolerance for exascale systems." US Department of Energy Advanced Scientific Computing Research, OS and Runtime Technical Council Workshop. 2012 .

[4] Constantinescu, Cristian. "Trends and challenges in VLSI circuit reliability." IEEE micro 4 (2003): 14-19.

[5] Gizopoulos, Dimitris, et al. "Architectures for online error detection and recovery in multicore processors." Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011. IEEE, 2011.

[6] Tselonis, Sotiris, Vasilis Dimitsas, and Dimitris Gizopoulos. "The functional and performance tolerance of gpus to permanent faults in registers." On-Line Testing Symposium (IOLTS), 2013 IEEE 19th International. IEEE, 2013.

[7] Wunderlich, Hans-Joachim, Claus Braun, and Sebastian Halder. "Efficacy and efficiency of algorithm-based fault-tolerance on GPUs." On-Line Testing Symposium (IOLTS), 2013 IEEE 19th International. IEEE, 2013.

[8] Guan, Qiang, et al. "Empirical studies of the soft error susceptibility ofsorting algorithms to statistical fault injection." Proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale. ACM, 2015.

[9] Schroeder, Bianca, and Garth Gibson. "A large-scale study of failures in high-performance computing systems." IEEE Transactions on Dependable and Secure Computing 7.4 (2010): 337-350.

[10] Di Carlo, Stefano, et al. "Fault mitigation strategies for CUDA GPUs." Test Conference (ITC), 2013 IEEE International. IEEE, 2013.

[11] Dimitrov, Martin, Mike Mantor, and Huiyang Zhou. "Understanding software approaches for GPGPU reliability." Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units. ACM, 2009.

[12] X. Xu, et. Al. HiAL-Ckpt: A hierarchical application-level checkpointing for cpu-gpu hybrid systems, 2010

[13] K.-H Huang and Abraham, "Algorithm-based fault tolerance for matrix operations, 1984

[14] C. Ding. Et. Al. "Matrix multiplication on GPUs with on-line fault tolerance" 2011

[15] Lyons, Robert E., and Wouter Vanderkulk. "The use of triple-modular redundancy to improve computer reliability." IBM Journal of Research and Development 6.2 (1962): 200-209.

[16] Bartlett, Wendy, and Lisa Spainhower. "Commercial fault tolerance: A tale of two systems." IEEE Transactions on Dependable Secure Computing, 1(1):87–96, 2004.

[17] Pullum, Laura L. Software fault tolerance techniques and implementation. Artech House, 2001.

[18] Ziade, Haissam, Rafic A. Ayoubi, and Raoul Velazco. "A survey on fault injection techniques." Int. Arab J. Inf. Technol. 1.2 (2004): 171-186.

[19] Chen, L. (1978). V-version programming: A fault-tolerance approach to reliability of software operation. FTCS-8, 1978, 6.

[20] B. Randell, "System Structure for Software Fault Tolerance," IEEE Trans, on Software Engineering, Vol. 1, No. 2, June 1975, pp.220-232

[21] Laprie, J. C., Arlat, J., Beounes, C., & Kanoun, K."Definition and analysis of hardware-and software-fault-tolerant architectures". *Computer*, 1990, 23.7: 39-51.

[22] Sabena, Davide, et al. "On the evaluation of soft-errors detection techniques for GPGPUs." Design and Test Symposium (IDT), 2013 8th International. IEEE, 2013.

[23] Sheaffer, Jeremy W., David P. Luebke, and Kevin Skadron. "A hardware redundancy and recovery mechanism for reliable scientific computation on graphics processors." Graphics Hardware. Vol. 2007.

[24] K.S. Yim, C. Pham, M. Saleheen, Z. Kalbarczyk, R. Iyer "Hauberk: Lightweight silent data corruption error detector for gpgpu " ,Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium, IPDPS '11, IEEE Computer Society, Washington, DC, USA (2011), pp. 287-300.

[25] Lei Zhang, Yinhe Han, Qiang Xu, and Xiaowei Li. Defect tolerance in homogeneous manycore processors using core-level redundancy with unified topology. In DATE '08: Proceedings of the conference on Design, automation and test in Europe, pages 891–896. ACM, 2008.

[26] M.-L.Li, et al., "Understanding the Propagation of Hard Errors to Software and Implications for Resilient System Design", ASPLOS 2008.

[27] J.-C. Laprie, Dependability: Basic Concepts and Terminology, 1992.

[28] Pradhan, Dhiraj K. *Fault-tolerant computer system design*. Vol. 132. Englewood Cliffs: Prentice-Hall, 1996.

[29] Dubrova, Elena. *Fault-tolerant design*. New York: Springer, 2013.

[30] DUGAN, J. Bechta; LYU, Michael R. System reliability analysis of an N-version programming application. *IEEE Transactions on Reliability*, 1994, 43.4: 513-519.

[31] B.Schroeder and Garth A Gibson. Understanding failures in petascale computers. Journal of Physics: Conference Series, 78, 2007

[32] Navarro, Cristobal A., Nancy Hitschfeld-Kahler, and Luis Mateu. "A survey on parallel computing and its applications in data-parallel problems using GPU architectures." *Communications in Computational Physics* 15.02 (2014): 285-329.

[33] D. B. Kirk and W. H. Wen-Mei. Programming massively parallel processors: a hands-on approach, 3rd edition. Morgan Kaufmann, 2016.

[34] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym. NVIDIA Tesla: A unified graphics and computing architecture. IEEE micro, 28(2), 2008.

[35] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, S. Buijssen, M. Grajewski, S. Tureka, Exploring weak scalability for FEM calculations on a GPU-enhanced cluster, Parallel Comput. 33 (Nov. 2007) 685–699.

[36] NVIDIA, "NVIDIA kepler K20 GPU datasheet," 2012.

[37] J. M. Yusof et al, "Exploring weak scalability for FEM calculations on a GPU-Enhanced cluster", 33.685–699. Nov, 2007.

[38] Yang, Chao-Tung, Chih-Lin Huang, and Cheng-Fang Lin. "Hybrid CUDA, OpenMP, and MPI parallel programming on multicore GPU clusters." Computer Physics Communications 182.1 (2011): 266-269.

[39] C.T. Yang, C.L. Huang and C.F. Lin, "Hybrid CUDA, OpenMP, and MPI parallel programming on multicore GPU". Computer Physics Communications. Pp. 266-269. 2011.

[40] M.-C. Hsueh, T. Tsai, and R. Iyer, "Fault injection techniques and tools," Computer, vol. 30, no. 4, pp. 75–82, 1997.

[41] Song, Ningfang, et al. "Fault injection methodology and tools." Electronics and Optoelectronics (ICEOE), 2011 International Conference on. Vol. 1. IEEE, 2011.

[42] Netti, Alessio, et al. "FINJ: A Fault Injection Tool for HPC Systems." arXiv preprint arXiv:1807.10056 (2018).

[43] Fiala, David, et al. "Detection and correction of silent data corruption for large-scale high-performance computing." Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society Press, 2012.

[44] Pellegrini, Andrea, et al. "CrashTest: A fast high-fidelity FPGA-based resiliency analysis framework." Computer Design, 2008. ICCD 2008. IEEE International Conference on. IEEE, 2008.

[45] Hari, Siva Kumar Sastry, et al. "mSWAT: low-cost hardware fault detection and diagnosis for multicore systems." Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on. IEEE, 2009.

[46] Li, Man-Lap, et al. "Trace-based microarchitecture-level diagnosis of permanent hardware faults." Dependable Systems and Networks With FTCS and DCC, 2008. DSN 2008. IEEE International Conference on. IEEE, 2008.

[47] Meixner, Albert, Michael E. Bauer, and Daniel Sorin. "Argus: Low-cost, comprehensive error detection in simple cores." Microarchitecture,

2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on. IEEE, 2007.

[48] K.S. Yim, C. Pham, M. Saleheen, Z. Kalbarczyk, R. Iyer Hauberk: Lightweight silent data corruption error detector for gpgpu Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium, IPDPS ’11, IEEE Computer Society, Washington, DC, USA (2011), pp. 287-300.

[49] M. Hiller, A. Jhumka, and N. Suri, “Propane: an environment for examining the propagation of errors in software,” ACM SIGSOFT Software Engineering Notes, vol. 27, no. 4, pp. 81‑85, 2002.

[50] FUHRMAN, Christopher P.; CHUTANI, Sailesh; NUSSBAUMER, Henri J. Hardware/software fault tolerance with multiple task modular redundancy. In: *Computers and Communications, 1995. Proceedings., IEEE Symposium on*. IEEE, 1995. p. 171-177.

Sensual Semantic Analysis for Effective Query Expansion

Muhammad Ahsan Raza[1], M. Rahmah[2], A. Noraziah[3]
Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Kuantan, Malaysia

Mahmood Ashraf[4]
Department of Information Technology
Bahauddin Zakariya University
Multan, Pakistan

***Abstract*—The information has evolved rapidly over the World Wide Web in the past few years. To satisfy information needs, users mostly submit a query via traditional search engines, which retrieve results on the basis of keyword matching principle. However, a keyword-based search cannot recognize the meanings of keywords and the semantic relationship among the terms in the user's query; thus, this technique cannot retrieve satisfactory results. The expansion of an initial query with relevant meaningful terms can solve this issue and enhance information retrieval. Generally, query expansion methods consider concepts that are semantically related to query terms within the ontology as candidates in expanding the initial query. An analysis of the correct sense of query terms, rather than only considering semantic relations, is necessary to overcome language ambiguity problems. In this work, we proposed a query expansion framework on the basis of query sense analysis and semantics mining using computer science domain ontology, followed by working prototype of the system. The experts analyzed the results of system prototype over test dataset and Web data, and found a remarkable improvement in the overall search performance. Furthermore, the proposed framework demonstrated better mean average precision and recall values than the baseline method.**

Keywords—Semantic computing; information retrieval; computational intelligence; ontology; term sense disambiguation

I. INTRODUCTION

At present, the volume of information over the World Wide Web (WWW) has been increasing continuously. Current search engines share this diverse information pool of the WWW and retrieve results by using simple keyword-based matching. These search engines cannot recognize the semantic relevance between search text and student query, thus receiving increased results that are irrelevant to computer science. In this situation, designing a system that interprets user search requirements correctly, rather than providing results by merely performing keyword-based matching, is challenging.

Query expansion is a technique that can be used for effective information searches to satisfy users' requirements. The query expansion process involves augmenting the initial user query with additional terms that are related to user requirements. Currently, among several query expansion techniques studied in [1], ontology browsing is considered a prominent query expansion technique. Ontology provides semantics to plain text [2]; thus, finding additional query-related concepts by exploring the semantic relations is useful in exploring semantic relations.

Focusing on computer science discipline, where data are unstructured and dispersed over WWW, this research work proposes an alternate sense-based semantic query expansion framework to overcome the word mismatch problem of keyword-based searches. Given a user query, the approach initially captures the set of senses for query terms. Then, the relevant concepts from ontology are extracted on the basis of term-sense data. Finally, the extracted concepts are used to expand the initial query for obtaining user-centric results.

Our approach extends the model presented in [3] via disambiguation of query term sense and semantic similarity strategy for selecting and ranking expanded terms.

The remainder of this paper is organized as follows. Existing techniques for query expansion based on ontology are reviewed in Section 2. Section 3 outlines the major steps of our approach alongwith the ontologies used in the query expansion method. The query expansion framework is discussed, and the functionality of each component of the framework is clarified in Section 4. Section 5 details the experiment results and analysis of this work. Section 6 presents the conclusion and highlights the future work.

II. RELATED WORK

Existing keyword-based search techniques have been used for retrieving information from large unstructured data on the WWW [4]. Such techniques retrieve results on the basis of matching the keywords from the user query. However, keyword-based techniques lack semantic orientation and cannot capture the user information requirements.

Query expansion is used to improve the performance of information retrieval system and retrieve results that are user-relevant. Ontology is useful in query expansion because it provides a means for discovering unstated concepts that can be used to expand the initial user query. Early works have explored the use of ontology in query expansion techniques that have been extended eventually in different ways, such as domain-specific ontology [5-7], general ontology [8-9] and linguistic expansion [10-11].

Bhogal, Macfarlane, & Smith in [12] have reviewed the role of ontology in discovering the terms for expanding seed query, whereas [13] provided a comprehensive overview of recent query expansion techniques for supporting an effective information retrieval. Authors in [14] contended that general and similar concepts related to the original query can be identified using thematic relations of ontology. The researchers used semantic relations and qualifiers (i.e., specified in seed query) to filter possible features for reformulation of a new expanded query. The work focused on geographical test data and queries and showed improvements in the accuracy of results. In [15], authors leveraged ontology to obtain the rarely occurred opinions about a product. The authors are of the view that such opinion targets have high chance of relatedness with frequently occurred targets. The proposed hybrid architecture showed improved results over existing techniques using semantic data.

Regarding semantic expansion, Gan & Hong [23] explored Wikipedia knowledgebase and three corpuses (CACM, ADI and CISI) to extract the terms relationships. They constructed a Markov network to select the relevant candidates for query expansion. Their experimental results showed that the proposed method outperforms the baseline model. Another application of query expansion includes word sense disambiguation, in which linguistic knowledge is exploited to select the correct word sense. For example, authors in [16] tested the use of WordNet (a famous thesaurus for providing word senses, e.g., set of synonyms) in query expansion. Their method achieved a 57% disambiguation rate using the standard expansion procedure.

However, our approach differs from previous works in three aspects: first, we exploit linguistic knowledge to disambiguate the term senses accurately to support vocabulary mismatch issues. Second, we generate computer science-relevant concepts from the ontology. Finally, the extracted concepts are evaluated and selected using a graph-based similarity method to formulate a precise expanded query that reflects the users' information requirements.

III. Overview of Approach

Our approach offers two expansion modes, namely, term sense disambiguation and semantic expansion. The former mode aims to solve the vocabulary mismatch problem, thus facilitating users to write textual queries in their own vocabulary. The latter expansion mode relies on ontology in selecting the concepts and relations that are relevant to user query. The rationale behind using two stages of expansion is that, the query itself makes the machine understand the user's demands.

The major steps in our approach are presented as follows:

Step1: Submit the initial query to the system as a set of terms Q.

Step2: Convert the initial query into a standardize query $Q' = \{q_1, q_2, q_3, . , q_K\}$ by removing the noise and stop-words, where $|Q'| = K$. For instance, query 'Algorithm for searching' can be represented as {{algorithm}, {searching}}.

Step3: Search thesaurus to extract set of senses $Sq'_i = \{s_1, s_2, s_3, \ldots, s_N\}$ for each term $q'_i \in Q'$, where $|Sq'_i| = N$ and $1 \geq i \leq$. K. Compute semantic similarity score for each sense in Sq'_i against q'_i and arrange senses in descending order of obtained scores. Moreover, include q'_i into corresponding Sq'_i as follows:

$$Sq'i = \{ \{q'i\} \cup \{s1, s2, s3, \ldots, sN\} \} \quad (1)$$

Step4: For each element of Sq'_i, browse ontology to find number of relevant concepts and add them in set $C_i = \{c_1, c_2, c_3, \ldots, c_M\}$, where $|C_i| = M$. Thus, we obtain vector $\overrightarrow{C_i}$, against each q'_i (see Fig. 1).

Step5: Calculate the semantic similarity score of each element of $\overrightarrow{C_i}$ against corresponding q'_i. Fig. 1 illustrates that each concept is assigned a similarity score.

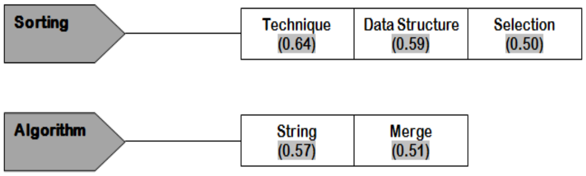


Fig. 1. Concept Vectors for Each Query Keyword.

Step6: Expand the user query Q′ with concepts that achieves high similarity score.

Step7: Submit the expanded query to the information retrieval system for results.

In the present work, we use WordNet thesaurus and computer science domain ontology to reformulate the initial query, in an attempt to understand user requirements in semantic manner.

A. WordNet Lexicon

Many researchers have focused on using the WordNet lexicon for query expansion work. The lexicon represents precise word relationships that are further categorized into 26 types, such as hyponym and synonym. Miller first introduced the WordNet lexicon in 1995 [17], while the latest available version is 3.1.

We use WordNet 3.1 and only focus on the synonymy relationships (namely, synsets) of the lexicon. These synsets provide a means for obtaining term senses to disambiguate user query.

B. Computer Science Domain Ontology

The use of computer technologies in our lives has caused the development of computer science as a distinct discipline. Computer science is an appealing discipline given the implementations that concern every aspect of life. Furthermore, this discipline has various sub-fields, such as database systems (the study of fundamental properties of relations and query processing) and programming languages (the study of approaches to describe problem-solving computations).

Ontology can be used in organizing the data in computer science discipline, thus enabling to browse relations among concepts semantically. We select the computer science domain ontology [18] developed at the University of Athens by Michael Sioutis and encoded in the web ontology language

(OWL). This ontology formally describes all branches of computer science (e.g., algorithms, artificial intelligence, programming languages, and data structures.) using relationships among these branches, such as hasPart and isPartOf.

Figure 2 depicts the portion of the graphical hierarchy of domain ontology. This portion indicates the concepts and their relations. For example, the programming methodology and languages concept is related to the computer science concept via isPartOf relationship.

We use computer science ontology to extract concepts that are semantically related to user search query. The search results in using such concepts when added to original user query are better than the original query.

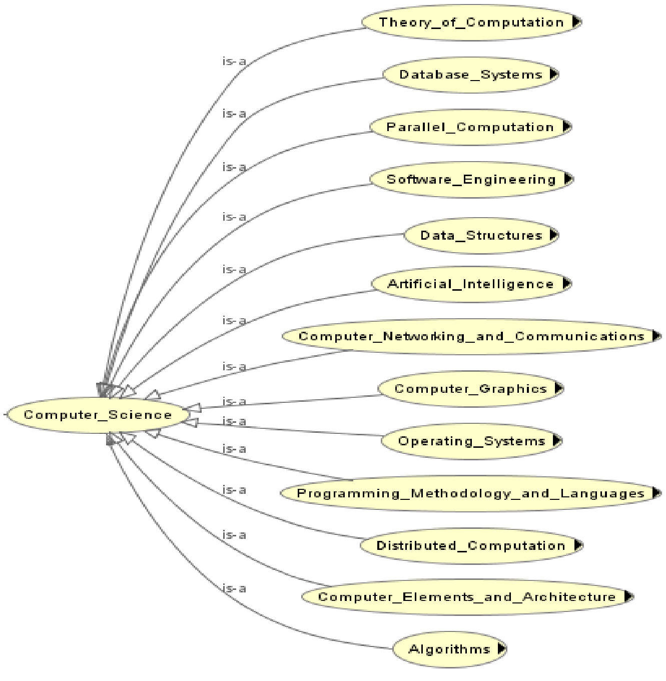


Fig. 2. OWLViz View of Computer Science Ontology.

IV. Query Expansion Framework

Figure 3 demonstrates an overview of the proposed framework for query expansion based on ontology. The framework is based on five main units, namely, user interface, query refinement, sensual semantic expansion, similarity inference, and query constructor components. These components are described in the following subsections, starting from the initial user query to generating final results. Our approach is based on expanding the initial query that focuses on sense disambiguation, computer science domain semantics, and use of semantic similarity method to filter the set of candidate expansion terms.

A. User Interfaces

The user poses initial search query Q and views the results returned by the information retrieval system via a user-view interface.

B. Query Refinement

Query refinement module adds several basic structures to the unstructured initial student query. The two basic operations of this component are tokenization and stemming. At the end of these operations, a pool of keywords (called standard query Q′) from the initial student query has become available.

1) *Tokenization*: Tokenization splits the query into words called tokens on the basis of a space character. Thus, we obtain two types of tokens, namely, word and space tokens. Furthermore, stop-words (e.g., the, a, and an) are removed from word tokens to obtain the query keywords.

2) *Stemming*: Stemming helps in identifying basic forms of query keywords by removing the affixes from each term. We use Porter stammer [19] to stem.

For example, for a given query Q = ‘an algorithm for sorting’, the query keywords after tokenization are {sorting, algorithm}, and the stammer provides us with the standard form of query Q′ as {sort, algorithm}.

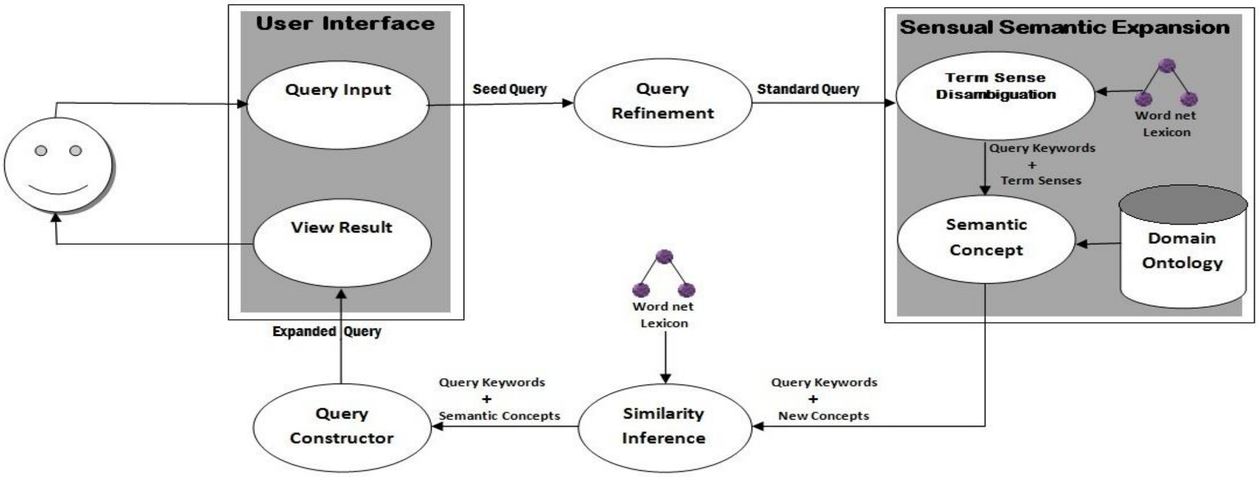


Fig. 3. Ontology based Query Expansion Model.

C. Sensual Semantic Expansion

Sensual semantic expansion (SEE) component supports the sense disambiguation of query Q′ and obtains a set of expansion terms (semantic concepts). The basic function is to match query term senses against computer science ontology concepts to find unstated concepts that imitate the user's interest. The SSE module processing is conducted via two phases.

1) Term sense detection: In this phase, multiple senses for each keyword of query Q′ are extracted via synsets of WordNet lexicon. These senses provide a means for referring to the same keyword in multiple ways. Moreover, this information helps in avoiding the retrieval of irrelevant candidate concepts for expansion.

2) Semantic concept identification: Given the query senses, this phase uses a knowledgebase (which in our case is computer science domain ontology) to discover the semantic concepts. Each sense of query keyword is matched with ontology concepts. If a match is found, then classes (concepts) related to matched ontology concept are extracted via hyponym and hypernym relationships.

Note that if match is found, then the SSE module omits ontology search for the remaining senses of a keyword; otherwise, it rejects a keyword. Algorithm1 describes the SSE sub-tasks. For each keyword of Q′, the algorithm obtains the sense vectors, and sense data are used to extract concepts that are related to initial query Q.

Algorithm 1 Algorithm for sensual semantic expansion

```
Input  :  Q′ , The set of query keywords
          DO, domain ontology
Output :  SC , Set of semantic concepts

FOR each keyword(i) in Q′
  // Term sense detection
  IF keyword is found in WordNet
    Compute synset for keyword(i) // set of synonyms
  END
  // Semantic concept identification
  FOR each sense(j) in synset
      IF sense(j) is found in DO
        // Traverse hypernym relationship
        // and hyponym relationship to one level
        Extract ontology concepts
        Add concepts in SC
        BREAK;  // omit search for remaining senses
      END
  END FOR
END FOR
```

D. Similarity Inference

The inclusion of query senses during term sense detection task may return computer science concepts (i.e., semantics) that are loosely related to user requirements. For example, in query Q of our example, the system must provide the concepts that represent different techniques of list sorting. Thus, the identified semantic concepts must be analyzed to check whether these concepts are representative of the original query Q or not.

In this phase, we adopt a semantic similarity measure for the following purposes: (1) to arrange the query senses on the basis of their scores and (2) to select among the set of candidate expansion concepts (with high scores) recognized by the SEE component. We use Zhou similarity measure [20] to evaluate the similarity score of each query Q′ keyword against corresponding senses and expansion concepts. The scores obtained using Equation (2) show the relatedness of concepts with query Q′. Moreover, we set k=0.5 to obtain a hybrid (i.e., path-based and information content-based) similarity value on the basis of the WordNet lexicon.

$$Sim\_score\ (w,c) = 1 - k \times A - (1-k) \times B \tag{2}$$

Where

$$A = \left(\frac{\log(len(w,c) + 1)}{\log(2 \times \max(depth_w\ ,depth_c) - 1)} \right)$$

And

$$B = \left(\frac{IC(w) + IC\ (c) - 2 \times IC\big(lso\ (w,c)\big)}{2} \right)$$

E. Query Constructor

The query constructor component formulates the expanded query. It receives the list of high-similarity-scored concepts and combines them with initial query Q to create an expanded query.

The query after the expansion is then automatically posted to the information retrieval system, which retrieves results for the user.

V. Experimental Results and Analysis

The above mentioned approach (called SSE) is implemented with a prototype of a system that uses tools, such as NetBeans, Jena API, and ARQ engine. We aim to retrieve the most relevant information for the users without requiring to navigate through irrelevant results. To obtain the search results, we use Atire search engine [21] as basic retrieval model. We evaluate our approach using Communications of the ACM (CACM) test collection, which consists of documents from the domain of computer science [22]. In addition, we have extracted 50 queries from the CACM topics and have selected the top 20 expansion terms in our experiments.

A. Evaluation Metrics

To measure the effectiveness of retrieval, we use two metrics, namely, mean average precision (MAP) and recall. The MAP indicates the accuracy of retrieved results, whereas recall denotes the completeness of results. We define these measures as follows.

$$MAP = \tfrac{1}{Q}\textstyle\sum_{i=1}^{Q}\tfrac{1}{R_i}\sum_{j}^{R_j} P(D_j) \tag{3}$$

$$Recall = \tfrac{1}{Q}\textstyle\sum_{i=1}^{Q}\tfrac{r_i}{R_i} \tag{4}$$

Where Q is the number of queries, R_i represents total number of relevant documents for *i-th* query, P refers to precision, D_j is the *j-th* document from top retrieved set of documents and r_i reflects the number of relevant documents retrieved for *i-th* query.

B. Results Analysis

In the experiment, 80 computer science students (10 undergraduate, 30 graduate, and 40 postgraduate) were divided into 4 groups randomly. All of these groups were trained to retrieve results via the Atire search engine for queries that were expanded using our approach (SSE) and for the unexpanded queries (called baseline). In addition, the groups were required to record the score of relevant results (i.e., documents relevant to a given query) achieved using the SSE and baseline method. For the sensitivity analysis, the groups were requested to measure the relevancy score separately for the top 50 and top 100 retrieved results of each submitted query.

For the performance analysis, three measurement variants, namely, MAP@50, MAP@100 and Recall@100 were calculated. The difference among these measures is based on the analysis of the top N retrieved documents (where N can be 50 or 100). The results for baseline and SSE model in terms of MAP and Recall are summarized in Table 1.

TABLE I. COMPARISON IN TERMS OF MAP@50, MAP@100 AND RECALL@100 MEASURES VIA ATIRE SEARCH ENGINE

Evaluation Measures	**Baseline**	**SSE Approach**
MAP@50	0.15	0.25
MAP@100	0.11	0.23
Recall@100	0.48	0.74

The SSE approach achieved a considerable improvement over the baseline method in the MAP and recall values. When top 50 documents are retrieved, the SSE approach shows improvement about +10% in MAP as those of baseline method. The results trend is found very similar for top 100 retrieved documents. Moreover, we realize that SSE method can improve the recall measure about +26%, when 1000 documents are retrieved. This observation further confirms the effectiveness of proposed SSE system.

The SSE approach is better than the baseline method given the following reasons: (1) SSE leverages user query senses at the initial stage. Therefore, the new approach helps in identifying the correct sense for the original query terms. (2)The SSE method avoids including unnecessary expansion terms by considering the computer science domain ontology and semantic similarity method.

The efficiency of the SSE procedure was further evaluated using Google, which is a search engine that is widely used by users. Table 2 reports the results measured with MAP and Recall for both Atire based and Google based SSE. An interesting observation is that the performance improvement in the Google-based SSE method is similar to the Atire-based SSE method for top 50 (MAP@50) and top 100 (MAP@100) retrieved documents. Therefore, the SSE approach can stably improve the retrieval accuracy for the Web-based search. By contrast, the Recall@100 result for the Google-based SSE is less substantial than the method implemented in the Atire toolkit but still much better than the baseline method (i.e., +15%).

TABLE II. Comparison in Terms of Map@50, Map@100 and Recall@100 Measures Via Atire and Google Search ENGINES

Evaluation Measures	**Atire Based SSE**	**Google Based SSE**
MAP@50	0.25	0.24
MAP@100	0.23	0.21
Recall@100	0.74	0.63

Finally, the results were plotted in a 2Dchart for the MAP and recall values. Fig. 4 displays that the SSE approach and Google-based SSE search outperform the baseline method. The results trend indicates that SSE method achieved better Recall value when Atire retrieval system is adopted, in contrast to Google-based retrieval. We believe the main reason for this is that the pool of expansion terms is kept small in size (20 terms). In Atire-based SSE, the fewer expansion terms provides an effective guidance in selection of relevant results.

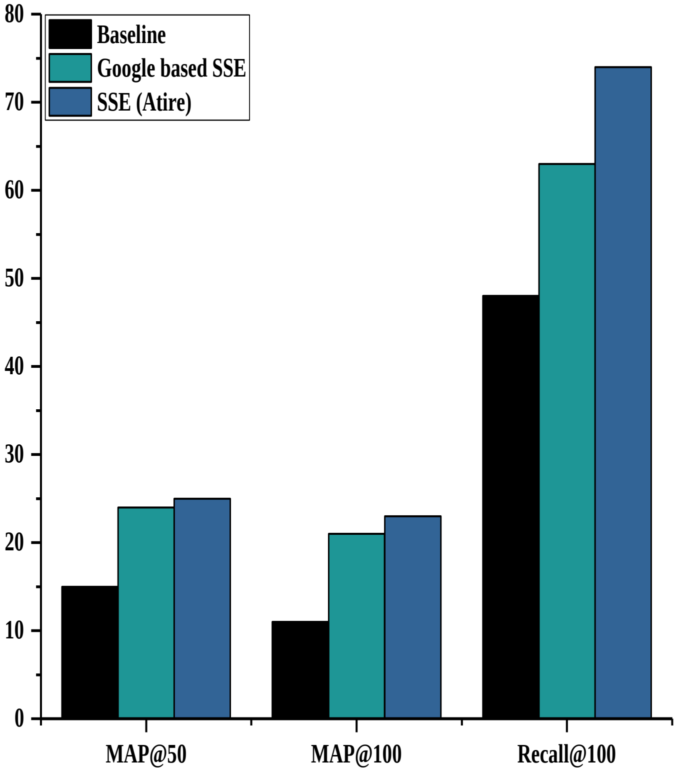


Fig. 4. Column Plots of MAP@50, MAP@100 and Recall@100 Measures.

VI. Conclusion and Future Work

In this work, we have addressed the problem of accurate search by focusing on reformulating the user query to become self-explanatory. We have proposed a sensual semantic framework for query expansion and have used semantic structures i.e., ontologies. In particular, the SSE method helps in extracting the correct sense of a query term from WordNet ontology. The semantic expansion process takes place by browsing computer science ontology for additional terms related to query terms and query senses. The generated expansion terms are then analyzed using similarity inference to select terms closely related to query senses. Experts have evaluated our prototype on the CACM collection and the Atire search engine. Our system has obtained the optimal results for MAP@50, MAP@100, and Recall@100 using test dataset. Moreover, we have tested the capability of SSE system on WWW using Google search engine. The difference between the Atire-based SSE and Google-based SSE methods for MAP@50 and MAP@100 is insignificant. This comparative analysis indicates that our approach is also useful in retrieving precise information from a diverse information pool of WWW. Our model has outperformed the baseline method, thereby indicating that the concept of query sense analysis along with semantic expansion can provide a breakthrough in retrieving relevant information for users.

Our future work includes enhancing our prototype for large standard ontologies (in contrast to domain-specific ontology) and making the prototype available to researchers to test its authenticity and detailed analysis in various domains.

References

[1] Carpineto, C., & Romano, G., “A survey of automatic query expansion in information retrieval,” *ACM Computing Surveys (CSUR), 44*(1), 1, 2012.

[2] Zhu, X.-H., Wu, T.-J., & Chen, H.-C., “An Interoperable Model for the Intelligent Content Object Based on a Knowledge Ontology and the SCORM Specification,” Journal of Educational Computing Research, 56(5), 723-749, 2018. doi: 10.1177/ 0735633117725764

[3] Raza, M.A., et al., “Query expansion using conceptual knowledge in Computer Science,” in 5th International Conference on Software Engineering & Computer Systems (ICSECS17), 2017. Langkawi Islang, Malaysia.

[4] Yonggang Qiu, and Hans-Peter Frei, “Concept based query expansion,” SIGIR conference on Research and development in information retrieval, pp. 160–169, New York, NY, USA, 1993.

[5] Waseem Alromima, Ibrahim F. Moawad , Rania Elgohary and Mostafa Aref , “Ontology-based Query Expansion for Arabic Text Retrieval,” International Journal of Advanced Computer Science and Applications(IJACSA), 7(8), 2016. http://dx.doi.org/10.14569/IJACSA. 2016.070830.

[6] Xinhua, L., Xutang, Z., and zhongkai, L., “A domain ontology-based information retrieval approach for technique preparation,” Physics Procedia. 25:1582-1588, 2012.

[7] Nacim Yanes, Sihem Ben Sassi, and Henda Hajjami Ben Ghezala, “Ontology-based recommender system for COTS components,” Journal of Systems and Software, Volume 132, 2017, Pages 283-297, ISSN 0164-1212.

[8] Adeel Ahmed and Syed Saif ur Rahman, “DBpedia based Ontological Concepts Driven Information Extraction from Unstructured Text,” International Journal of Advanced Computer Science and Applications(ijacsa), 8(9), 2017. http://dx.doi.org/10.14569/IJACSA. 2017.080954.

[9] Yuanfeng He, Yuanxi Li, Jiajia Lei, C.H.C Leung, “A framework of query expansion for image retrieval based on knowledge base and concept similarity,” Neurocomputing, Volume 204, 2016, Pages 26-32, ISSN 0925-2312.

[10] C. H. C Leung, and Alfredo Milani, “Collective evolutionary concept distance based query expansion for effective web document retrieval,” 2017.

[11] Bhawani Selvaretnam, and Mohammed Belkhatir, “A linguistically driven framework for query expansion via grammatical constituent highlighting and role-based concept weighting,” Information Processing & Management, Volume 52, Issue 2, 2016, Pages 174-192, ISSN 0306-4573.

[12] Bhogal, J., A. Macfarlane and P. Smith, “A review of ontology based query expansion,” Information Processing and Management 43(4): 866-886, 2007.

[13] Azad, H.K. and A. Deepak, “Query expansion techniques for information retrieval: a survey,” CoRR, 2017. abs/1708.00247.

[14] Mauro, N., et al., “Concept-aware geographic information retrieval,” in Proceedings of the International Conference on Web Intelligence 2017, ACM: Leipzig, Germany. p. 34-41.

[15] khan, K., Ullah, A., & Baharudin, B. “Pattern and semantic analysis to improve unsupervised techniques for opinion target identification,” Kuwait Journal of Science, 43(1), 129-149, 2016.

[16] Crimp, R. and A. Trotman, “Automatic term reweighting for query expansion,” in Proceedings of the 22nd Australasian Document Computing Symposium 2017, ACM: Brisbane, QLD, Australia. p. 1-4.

[17] G. A. Miller. “WordNet: A lexical database for English,” CACM 38, 11, 39–41, 1995.

[18] Sioutis M. “Computer Science Ontology,” Department of Informatics and Telecommunications, University of Athens, 22 June 2009. http://cgi.di.uoa.gr/~std04153

[19] Porter, M. F., “An algorithm for suffix stripping. In Readings in Information Retrieval,” K. S. Jones and P. Willett Eds., Morgan Kaufmann, 313–316, 1997.

[20] Zhou, Z., Wang, Y. and Gu, J., “New model of semantic similarity measuring in Wordnet,” in Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, 2008, November 17-19, Xiamen, China.

[21] A. Trotman, C. L. A. Clarke, I. Ounis, S. Culpepper, M.-A. Cartright, and S. Geva, “Open Source Information Retrieval,” A Report on the SIGIR 2012Workshop. SIGIR Forum 46, 2, 95–101, 2012.

[22] Salton, G., Fox, E. A., & Wu, H., “Extended Boolean information retrieval,” Commun. ACM, 26(11), 1022-1036, 1983. doi: 10.1145/182.358466

[23] Gan, L., & Hong, H., “Improving Query Expansion for Information Retrieval Using Wikipedia,”. *8*(3), 27-40, 2015. doi: 10.14257/ijdta.2015.8.3.03.

Reading the Moving Text in Animated Text-Based CAPTCHAs

Syed Safdar Ali Shah[1], Riaz Ahmed Shaikh[2], Rafaqat Hussain Arain[3]
Department of Computer Science
Shah Abdul Latif University
Khairpur, Pakistan

***Abstract*—Having based on hard AI problems, CAPTCHA (Completely Automated Public Turing test to tell the Computers and Humans Apart) is a hot research topic in the field of computer vision and artificial intelligence. CAPTCHA is a challenge-response test conducted to single out humans and bots. It is ubiquitously implemented on the web since its introduction. As text-based CAPTCHAs are successfully broken by various researchers therefore several design variants have been proposed and implemented in order to further strengthen it. Animated Text-based CAPTCHAs are one of the design variant of it and are based on the difficulty of reading the moving text. They are based on zero knowledge per frame principle. Although it's still easy for humans to read animated text but it's a challenge for machines. As proposals for animated CAPTCHAs are on the rise so there is a strong need to scrutinize their strength against automated attacks. In this research, such CAPTCHAs are investigated to verify their robustness against automated attacks. The proposed methods proved that these CAPTCHAs are vulnerable and they do not guarantee the robustness against automated attacks. The proposed frame selection, noise removal, segmentation and recognition methods have successfully decoded these CAPTCHAs with an overall precision, segmentation accuracy and recognition rate of up to 53.8%, 92.9% and 93.5% respectively.**

Keywords—Bots; CAPTCHAs; ANNs; animations; image processing; HIPs; machine learning

I. Introduction

The acronym CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart. This term was coined by Ahn et al. in their pivotal publication in 2000 to thwart the web against bots [1]. CAPTCHAs are also known as HIPs (Human Interaction Proofs). According to one report by BBC, 61% of web traffic is generated by bots [2]. Therefore stopping the bots and securing the web traffic is indispensable. Now CAPTCHA is a standard security mechanism on the web to identify the human interaction. It is based on hard AI (Artificial Intelligence) problems. These are the problems which are supposed to be fairly easy for humans but extremely difficult for machines. As the CAPTCHAs are based on hard AI problems, they have emerged as a hot research topic in the fields of computer vision and artificial intelligence. Additionally as 'P' stands for public [3], means the underlying algorithm to create the test should be open to research community as stated by Ahn et al. This statement by the pioneers of the field motivates the researchers to work in this field.

Over the years, many design variants are introduced by CAPTCHA designers. On the other hand, attackers have successfully decoded these CAPTCHAs. As the existing CAPTCHA schemes are broken by the attackers, new design variants are introduced by designers. Therefore it is an ongoing war since the introduction of first CAPTCHA [4]. However in both cases it's a victory. If the CAPTCHA is successfully broken then a hard AI problem is solved which leads to one step forward in machine learning. On the other hand if a CAPTCHA is found resistant against automated attacks then a way to distinguish between human and machines is devised which provides a security mechanism on the web.

A large number of static text-based CAPTCHAs are successfully broken by various researchers [5-8]. Therefore a growing number of animated CAPTCHAs are proposed by many researchers. Animated text-based CAPTCHA are an alternative to static Text-based CAPTCHAs. In this type of CAPTCHA, the user is asked to read the moving text as shown in Figure 1. This type of test presents animated text typically in Graphics Interchange Format (GIF) images. However other formats like flash files and streaming videos can also be used to present animated CAPTCHAs. The information is usually, spread in multiple frames rather than presenting it in a single frame like 2D static CAPTHCAs. It is assumed that adding the time dimension in this type of CAPTCHAs makes it more secure than its counterparts, i.e. 2D static CAPTCHAs. As various attackers have successfully attacked the static text-based CAPTCHAs but very little research is carried out to verify the robustness of animated text-based CAPTCHAs. In this research, the robustness of these CAPTCHAs is verified. There is a dual benefit of this research, it not only helps to identify the design flaws in current animated CAPTCHAs but the proposed algorithms can also be used to read the moving text. By implementing the proposed frame selection, noise removal, segmentation and machine learning methods, targeted CAPTCHAs are successfully broken as discussed in section 3.

Rest of the paper is organized as follows; Section 2 presents literature review of the field. Section 3 presents the proposed frame selection, noise removal, and segmentation and recognition methods. Section 4 presents the results of the proposed methods, finally section 5 presents the conclusion and future work.

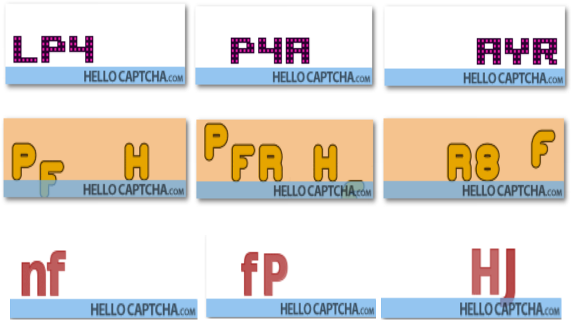


Fig. 1. Animated Text-Based Captchas.

II. Literature Review

Since its introduction, numerous design variants of CAPTCHAs are introduced by researchers. However text-based CAPTCHAs are still most prevalent due to simplicity and ease of use [9]. As text-based CAPTCHAs are successfully attacked by many researchers, therefore they have evolved over the years. Various extraction, segmentation and recognition resistance schemes have been introduced to improve their robustness against automated attacks. In spite of many resistance schemes, various popular CAPTCHAs offered by Yahoo, Google and MSN have been successfully broken [7][10][11]. It has led to the developers to introduce new design alternatives. Animated text-based CAPTCHA is an alternative to static text-based CAPTCHAs. In this CAPTCHA the user is asked to read the text which is spread in various frames.

Cui et al. termed it as Zero Knowledge Per Frame Principle [12]. They have proposed an animated CAPTCHA scheme containing moving letters on a noisy background. Fischer and Herfet proposed an idea of presenting the text on a deforming surface [13]. Chow and Susilo proposed an animated CAPTCHA scheme based on motion parallax [14]. In this scheme, segmentation resistance is used in moving text, humans can still identify the individual characters while they move However it is supposed as a difficult job for machines. Creo Group has introduced a HelloCAPTCHA scheme, which spreads the information in various frames [15]. Naumann et al. introduced a similar CAPTCHA scheme where letters and other sketches move on a noise background, they become observable while moving in different areas of an image [16].

Ince et al. introduced an interesting 3D CAPTCHA scheme where the users are presented with a randomly selected text [17]. The user is asked to type the characters/numbers in color input boxes read from each side of a multicolor 3D cube. Chaudhari et al. presented an idea of a 3D drag n drop CAPTCHA [18]. The user is presented with a randomly generated 3D text. Instead of typing the test, the user is asked to drag n drop the individual letters in boxes. Susilo et al. introduced a CAPTCHA which is built from stereoscopic 3D images [19]. They stated that the distorted and overlapped 3D text in stereoscopic images will increase the complexity against automated attacks. Imsamai and Suphakant proposed a 3D CAPTCHA scheme, where multiple factors like rotation, overlapping, distributed noise etc. were added in alphanumeric characters to improve its robustness against bots [20].

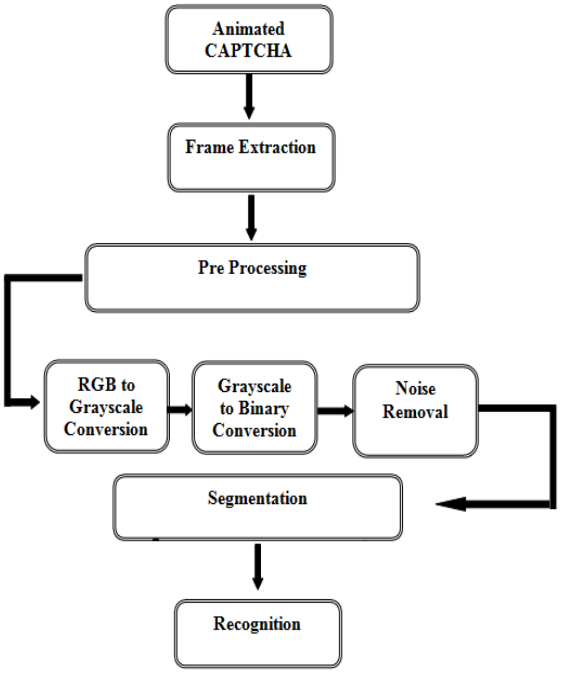


Fig. 2. Sytem Diagram of the Proposed Algortihm.

III. Proposed Method

In this research, HelloCAPTCHA[15] is selected as representative of animated CAPTCHA scheme because it offers a variety of design variants of animated CAPTCHAs. Three different schemes were selected, i.e. flitter, Popup and Smarties CAPTCHAs as shown in Figure 1. All the challenged images are GIF images containing a certain number of frames. These frames are displayed after a fixed duration of time. In this way, animation of characters is achieved. The system diagram of the proposed method is shown in Figure 2.

A. *Frame Selection*

As the animated CAPTCHAs present the challenge in multiple frames instead of presenting them in a single frame like static text based CAPTCHAs. Time dimension is therefore added as an extra layer of security. Therefore the selection of frames containing the challenged images of characters is a research problem. Once the required frames can be selected, the problem can be reduced to static CATPCHA. We have analyzed the attacked CAPTCHAs and serious design weaknesses were found in them. These design weaknesses were exploited to break these CAPTCHAs successfully with high precision as discussed in section 4.

All attacked types of CAPTCHAs are displayed in the size of 180x60 RGB images. Each challenge consists of 6 alphanumeric characters. Flitter CAPTCHAs consists of 173 to 177 frames. Characters appear after specified interval of time at fixed columns in the animation. The major weakness in this design is the fixation of columns and time intervals to display the individual characters. After a certain time period, the individual characters are displayed and we measured this time in flitter images as approx. 55 mil Seconds. Another major weakness is to display a complete character after every

20th frame. These weaknesses are exploited to extract characters from the animated CAPTCHAs. In Popup CAPTCHA scheme, the characters popup in the image at certain columns and then disappear. The animation is achieved by means of moving 80 to 85 frames at regular intervals of time, although the characters popup randomly but stay in the image for a certain period of time in order to catch the human's attention. The individual characters remain appeared in the fixed columns and hence extracted by calculating the amount of time. Smarties CAPTCHAs present each character after a certain interval of time. Every character is displayed at the screen at fixed locations. It remains appeared at that location for a certain period of time and then disappears. Once the 3rd character is appeared then the first is disappeared. The animation is achieved by means of moving 190 to 200 frames in a GIF file. Multiple such CAPTCHAs are analyzed and regular patterns in their appearances were found. These design weaknesses in attacked CAPTCHAs are exploited to successfully decode them. Once the required frames are extracted and labelled then further operations are performed on them, like preprocessing, segmentation and finally the recognition.

B. *Preprocessing*

The obtained frames contain the disconnected characters and the step of segmentation is fairly simple in the selected types of CAPTCHAs. These images are firstly converted to grayscale images as shown in Figure 3.

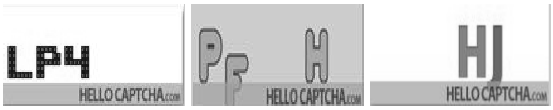


Fig. 3. RGB to Grayscale Conversion.

The obtained Grayscale images are converted to binary images using Equation No.1

$$Y = 0.2989*R + 0.5870*G + 0.1140*B \quad (1)$$

Figure 4 shows the results of grayscale to binary conversion.

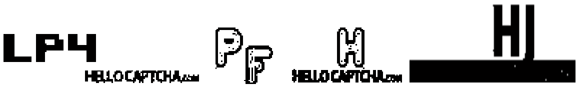


Fig. 4. Grayscale to Binary Conversion.

The obtained binary images contain noise as well footer of the designer's name. This noise can affect the results in the later steps of segmentation and recognition. The footer containing 'HelloCAPTCHA.com' always appears at a fixed location and it is easily removed by calculating its area. Rest of the noise is salt and pepper noise. A threshold value 't' is used to remove all the pixels having values smaller than 't'. Figure 5 shows the results after noise removal.

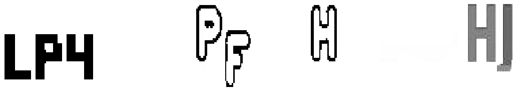


Fig. 5. Results after Noise Removal.

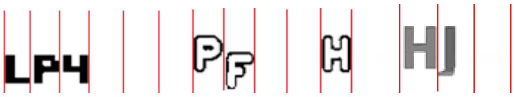


Fig. 6. Character Segments after Segmenation.

C. *Segmentation*

Segmentation aims to separate the individual characters. In the obtained binary images (after noise removal) there is no overlapping of characters. These images therefore can be easily segmented using the condition of blank columns. There are multiple blank columns (columns containing no black pixels) in the binary images and we can obtain the character segments as shown in Figure 6.

The character segments are labeled to store their positions in the string in order to reconstruct the string as output of the process.

D. *Recognition*

In the previous step of segmentation, 200 samples for each type of CAPTCHA were segmented. Therefore Approx. 1200 images of individual characters were obtained. These images of individual characters are used to train an Artificial Neural Network (ANN) using Matlab 9.2. The said ANN is trained using Scaled Conjugate Gradient Algorithm with backpropagation. In order to calculate the performance of the network cross entropy is used for the given targets and outputs. The data is randomly divided into training, validation and testing datasets as 70%, 15% and 15% respectively.

The segmented images of individual characters are normalized. ANN is constructed by calculating the feature vectors of the normalized images by calculating the local and global features of character skeleton [21].

IV. RESULTS AND DISCUSSIONS

Overall success rate of the proposed algorithm depends on segmentation accuracy, recognition success rate of the classifier (ANN) and the number of characters in a challenged image. Therefore Equation 2 is used to calculate the overall precision [22].

$$OP\ (\%) = SSR\%\ (\%) \quad (2)$$

Where OP is the Overall Precision, SSR is the Segmentation Success Rate, SRR is the Success Recognition Rate. SSR depends on the numbers of characters correctly segmented in a dataset of images. For example if a segmentation algorithm can segment 500 characters in 100 images of Flitter CAPTCHAs (100*6, as there are 6 characters in each image) then the segmentation accuracy would be 500/600 = 0.833 or 83.3%. SRR depends on the accuracy of the classifier. Table 1 shows the results of the proposed algorithm.

TABLE I. RESULTS OF THE PROPOSED ALGORITHM

CAPTCHA	SSR (%)	SRR (%)	OP (%)
Flitter	73.3	93.5	48.97
Popup	82.5	89.3	41.83
Smarties	92.9	91.3	53.8

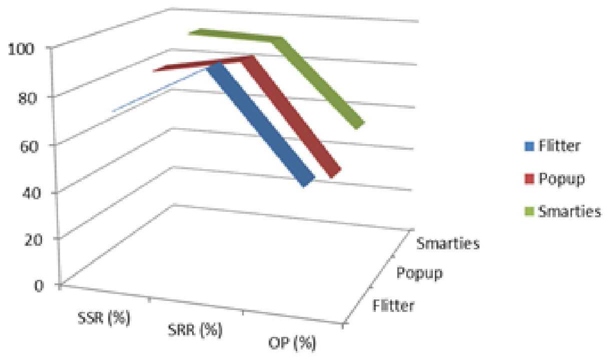


Fig. 7. Segmentation, Recognition and OP Rates.

As mentioned in [23], the CAPTCHA is assumed to be broken if an automated program can decode it with an accuracy of even 1%. Figure 7 displays the results in a graph where it can be observed that our segmentation, recognition and overall precision rates are clearly far beyond this imagination of an ideal CAPTCHA.

V. Conclusion and Future Work

In this paper, three popular animated CAPTCHA schemes offered by HelloCAPTCHA were successfully decoded. An overall precision, segmentation accuracy and recognition rate of up to 53.8%, 92.9% and 93.5% respectively were achieved. Simple but robust image processing techniques were applied to successfully decode them with high accuracy. Serious design weaknesses were exploited in the attacked CAPTCHA schemes. Regular patterns such as fixed columns for certain number of characters and regular time intervals for the appearances of certain characters makes these schemes vulnerable against automated attacks. Furthermore it was validated that the addition of time dimension does not guarantee the robustness against automated scripts.

In future, the robustness of other animated CAPTCHA schemes can be verified, which offer segmentation resistant principles along with addition of time dimensions in their proposed schemes.

References

[1] L. V. Ahn, M. Blum, J. Langford. Telling humans and computers apart automatically. Communications of the ACM, 2004, 47(2): 56-60.

[2] British Broadcasting Corporation. BBC News Technology. http://www.bbc.co.uk/news/technolgy-25346235, Dec.2013.

[3] K. A. Kluever and R. Zanibbi. Balancing usability and security in a video CAPTCHA. SOUPS, ACM international conference proceeding series. ACM. 2009.

[4] R. Hussain, H. Gao, R.A. Shaikh. Segmentation of connected characters in Text-based CAPTCHAs for intelligent character recognition. Multimedia Tools and Applications, 76(24), pp. 25547-25561, 2017.

[5] G. Mori, J. Malik. Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. International conference on Computer Vision and Pattern Recognition Proceedings, Washington, 2003, 134-141.

[6] K. Chellapilla, P. Y. Simard. Using machine learning to break visual human interaction proofs. 17th International Conference on Neural Information Processing Systems Proceedings, Columbia, 2004, 265-272

[7] A. S. El Ahmad, J. Yan, M. Tayara. The robustness of Google CAPTCHA's [EB/OL]. Computing Science Newcastle University, 2011.

[8] O. Starostenko, C. Cruz-Perez, F. Uceda-Ponga, et al. Breaking text-based CAPTCHAs with variable word and character orientation. Pattern Recognition, 2015, 48(4): 1101-1112.

[9] J. Yan, A. S. El Ahmad. Usability of captchas or usability issues in captcha design. 4th symposium on Usable privacy and security Proceedings, New York, 2008, 44–52.

[10] S. Y. Huang, Y. K. Lee, G. Bell, Z. H. Ou. An efficient segmentation algorithm for CAPTCHAs with line cluttering and character warping .Multimedia Tools and Applications, 2010, 48(2): 267-289.

[11] H. Gao, X. Wang, F. Cao, et al. Robustness of text-based completely automated public Turing test to tell computers and humans apart. IET Information Security, 2016, 10(1): 45-52.

[12] J. S. Cui, J. T. Mei, W. Z. Zhang, et al. A CAPTCHA implementation based on moving objects recognition problem. International conference on E-Business and E-Government Proceedings, 2010 Guangzhou, 2010, 1277-1280.

[13] I. Fischer, H. Thorsten. Visual CAPTCHAs for document authentication[C]. IEEE workshop on Multimedia Signal Processing Proceedings, Victoria, 2006, 471-474.

[14] Y. W. Chow, W. Susilo. AniCAP: An animated 3D CAPTCHA scheme based on motion parallax. International Conference on Cryptology and Network Security Proceedings, Sanya, 2011, 255-271.

[15] Crew Group. HelloCAPTCHA vs Spambots. http://www.hellocaptcha.com, 2018, last viewed on 10-Oct-2018.

[16] A. B. Naumann, F. T. Franke, B. Christian. Investigating CAPTCHAs Based on Visual Phenomena. IFIP Conference on Human-Computer Interaction Proceedings, Uppsala, 2009, 745-748.

[17] I. F. Ince, Y. B. Salman, M. E. Yildirim, et al. Execution time prediction for 3d interactive captcha by keystroke level model. Fourth International Conference on Computer Sciences and Convergence Information Technology Proceedings, Seoul, 2009, 1057-1061.

[18] S. K. Chaudhari, A. R. Deshpande, S. B. Bendale, et al. 3D drag-n-drop CAPTCHA enhanced security through CAPTCHA. International Conference & Workshop on Emerging Trends in Technology Proceedings, Mumbai, 2011, 598-601.

[19] W. Susilo, C. Yang-Wai, Z. Hua-Yu. Ste3d-cap: Stereoscopic 3d captcha. International Conference on Cryptology and Network Security, Kuala Lumpur Proceedings, 2010, 221-240.

[20] M. Imsamai, P. Suphakant. 3D CAPTCHA: A next generation of the CAPTCHA. International conference on Information Science and Applications, (ICISA) Proceedings, Ho Chi Minh, 2010, 1-8.

[21] D. D. Gaurav, & R. Ramesh. "A feature extraction technique based on character geometry for character recognition" arXiv preprint arXiv, 2012, 1202.3884.

[22] Starostenko, O., Cruz-Perez, C., Uceda-Ponga, F., & Alarcon-Aquino, V. (2015). "Breaking text-based CAPTCHAs with variable word and character orientation" *Pattern Recognition*, *48*(4): 1101-1112.

[23] Yan, J., & El Ahmad, A. S. (2008). A Low-cost Attack on a Microsoft CAPTCHA. In Proceedings of the 15th ACM conference on Computer and communications security (pp. 543-554) ACM.

Improvement of the Vertical Handover Decision and Quality of Service in Heterogeneous Wireless Networks using Software Defined Network

Fatima Laassiri[1, a], Mohamed Moughit[2, b], Noureddine Idboufker[3, c]
[a, b]IR2M Laboratory, FST, Univ Hassan UH1- Settat, Morocco
[b]EEA&TI Laboratory, FST, Univ Hassan, Mohammedia, Morocco
[b]National Schools of Applied Sciences Khouribga, Univ Hassan 1, UH1- Settat, Morocco
[c]National School of Applied Sciences, Univ Cadi Ayyad Marrakech, Morocco

***Abstract*—The development of wireless networks brings people great convenience. More state-of-the-art communication protocols of wireless networks are getting mature. People attach more importance to the connections between heterogeneous wireless networks as well as the transparency of transmission quality guarantees. Wireless networks are an emerging solution based on users' access to information and services, regardless of their geographic location. The success of these networks in recent years has generated great interest from individuals, businesses and industry. Although there are several access technologies available to the user such as IEEE standards (802.11, and 802.16).SDN is a new network paradigm used to simplify network management, reducing the complexity of network technology. The following work aims to expose a simulation implemented under OMNeT 4.6 ++, to improve the Handover performance between two technologies WiFi and WiMAX. This paper proposes a decision algorithm for a heterogeneous vertical handover between WiFi access points and WiMAX network. The inputs to the algorithm are WiFi RSS, bit rate, jitter, and estimated TCP end-to-end delay.**

Keywords—Heterogeneous network; vertical handover; WiMAX; WiFi; IEEE 802.16; IEEE 802.11; OMNeT4.6

I. Introduction

Recently, the Software-Defined Network (SDN) plays an important role because of its flexibility and ease of transportation. Worldwide Interoperability for Microwave Access (WiMAX)[1] or Global Interoperability for Microwave Access is the promising Fourth Generation (4G)[2] network to meet the needs of customers. It is a telecommunication technology that it provides software-defined data for several distances from a point-to-point links to all cell-type accesses, that it allows the connection between mobile and fixed networks. The coverage area of WiMAX is high compared to other technologies; it offers good support and good stability.

Wireless Fidelity (WiFi) [3] provides broadband connectivity for local area networks, while WiMAX provides broadband coverage in the metropolitan area with guaranteed quality of service. WiFi access points usually offer free access to good time users, while WiMAX overlay networks offer paid access to users. It is therefore necessary to provide WiFi connectivity for such a long time and to allow a roaming mobile device to switch to a superimposed WiMAX network only when the WiFi services are out of range or when its quality of service becomes unacceptable, which means that the WiMAX coverage is supposed to be always available and that the mobile terminal has to switch between WiFi depending on the availability of WiFi access points.

It offers another aspect of how conventional remote systems have been described, due to the effects of innovation on the public and its impact on the environnent. The mechanical points of interest of WiMAX are central points of unprecedented WiMAX radio change, it works in the field of security and quality, and it has an open access base focused on IP access. This innovation can be used for various applications, WiMAX is a scalable remote correspondence system capable of providing high-speed remote access with high data rate of 4G over a long separation in a point-to-multipoint and visible or an unobservable path condition.

This article presents an evaluation and improvement in Handover vertical [4] performance between WiFi and WiMAX via a decision algorithm for the heterogeneous transfer between WiFi access points and a WiMAX networks. The inputs to the algorithm are WiFi Received Signal Strength (RSS)[5], bit rate, jitter, and estimated Transmission Control Protocol (TCP)[6] end-to-end delay.

II. State of the Art

Wireless networks at the forefront of microwave technology and useful in broadband access, as a central innovation for the IEEE 802.16 reference group, are advancing in 4G discovery. With the current presentation of portability management systems in the IEEE 802.16e standard, it is currently competing with current and future ages of remote advances to provide ubiquitous recording arrangements. Nevertheless, the establishment of a decent versatile structure depends to a large extent on the ability to make quick and consistent transfers regardless of the situation of the building being sent. Since the IEEE has characterized the mobile WiMAX MAC layer transfer. (IEEE 802.16e) administration structure and the WiFi MAC layer (IEEE 802.11), the WiMAX Forum and WiMAX Network Workgroup (NWG) [7] are processing the improvement of the upper layers. That is, the path to commercialization of an undeniable structure of versatility, that it is in charge of investigating the difficulties. This focuses on potential research issues related to the transfer into the current and future wireless portability structure. An examination of these issues in the MAC, Network and Cross-Layer situations is presented alongside the exchange of distinctive answers to these difficulties.

Subscriber Station (SS) [8] gives the link to WiMAX. Many websites are used to provide this method, but this is not a complete view of the tools available as certified that they are set in mobile Internet devices and various private labelled tools, laptops.

It assembles Wi-Fi activities because of the same Wireless Network. It also gives network connectivity to businesses and at home without the help of external devices. For this, it uses WiMAX carriers. It is scaled a few kilometers. It varies at the scale of the city. It is given with a subscriber unit. It helps the customer to connect to the Internet and other accesses.

According to research carried out by the "Mojtaba Seyedzadegan"[9] on the overview of wireless networks, its architecture deals with the supply of data. It is entirely based on IEEE 802.16. It supports two things one is the alternative broadband cable and the other is DSL.

On the basis of the existing works of the following researchers: Marceau Coupechoux, Jean-Marc Kifel, Philippe Godlewski that they accumulate the wireless ones in what they follow

- The users arrive at random times.
- The location of the users is taken into account.
- The choice of the network takes place at the arrival of the user.
- The decision is made by a dynamic programming algorithm. Limitations
- Digital resolution is limited by the size of the problem.
- The network cells are concentric.
- The following researchers: Srinivas Shakkotai, Eitan Altman, Anurag Kumare, they present the following states:
- The user population is constant over time.
- Users can connect to multiple WiFi networks simultaneously (Multi-homing).
- The model is macroscopic: The populations follow a dynamic that converges to a state that maximizes the sum of user flows. Limitations
- The cells belong to the same technology (Wi-Fi).
- The algorithm to obtain the dynamics is not provided.

III. Problem and Solutions

Wi-Fi and WiMAX refer to certain types of Wireless Networks Wireless Local Area Network (WLAN)[10] and Wireless Metropolitan Area Network (WMAN)[10], using 802.11 [11] and 802.16[12] specifications. They are widely used by businesses.

Because of the increasing demand, users ubiquitous access to wireless services, which he led to the deployment of a forced use of this wireless technology such as Wi-Fi, it offers a level of quality in the range, but the problem is that the number of devices is increasing, which reduces the performance of travel time Handover and QoS[13] for all these reasons, this work proposes a solution for the improvement of vertical Handover by the creation of an algorithm which is implemented at the level of SDN controller, to better optimize the performances.

OpenFlow [14] is a first attempt to develop a wireless SDN platform with Stanford University. It separates the control plane from the data path and produces slices of network using FlowVisor [15] to isolate the different flows.

The underlying infrastructure is configured with SNMPVisor, a line interface command for configuring data path elements with Simple Network Management Protocol (SNMP) [16]. In other words, OpenFlow allows multiple different experiences and services to run simultaneously on a physical network.

Researchers are currently looking to implement the OpenFlow to WiFi access points and WiMAX base stations for traffic control and use a network controller, to communicate with OpenFlow devices and to provide global network views. FlowVisor can be considered a transparent proxy [17] for OpenFlow. It slices the network by selectively rewriting or dropping OpenFlow messages to delegate control of different streams with different controllers. The structure of OpenFlow is shown in figure 1 to separate the traffic of different users with multiple transmission policies.

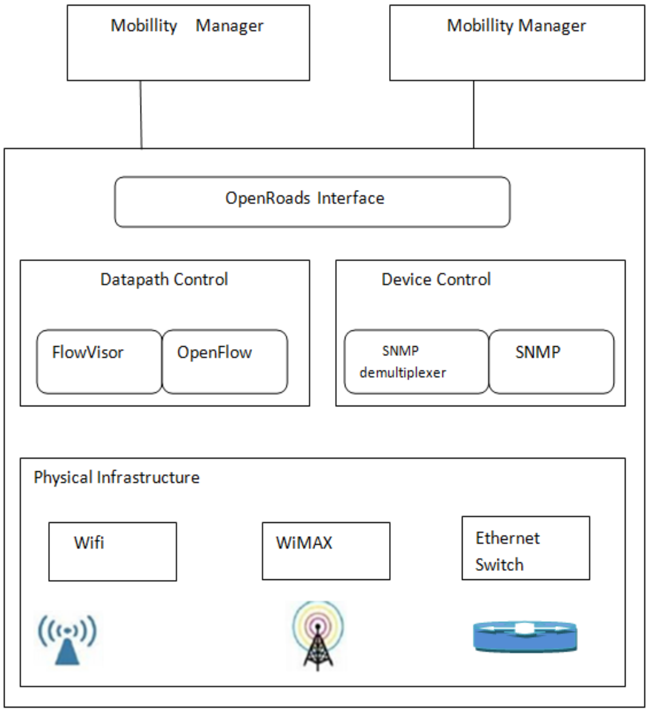


Fig. 1. Openflow Architecture.

Currently, the radio access network uses distributed algorithms to manage the limited spectrums and transmit Handover. While the decision in a little environment with few base stations could be simple to make, it would be harder to quickly pick the best candidate for deployment ladders. To manage growing mobile base station traffic, this research proposes that Software Defined Wireless Network (SDWN) [18] is a centralized system for software-defined radio access networks, to effectively perform transfers and allocate spectrum resources as well as to set spectrum values.

In this articlee, we propose a decision algorithm for a heterogeneous vertical Handover between WiFi access points and a WiMAX network capitalize. The inputs to the algorithm are WiFi RSS, bit rates, jitter, and estimated TCP end-to-end delay.

WiFi provides broadband connectivity for local area networks, while the WiMAX network provides broadband coverage in the metropolitan area with guaranteed quality of service. WiFi access points typically offer free access to good-track users, while WiMAX overlay networks offer paid access to users. It is therefore necessary to provide WiFi connectivity for as long as possible and to allow a roaming mobile device to switch over to a WiMAX accumulate network only when WiFi services are out of range or when its quality of service becomes unacceptable, which it means that WiMAX coverage is supposed to be always available and that the mobile terminal has to switch between WiFi depending on the availability of WiFi access points.

IV. Contribution

The following algorithm (Figure 2) shows the calculation steps implemented when approaching a mobile terminal, and when moving away from a WiFi access point, in both cases, the flow samples WiFi feeds are measured at regular time intervals and a moving average value Media RSS (MRSS) [19] is calculated for each RSS sample at the SDN controller level.

Considering first the case where the mobile terminal approaches the WiFi access point, if the calculated value of the MRSS is less than or equal to the sensitivity of the receiver, it means that the mobile terminal has already left the WiFi reception area. In this case, the WiFi Handover to WiMAX is immediately launched and centralized by the SDN controller. On the other hand, the MRSS value exceeds the property of the receiver, the position and the current speed of the mobile are calculated. These two values make it possible to determine the remaining time before the mobile terminal crosses the WiFi reception boundary. If this time exceeds the specified end-to-end TCP handshake delay, a prerelease routine is initiated by the SDN controller to ensure transparent Handover to the WiMAX network. This Handover pre-break routine has been adopted to ensure that Internet connectivity is maintained throughout the Handover process.

Considering thereafter, the case where the mobile terminal moves away from the WiFi access point. If the calculated value of the MRSS is greater than or equal to the sensitivity of the receiver, it means that the mobile terminal is already in the WiFi reception area. In this case, the WiFi WiMAX Handover is immediately launched. If, on the other hand, the MRSS value is lower than the sensitivity of the receiver, the current position and speed of the mobile are calculated. These two values make it possible to determine the remaining time before the mobile terminal enters the WiFi reception area. If this time exceeds the specified end-to-end TCP transfer latency, the transfer routine before the WiFi network cutoff is initiated. This routine makes it possible to provide connectivity to the Internet via the WiFi network as soon as the mobile terminal reaches the limit of its reception area. This ensures that the mobile terminal maximizes connectivity via the WiFi network.

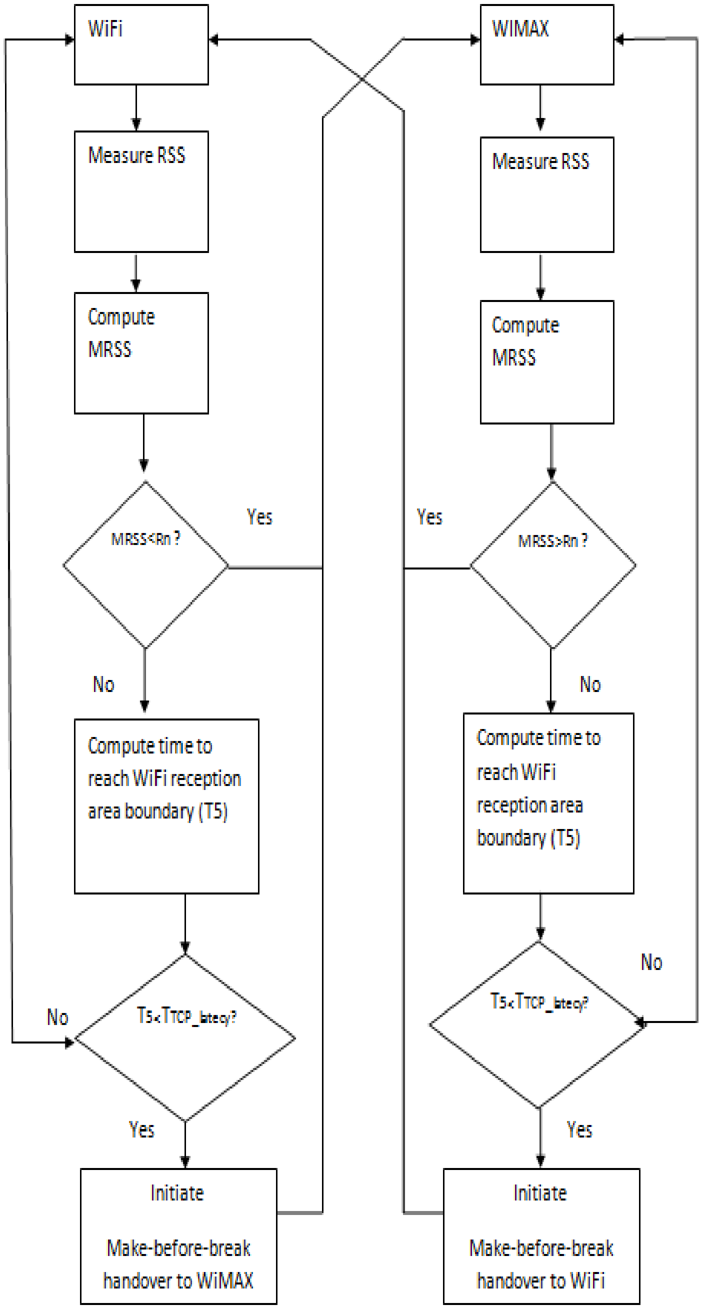

WiFi
WIMAX
Measure RSS
Measure RSS
Compute MRSS
Compute MRSS
MRSS<Rn ?
Yes
Yes
MRSS>Rn ?
No
No
Compute time to reach WiFi reception area boundary (T5)
Compute time to reach WiFi reception area boundary (T5)
No
No
T5<TTCP_latecy?
T5<TTCP_latecy?
Yes
Yes
Initiate Make-before-break handover to WiMAX
Initiate Make-before-break handover to WiFi

Fig. 2. The Flowchart of the Proposed Vertical Handover Algorithm.

V. Difference between WiMAX and WiFi

The following table presents a difference between WiFi technology and WiMAX.

TABLE I. DIFFERENCE BETWEEN WIMAX AND WIFI

Parameter	WiFi (802.11n)	WiMAX(802.16e)
Method of access	OFDM	OFDMA
Frequency band	2.4 or 5 GHZ	2.3-2.4, 2.496- 2.69, 3,3-3.8 GHZ
Max Throughput Downlink Uplink	540 Mbps	75 Mbps 25 Mbps
Cellule coverage	400m	2-7Km
Cellule capacity	32	100-200
Supported mobility	4Km/h	120Km/h
Service cost	Low	High
Energy consumption	Medium	High
Security	Medium	High
Quality of service	Yes	Yes

VI. Wireless Connection Modes

Two modes exist

The infrastructure mode is the most common. The machines communicate via an access point that it is only a kind of WiFi HUB.

The Ad Hoc[20] mode is more particular, since it makes it possible to dispense with the access point. Security possibilities in this mode are, however, almost non-existent. This mode should be avoided.

VII. Methods and Implementation of WiMAX and WiFi using SDN

The transparent Handover algorithm has been evaluated both in the case where a mobile terminal approach one WiFi access point to another. In each case, simulations were performed for typical speeds that the mobile terminal could encounter. The typical handover latency between WiFi and WiMAX for an end-to-end TCP connection varies from 450 ms to 1 second.

The distance from the WiFi reception limit at which the transfer process started has been recorded for transfer delays ranging from 10 ms to 1000 ms, this distance was the minimum distance, determined by our algorithm, for the Handover to be done transparently. The maximum reception distance of the WiFi access point is obtained using the path loss equation (1)

PlossdB =

$$20\log10\,(4\pi / \lambda) + 10\rho\,\log10d \qquad (1)$$

Where PlossdB is the signal loss in dB between the transmitter and the receiver;

λ is the wavelength of the WiFi signal

ρ is a path loss constant

d is the distance between the transmitter and the receiver.

The maximum reception distance for WiFi reception occurs when the path loss calculated by equation (1) above is equal to the difference between transmit power and receiver sensitivity. This algorithm uses a WiFi transmission power of + 17 dBm, an outside path loss factor of 2 and a receiver sensitivity of -76 dBm. Substituting these values in Equation (1) gives a maximum receiving distance of 994 meters.

This work consists of two vertical Handover scenarios between WiFi and WiMAX with and without SDN including

Scenario 1(Table 2, 3, 4): Vertical handover between WiFi and WiMAX without SDN: 3 AP, 7 WiFi Clients, 7 WiMAX Clients, and 3 BS. (Figure 3)

Scenario 2 (Table 2, 3, 4) Vertical handover between WiFi and WiMAX with SDN: 3 APs, 7 WiFi Clients, 7 WiMAX Clients, and 3 BS. (Figure 4)

The trajectory is defined as follows:

For the WiMAX scenario, the mobility speed is set at 20 km / hour;

Two meters per second in a WiFi network;

The WiFi 802.11b standard with a bandwidth of 2Mb was used to reach a theoretical coverage area of 400 meters.

The simulation parameters used in WiFi scenarios are listed in the table below:

TABLE II. BASE STATION PARAMETERS

Parameter	Value
Antenna Gain	15 dBi
Number of transmitters	SISO
Maximal transmission power	500 mW
PHY profile	OFDM
Maximal power density	-60 dBm
Minimal power density	-110 dBm
The resource retention time	200 msec

TABLE III. ACCESS POINT SETTINGS

Parameter	Value
PHY mode	Direct Sequence
Throughput	2 Mbps
Transmission power	0,005 W
Beacon interval	0,02 Secs
Buffer size	256 Kilobits

TABLE IV. APPLICATION SETTINGS

Parameter	**Value**
Traffic	VOIP
Codec	G729A
Voice frames per packet	1
Traffic generation	Continuously and infinite (from the start to end of the simulation)

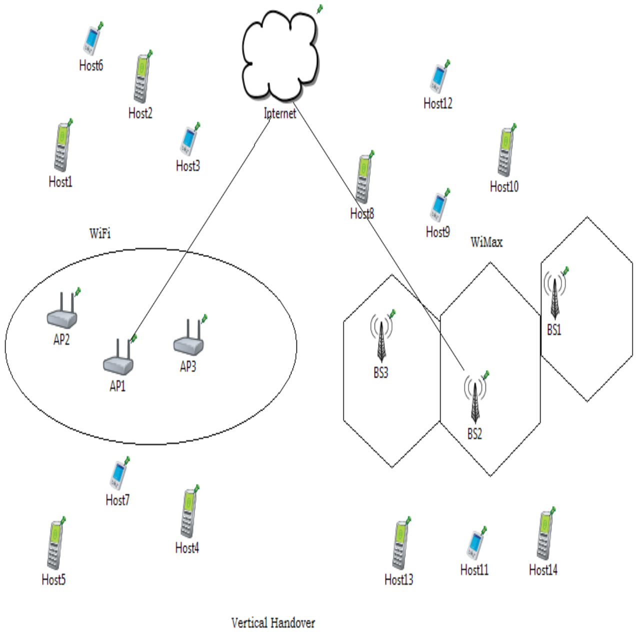


Fig. 3. Vertical and Horizontal Handover without SDN.

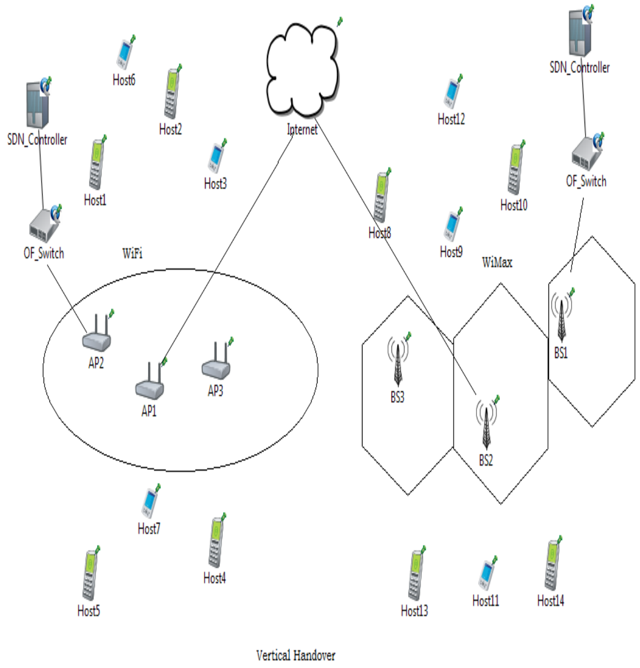


Fig. 4. Vertical and Horizontal Handover with SDN.

In order to take into account the mobility of users, the standard implements a Handover procedure that can be used in the following cases:

When the mobile station MS (Mobile Station) can be taken into account with a better signal quality by another base station (terminal movement, signal attenuation or interference).

When the mobile terminal can be taken into account with a better QoS by another base station (Load balancing [21], admission control, or QoS expectations).

VIII. IMPROVEMENT IN HANDOVER

Reducing the duration of a Handover is one of the goals. This can be done by developing a Handover protocol. Researchers are interested in reducing the duration of Handover by offering a cooperative Handover to prematurely prepare the Handover with the target network. The researchers are looking for a seamless vertical handover. For this, they present a new concept called "Takeover". This is to allow a neighbouring node in the recovery area to process requests from a mobile node that it wants to make a Handover before it can implement it. This is called Cooperative Handover as the nodes help each other with the Handover. This reduces the Handover time: Pre-authentication and preregistration time (Using a mobile IP registration procedure). In addition, a protocol for the Takeover has been developed and applied. Handover's decision is based on the signal quality of the two base stations. On the other hand, the proposed system requires more signalling and also more processing by the neighbouring node. This finally requires operation in mesh mode (Direct communication between terminals).

IX. MOBILITY MANAGEMENT

In fixed broadband access, WiMAX distinguishes four types of mobility related to the circumstances of use:

Nomad: In this case, a fixed place is assigned to the user for the use of the services. In order to connect to a different location, the user must make a break or disconnect [22];

Portable: Using nomadic access with a portable device, portability is ensured for the user at the cost of the Handover [23];

Simple mobility: Almost uninterrupted Handovers are made with users who can reach a speed of 60 km / h in very short time intervals [24];

Full mobility: The user can move up to a speed of 120 km / h [25].

The Handover is the mechanism that ensures the continuity of the connection of a subscriber’s station during its movement of the coverage area from one base station to another.

X. HANDOVER PROCESS

The Handover process can be carried out in three phases:

Transfer Information Collection: Also known as System Discovery, the information required to identify the need transfer is collected. In this phase, information from all neighbouring networks is collected, and they can also be called the system discovery phase.

Transfer decision: This process finds the appropriate candidate network to which the mobile terminal "MT" can be transferred according to certain decision algorithms.

Transfer execution: Finally, the signalling exchange for the establishment of the new communication path was carried out with the rerouting of data via this path.

Network Selection: The Horizontal Transfer Decision Algorithm includes the strength of the received RSS signal. For the Vertical transfer decision, many criteria can be taken into account, such as: Cost of service, power consumption, mobile device speed, and user preference.

XI. Results and Discussion

The following figure (5) shows that WiMAX does not require more delay, compared to WiFi, for WiFi to WiMAX transfer or vice versa, compared to WiFi, this is quite logical because the WiMAX network was designed primarily to connect a very large number of users with a high bandwidth with a minimum of time.

Figures 6 illustrates the network delay of WiFi and WiMAX scenarios using SDN In WiMAX, and WiFi, they represent the end-to-end delay of all packets received by WiMAX or WiFi MACs, all nodes of the network, and they transmit to the upper layer.

MN can discover a suitable WiFi network via a request-response

Transfer to WiFi based on a number of factors such as the quality of service, power, cost, etc. If this discovery is successful, the mobile initiates the transfer procedures.

After transferring to WiFi, the MN can choose to set the WiMAX radio to sleep mode. This allows the MN to quickly return to WiMAX in case of WiFi, the coverage is abruptly degraded. Figure 7 shows the process of transfer between WiMAX and WiFi.

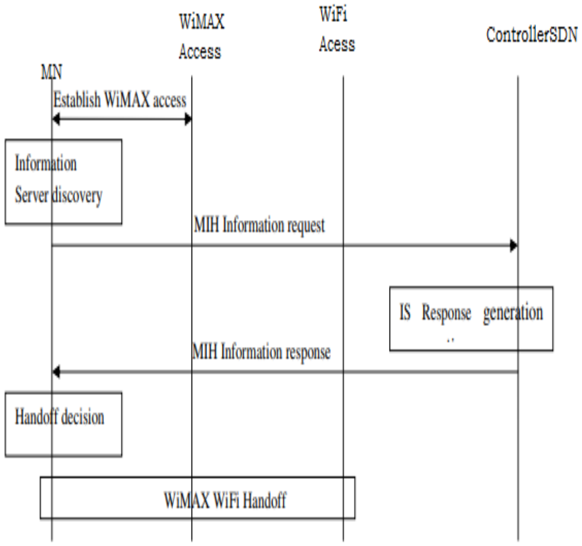


Fig. 5. Handover from WiMAX to WiFi using SDN Controller.

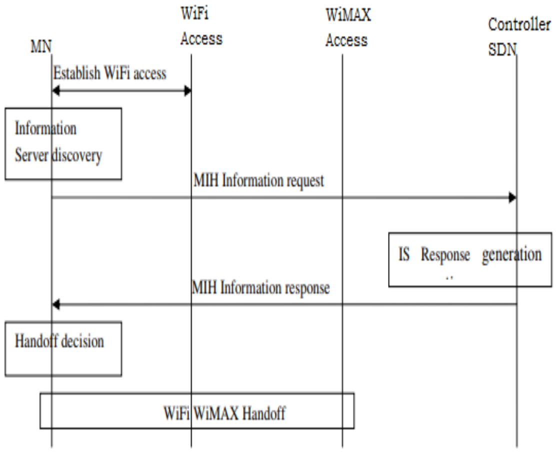


Fig. 6. Handover from WiFi to WiMAX using SDN Controller.

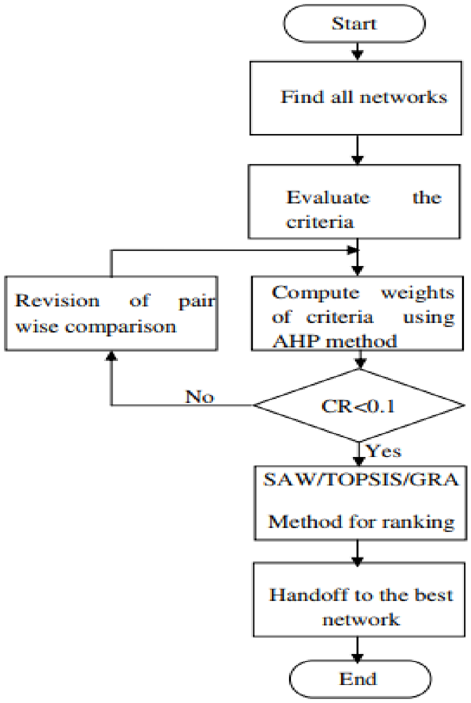

Start
Find all networks
Evaluate the criteria
Revision of pair wise comparison
Compute weights of criteria using AHP method
No
CR<0.1
Yes
SAW/TOPSIS/GRA
Method for ranking
Handoff to the best network
End

Fig. 7. Multi-Criterion Handover Decision Algorithm.

A. *Delay between Wi-Fi and WiMAX Networks with and without SDN*

Figure 8 illustrates the network delay by WiFi and WiMAX scenarios with and without SDN using OMNeT++4.6, In WiMAX, and WiFi, they represent the end-to-end delay of all packets received by WiMAX or WiFi MACs of all nodes of the network, and they transmit to the upper layer.

According to the following figure, we observe that WiMAX does not require more delay, compared to WiFi to make the transfer from WiMAX to WiMAX or vice versa, compared to WiFi, This is quite logical because the WiMAX network has been designed primarily to connect a very large number of users with high bandwidth with a minimum of time.

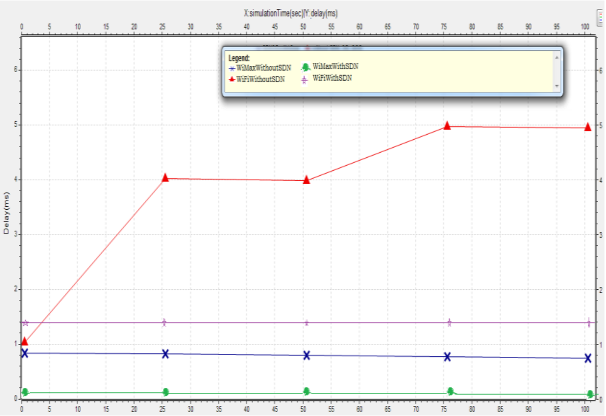


Fig. 8. Time delay between WiFi and WiMAX Networks with and without SDN.

B. *Throughout between WiFi and WiMAX Networks with and without SDN*

WiMAX is intended to connect a large number of clients, that's why we find its speed is higher or without SDN and more scalable with the implementation the SDN. As a results (Figure 9), WiMAX remains the most scalable compared to WiFi network with and without SDN.

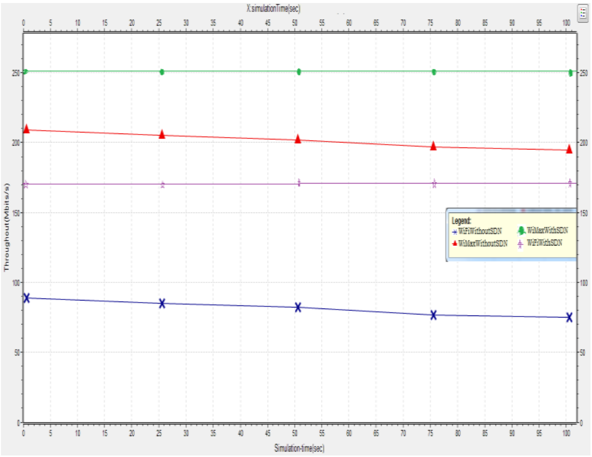


Fig. 9. Throughout between WiFi and WiMAX Networks with and without SDN.

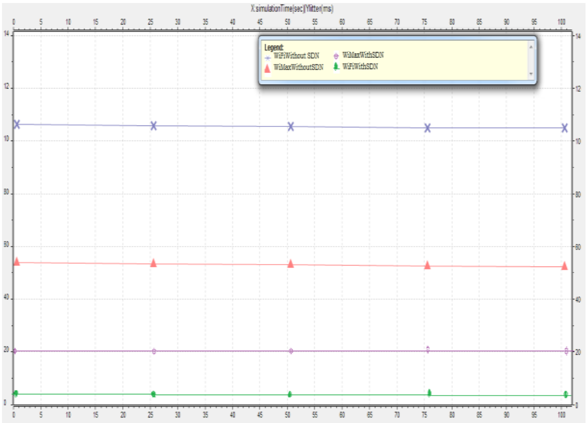


Fig. 10. Jitter between WiFi and WiMAX Networks with and without SDN.

C. *Jitter between WiFi and WiMAX Networks with and without SDN*

The results in Figure 10 shows that the jitter presented by WiFi and WiMAX scenarios with SDN is low compared to without SDN which requires more jitters using OMNeT++4.6, which justifies the impact of the addition of SDN is important.

XII. Conclusion

This article presents performance improvement and evaluation of vertical handover between the WIFI and WiMAX by the implementation of a new algorithm at the level of the SDN controller which allows the handover decision by RSS, bit rate, jitter, and estimated delay using OMNeT++4.6. This algorithm is better placed to anticipate and initiate the transfer before it is required. This would include identifying and configuring wireless and TCP/ IP connections to the next access point before the actual handshake process.

References

[1] Jeffrey G. Andrews, Arunabha Ghosh, Rias Muhamed ''Fundamentals of WiMAX Understanding Broadband Wireless Networking'', 2007

[2] Jolly Parikh, Anuradha Basu, ''LTE Advanced: The 4G Mobile Broadband Technology'', International Journal of Computer Applications (0975 – 8887) Volume 13– No.5, January 2011

[3] Fèten RIDENE RAISSI, et Adel RAISSI, ''Authentification dans les Réseaux Wifi par le protocole radius'', Thesis, 2010

[4] Gamal Abdel , Fadeel Mohamed Khalaf, Hesham Zarief Badr, ''A comprehensive approach to vertical handoff in heterogeneous wireless networks'', Journal of King Saud University – Computer and Information Sciences (2013) 25, 197–205, November 2012

[5] Slavisa Tomic, Marko Beko, Rui Dinis, ''RSS-Based Localization in Wireless Sensor Networks Using Convex Relaxation: Noncooperative and Cooperative Schemes'' , IEEE Transactions on Vehicular Technology, Volume: 64, Issue: 5 , May 2015

[6] R. Les Cottrell, Saad Ansari, Parakram Khandpur, Ruchi Gupta, Richard Hughes-Jones, Michael Chen, Larry McIntosh, Frank Leers, ''Characterization and evaluation of TCP and UDP-based transport on real networks'', Volume 61, Issue 1–2, pp 5–20, February 2006

[7] Seok-Yee Tang, Peter Muller, Hamid R. Sharif, ''WiMAX Security and Quality of Service, A John Wiley and Sons, Ltd., Publication, 2007

[8] Hung-Yu Wei ; S. Ganguly ; R. Izmailov ; Z.J. Haas, ''Interference-aware IEEE 802.16 WiMAX mesh networks'', 2005 IEEE 61st Vehicular Technology Conference, June 2005

[9] Mojtaba Seyedzadegan, and Mohamed Othman, ''IEEE 802.16: WiMAX Overview, WiMAX Architecture'', International Journal of Computer Theory and Engineering, Vol. 5, No. 5, DOI: 10.7763/IJCTE.2013.V5.796, October 2013

[10] Haihong Zheng, Shashikant Maheshwari, Basavaraj Patil, Srinivas Sreemanthula, ''Framework for internetworking between WMAN and WLAN networks'', 2008

[11] Paul Mühlethaler, ''802.11 et les réseaux sans fil'', book, EAN13 : 9782212111545, 2002

[12] Wen-Hsing Kuo, Wanjiun Liao, Tehuang Liu, ''Adaptive Resource Allocation for Layer-Encoded IPTV Multicasting in IEEE 802.16 WiMAX Wireless Networks'', IEEE Transactions on Multimedia, Volume: 13 , Issue: 1 , Feb. 2011

[13] Mohamed El Mahdi Boumezzough, Noureddine Idboufker, Abdellah Ait Ouahman, ''Evaluation of SIP Call Setup Delay for VoIP in IMS'', IEEE International Conference on Advanced Infocomm Technology, ICAIT 2012: Advanced Infocomm Technology pp 16-24, 2012

[14] Fatima LAASSIRI, Mohamed MOUGHIT, Noureddine IDBOUFKER, ''Evaluation of the QoS parameters in different SDN architecture using OMNeT 4.6++'', 2017 18th International Conference on Sciences and

Techniques of Automatic Control and Computer Engineering (STA), 2017

[15] Raghunath Deshpande, ‘‘Overview of Different Approaches for Leveraging SDN in Mobile Networks’’, https://www.researchgate.net/publication/282322426, February 2015

[16] B. Wijnen, R. Presuhn, K. McCloghrie, ‘‘View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)’’, Network Working Group, Copyright (C) The Internet Society (2002)

[17] Anders Moberg, Dmitry M. Sonechkin, Karin Holmgren, Nina M. Datsenko, Wibjörn Karlén, ‘‘Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data’’, Nature volume433, pages613–617, February 2005

[18] Honglin Hu ; Hsiao-Hwa Chen ; Peter Mueller ; Rose Qingyang Hu ; Yun Rui , ‘‘Software defined wireless networks (SDWN): Part 1 [Guest Editorial]’’, IEEE Communications Magazine, Volume: 53 , Issue: 11 , November 2015

[19] Soujanya Bhumkar, Mayank Mehta, Josh Schwarzapel, Austin Shoemaker, ‘‘Method and system for displaying photos, videos, RSS and other media content in full-screen immersive view and grid-view using a browser feature’’, 2007

[20] Charles E. Perkins, ‘‘Ad Hoc Networking’’, CE Perkins - 2001 - academia.edu [21] M.E. Baran, F.F. Wu, ‘‘Network reconfiguration in distribution systems for loss reduction and load balancing’’, IEEE Transactions on Power Delivery, Volume: 4 , Issue: 2 , Apr 1989

[21] Niranjan Suri, Jeffrey M. Bradshaw, Maggie R. Breedy, Paul T. Groth, Gregory A. Hill, Renia Jeffers, Timothy S. Mitrovich, Brian R. Pouliot, David S. Smith, ‘‘toward a strong and safe mobile agent system’’, ISBN:1-58113-230-1, Barcelona, Spain — June 03 - 07, 2000

[22] Mona Laroussi, Alain Derycke, Trigone-CIREL, ‘‘new e-learning services based on mobile and ubiquitous computing: Ubi-learn project’’, International Conference on Computer Aided Learning in Engineering education, 16-18 février 2004, 2004, Grenoble, France. 6 p., 2004.

[23] Zdenek Becvar, Jan Zelenka ‘‘Handovers in the Mobile WiMAX’’,https://www.researchgate.net/profile/Zdenek_Becvar/publication/229049396_Handovers_in_the_Mobile_WiMAX/links/00b7d51521 9902288b000000.pdf 2006

[24] Niranjan Suri, Jeffrey M. Bradshaw, Maggie R. Breedy, Paul T. Groth, Gregory A. Hill, Renia Jeffers, ‘‘Strong Mobility and Fine-Grained Resource Control in NOMADS’’, LNCS, volume 1882, , pp. 2-15, Springer-Verlag Berlin Heidelberg, 2000

The Utilization of Feature based Viola-Jones Method for Face Detection in Invariant Rotation

Tioh Keat Soon[1], Abd Samad Hasan Basari[2], Nuzulha Khilwani Ibrahim[3], Burairah Hussin[4],
Ariff Idris[5], Noorayisahbe Mohd Yaacob[6]
[1,2,3,5,6] BIOCORE and [4]OptiMAS Research Group, Center for Advanced Computing Technology (C-ACT)
Department of Intelligent Computing and Analytics
Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Mustafa Almahdi Algaet[7]
Department of Computer, Faculty of Education,
Elmergib University,
Alkhums, Libya

Norazira A. Jalil[8]
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman (Perak Campus),
Jalan Universiti Bandar Barat, 31900 Kampar, Perak

***Abstract*—Faces in an image consists of complex structures in object detection. The components of a face, which includes the eyes, nose and mouth of a person differs from that of ordinary objects, thus making face detecting a complex process. Some of the challenges encounter posed in face detection of unconstrained images includes background variation, pose variation, facial expression, occlusion and noise. Current research of Viola-Jones (V-J) face detection is limited to only 45 degrees in-plane rotation. This paper proposes only one technique for the V-J detection face in unconstrained images, which V-J face detection with invariant rotation. The technique begins by rotating the given image file with each step 30 degrees until 360 degrees. Each step of adding 30 degrees from origin, V-J face detection is applied, which covers more angles of a rotated face in unconstrained images. Robust detection in rotation invariant used in the above techniques will aid in the detecting of rotated faces in images. The images that have been utilized for testing and evaluation in this paper are from CMU dataset with 12 rotations on each image. Therefore, there are 12 test patterns generated. These images have been measured through the correct detection rate, true positive and false positive. This paper shows that the proposed V-J face detection technique in unconstrained images have the ability to detect rotated faces with high accuracy in correct detection rate. To summarize, V-J face detection in unconstrained images with proposed variation of rotation is the method utilized in this paper. This proposed enhancement improves the current V-J face detection method and further increase the accuracy of face detection in unconstrained images.**

***Keywords*—Face detection; V-J face detection; unconstrained images; bicubic interpolation; SIFT**

I. Introduction

There are mainly two methods that can be utilized for face detection. It is either a feature-based method or image-based method [1]. In this paper, feature based method is selected for face detection. Feature based method include skin colour, facial features and blob features. The advantages of the feature-based method are due to its rotation independence, scale independence, and quick execution time compare to image-based method [2]. Face detection is widely used in a multitude of preliminary applications. Face detection is utilized to locate a face or faces in an image. Face recognition, on the other hand, is utilized to find out who the person in the image is after face detection has been performed. Therefore, the preliminary accuracy of face detection is crucial to support face recognition. It is similar to the application of CCTV surveillance with built-in face recognition for the security purposes. Improving face detection for CCTV surveillance cameras will ensure that the faces of people can be easily identified compared to modern CCTV footage that produces blurry images. Modern cameras come with built-in functions to auto-focus on the face region. By being able to detect the face accurately, only then can unwanted red-eye effects be corrected. Another utilization of face detection can be seen in marketing methods in order to gather information on the types of customers that frequently pass by certain areas. The proposed V-J face detection method allows the detection of the faces of customers in different angles on the same plane. By utilizing customer classification, businesses can predict what type of customers is interested in certain product for advertisement purposes.

II. Related Work

According to [3], there are four categories under face detection methods. They are the feature invariant approach, knowledge-based method, template-based method and appearance-based method.

Knowledge based method is known as the Rule based method. This method translates human knowledge of face features into a set of rules. These rules include the relationship of facial features. For instance, the intensity of the eye is darker than the forehead of face. Another example of frontal face in images is often with 2 symmetrical eyes, a nose and a mouth. The features then are represented as the distance and positions. The limitation of Rule based method is that it may lead to high false positive if it is too general whereas false negative may increase if it is too detailed. Hierarchical knowledge based is introduced to overcome the problems. However, it has the limited solution to find multiple faces in a complex image with the solution alone.

In contrast to knowledge based, feature invariant method aims to find structures of face features regardless of lighting conditions, different scaling and angles in complex images. Numerous feature invariants have been proposed such as human skin color, blob detection and moment to detect face features which then moves forward to classify face region. Merits of face detection based on human skin colour have a faster execution time in face detection despite different scaling and angles. Usually, it is utilized in preliminary process for dimension reduction to improve the speed detection. There are several types of colour space. Usually, there are 6 colour spaces for skin colour face detection. Face detection that based on colour space are YCbCr, RGB, HSI, nRG, HSV, and CIE. However, skin colour based method has skin colour-like background challenge. The author in [4] proposed to use skin colour with edges. YCbCr colour space was selected and classified skin or non-skin types by using Gaussian Mixture Model. Sobel edge detection was utilized after binary process. At least 3 'holes' was created with Euler formula if a face has been detected. However, the researcher found out that fault detection was due to over-bounding if other regions are similar to skin colour. In order to resolve the challenge, skin modeling coefficient matrix technique and improved Gaussian distribution are proposed. The author in [5] explicit defined algorithm is chosen for the development due to simplicity and speed performance. Skin modeling coefficient matrix is used for segmentation process skin pixels or non-skin pixels. The Robert edge detection method was performed before post-processing. Connected component analysis is performed after post-processing. Finally, 2 conditions of aspect ratio must be fulfilled to classify face or non-face. The author in [6] proposed to use skin-colour model (RGB colour space), facial features (labial feature and holes feature) and improved Gaussian distribution model to detect multiple faces with good performance and remove skin colour-like background. Another challenge of skin colour is to resolve illumination problem. The author in [7] proposed to use the skin colour (HSV) in different range for indoor and outdoor environment. Erosion method was used to remove small non-face objects after it was converted to binary black and white. Active snake contour method was selected to detect maximum contour area. However, the researcher suggests changing to automatic threshold for better detection rate especially outdoor environment if the illumination is too bright. Noise challenge is further removed. The author in [4] proposed low pass filter was used to eliminate noises. Threshold value was determined via average sum of median and maximum values column scanning. Blob is also considered as an interest point in face detection. It has been widely used for face detection. For instance, for the blobs are Haar features, corner detection, Laplacian of Gaussian, Difference of Gaussian and component labeling. Later, the blobs are further analyzed by extracting the information of shapes of objects that are present in the image. This technique is also referring to image segmentation. Result of feature extraction is to identify the number of different objects, region information and other salient features. At an early stage, [8] was one of the first corner detection being carried out for interest point. It was improved by [9] to remove the noise. The author in [9] applied Gaussian to autocorrelation matrix for corner detection. However, it is limited to scaling invariant. The author in [10] proposed SIFT to overcome the scaling variant problem. It was dependent on the sigma or standard deviation. The author in [11] showed that the Laplacian response is decayed when the standard deviation or scale getting bigger. Superposition of two ripples results in the maximum response becoming blob-like. To keep the Laplacian response the same across the scale, second order of Gaussian must multiply by σ^2. In [10], the author proposed to use Difference of Gaussian (DoG), which is approximate for LoG. Optimization is improved by using DoG. Corner detection results in rotation invariant but not in scale variant [9]. Scale space theory was introduced and there are two important steps. These two steps are known as i) feature detection and ii) finding maxima and minima extrema. The author in [11] introduced automatic scale selection. There are many blob detections based on LoG or DoG in scale space. For instance, Determinant of Hessian (DoH), SIFT [10], Harris Laplacian [12], Hessian Laplacian [11], and Harris Affine Region [13]. The author in [14] proposed in-plane angle estimation for face images from multi-poses by applying SIFT to 2 reference points, which are midpoint of eyes and nose. The appearance descriptors consist of SIFT descriptors such as location, scale and orientation of reference points. 2 hypotheses were used to determine face or non-face via Bayesian classifier. The proposed result outperformed in terms of low false face detection rate, low in-plane rotation error and speed performance. The author in [15] proposed an improved Haar-like feature so called Haar Contrast Feature, which efficiently for object detection under various illuminations with the Haar Wavelet based. The LoG can be represented by Haar Wavelet which proposed by [16]. Computation of Haar Wavelet can be done by utilizing integral imaging method. This method has speed up the process. The author in [17] proposed heterogeneous feature descriptors and feature selection for efficient and accurate face detection. To address the issue of distinctive representation for face patterns, the researcher proposed complementary feature descriptors Generalized Haar-like descriptor, Multi-Block Local Binary Patterns descriptor and Speeded-Up Robust Features (SURF) descriptor. Particle Swarm Optimization (PSO) algorithm was integrated into the Adaboost framework as feature selection and classifier learning. A three-stage hierarchical classifier structure and nonlinear support vector machine (SVM) classifier were used to rapidly remove non-face patterns. The experiment was tested on CMU+MIT data set. The proposed solution also worked well for faces with Yaw rotation between ±22.5°. The results show robustness and efficiency of the proposed solution with other state-of-the-art algorithms. The author in [18] proposed to use normalized RGB colour space to determine skin. Blob detection (Connected Component) was used later. The researcher [19] made a time consumption and accuracy comparative study of SIFT and its variants such as GSIFT, PCA-SIFT, SURF, ASIFT, and CSFIT in 4 situations. The results showed that, in scale and rotation situations, SIFT and CSIFT performed better compared to other variants. In affine image, ASIFT performed better compared to others. SURF gain the fastest speed performance compared to others. In blur or illumination image, GSIFT performed better compared to others.

There are two techniques of face detection based on template matching. Template image includes either the face as a whole or face features separately. The stored predefined face features with eye, mouth, and nose are known as deformable template matching. It is then the stored predefined face features are correlated with the input face. For instance, template is matched with the input image through slide windows. However, it is limited to achieve better result with the variant of scale and pose. Deformable template is introduced to overcome the problems. In [20] enhanced the winner-update algorithm (WUA) with winner-update and integral image (WUI) for fast and full search algorithm. These algorithms were used for reducing the computational complexity. By exploiting the integral image, the method gained the speed performance. The author in [21] adopted template matching method for design pattern detection. This method was not only utilized to detect exact pattern, but on variation of patterns as well, based on normalized cross correlation.

Appearance based method is similar to template matching but learning from a set of stored example face images. This method depends on statistical analysis and machine learning. For instance, statistical analysis based on probability to determine a face or not. There are a lot of machine learning in this method such as logistic regression, discrimination analysis and other binary classification. Based on appearance method, statistical analysis is also known as feature representation. Example of feature representation methods are Haar feature, skin colour and shape. Usually machine learning used in face detection is mainly for feature selection. Most feature selection methods are based on machine learning such as Adaboost, neural-network and SVM. Pattern classification is a method to classify pattern vectors into several classes. It is often referred to machine learning in artificial intelligence field. The methods could be classified into supervised learning, unsupervised learning, reinforcement learning, evolutionary learning, and ensemble learning. Supervised learning is usually having past historical data and class of the subset of the data. Unsupervised learning is same as supervised learning but without knowing the class of the subset of the data. One of the famous ensembles learning method is Adaboost. It was utilized for the face detection classification. The method that helped to improve the training time during performing Adaboost is the Cascaded method. Most recent research is focusing on machine learning. They are multilayer Neural Network, Support Vector Machine, Adaboost, Hybrid Adaboost and Support Vector Machine, Model Based, Discriminant Analysis method and Deep Learning.

A recent progress of face detection is rotation invariant, fast speed detection, quality of the image which includes illumination, noise and blur. According to knowledge-based method, [22-23] proposed morphological technique to detect face. It is limited to accuracy of edge detection and multi face. The authors in [24-27] focus on deep convolutional Neural Network. However, appearance based requires more data to do the training and it is time consuming. The authors in [28-34] focus on V-J face detection. Survey study [35] on several techniques regarding the extraction and learning algorithms including Local Binary Pattern (LBP), Adaboost algorithm, SNOW classifier, SMQT features and Neural Network-Based face detection. It shows that V-J face detection is faster and accurate for frontal face detection.

Related works of V-J face detection. The author in [36] proposed real-time face detection. The author further proposed pose estimation during the Haar features training which covers (±15°), 30° covers (15° - 45°), 60° covers (45° - 75°), 90° covers (75° - 105°) until 360° with 12 detectors. However, training on rotation of Haar features require longer training time. Viola et al., 2004 continued to propose more robust real-time face detection but limited to ±15° only. The author in [37] proposed rotate input sub-windows with ±30° which could cover up to ±45° from 0°. The author in [14] proposed feature transform which covers ±10° only. Based on the previous authors, most of them focused on mainly speed and only a minor contribution to accuracy. Many of them are based on training method. In 2004, Viola and Jones [38] took about 2 weeks to complete the training, which was only limited to ±15° in-plane rotation. This thesis extends the rotation method from Li and Yang [37] which covers 360°. The method gained significantly better accuracy on in-plane rotation with low false positive without a longer training required.

Studies in [24-25, 39-41] tested the face detection from Face Detection Dataset and Benchmark (FDDB). It contains more than 5000 unconstrained faces such as large appearance variation in pose, occlusion, expression, illumination, and imaging conditions. Study [24] did the training dataset from Feret, PIE database which contains face with different poses, frontal, left/right half profile, and 0 till 30 degree in-plane rotations. Study [42] tested face detection from Feret database. Study [22] tested FEI database contains facial images including facial expressions, occlusion, lighting conditions, and background complexities. Studies [23, 40, 43] tested on IMM frontal face database contains variance of lighting conditions which was recorded in 2005 by Fagertun and Stegman at Technical University of Denmark. Study [23] tested face detection from FEI database. Studies [24, 44, 46] tested BioID database which consists of 1521 grey images with 384x286 pixels dimensions. Studies [4, 40, 32] tested Bao database that contains family images. Study [32] tested LFW database. Study [40, 43] tested Caltech database. Study [47] tested with XM2VTS contains occlusion faces. Study [5, 14, 44, 45, 48] tested and training from MIT+CMU database. Studies [14, 40] evaluated the testing from CMU dataset which contains total 50 face images with in-plane rotation and some with multiple faces. Study [14] tested with good quality images where some with poor quality of images were removed, left 40 images and 65 faces.

III. Proposed V-J Face Detection in Unconstrained Images

This paper follows pattern recognition methodology by [49], which refer to the Figure 1. According to [49], there are 2 ways to do the recognition, either classification phase or training phase. Training phase can be incorporated into classification method. The included experiment has been based on our previous publication in [50].

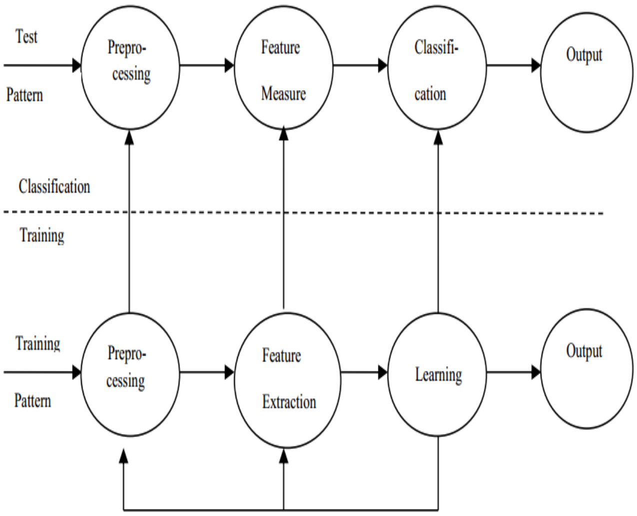


Fig. 1. Pattern Recognition Methodology

The methodology starts with pre-processing the enhanced in-plane rotation image file so that faces in different angles could be detected. Then, the evaluation of V-J face detection is evaluated in established databases to proof it is more accurate in V-J face detection in unconstrained images. There is only one main part research design of V-J face detection in unconstrained images in this thesis. It involves the enhanced rotation in V-J face detection only. The data type utilized in this paper is grey colour images. The rotated face images are from established database CMU. CMU consists of grey colour with 50 rotated face images only. The CMU images have different size of image files with grey format. Study [37] utilized 50 CMU input images as image dataset. The enhanced rotation of V-J face detection consists of two parts 1) Rotation process, 2) Face detection process.

A. Rotation Process

The parameter of θ is represented as radian of the rotation. A negative angle represents clockwise rotation whereas positive value represents anticlockwise rotation. The enhancement of V-J face detection is carried out with rotated 30°, 60°, 90°, 120°, 150°, 180°, 210°, 240°, 270°, 300°, 330° and 0°. In mathematics, rotation [51] is formulated as (1).

$$[x', y'] = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} [x, y]^T \tag{1}$$

B. Face Detection Process Units

Equation (2) shows the grey colour conversion.

$$\mathrm{grayColour} = (\mathrm{R} + \mathrm{G} + \mathrm{B})/3 \tag{2}$$

In Viola Jones's face detection, Figure 2 shows Haar features are represented in two-rectangle features, three-rectangle features and four-rectangle features. The value of the rectangle feature is the sum of difference between black region and white region.

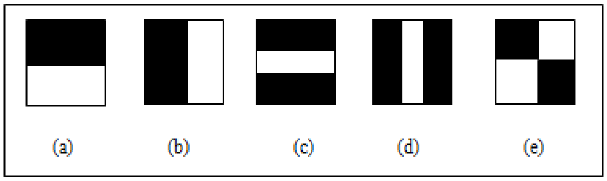


Fig. 2. Haar Features

In order to speed up computation of rectangle-feature, Viola Jones proposed integral image. There are four regions (Figure 4) represent A, B, C, and D. To compute D region, D=(X4,Y4)+(X1,Y1)-(X2,Y2)-(X3,Y3). Figure 3 shows integral images calculation.

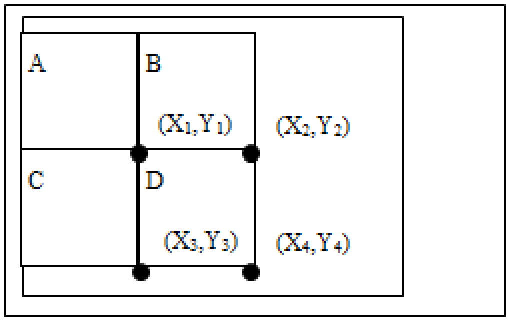


Fig. 3. Integral Image Calculation

Equation (3) shows variance normalization

$$\sigma^2 = \mathrm{M}^2 - \frac{1}{\mathrm{N}}\sum \mathrm{X}^2 \tag{3}$$

Where σ is the standard deviation,

M is the mean,

N is the region size and

X is the pixel value within the region.

Adaboost is the machine learning algorithm. It will form a strong classifier. The Adaboost performed the feature selection for Haar features in face detection. Figure 4 shows sample of rotation from CMU.

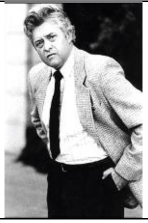
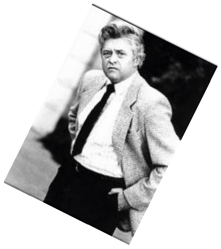
No	Angles	Samples
(a)	0°	
(b)	30°	

Fig. 4. Sample of Rotation

Figure 5 shows proposed V-J face detection in unconstrained images.

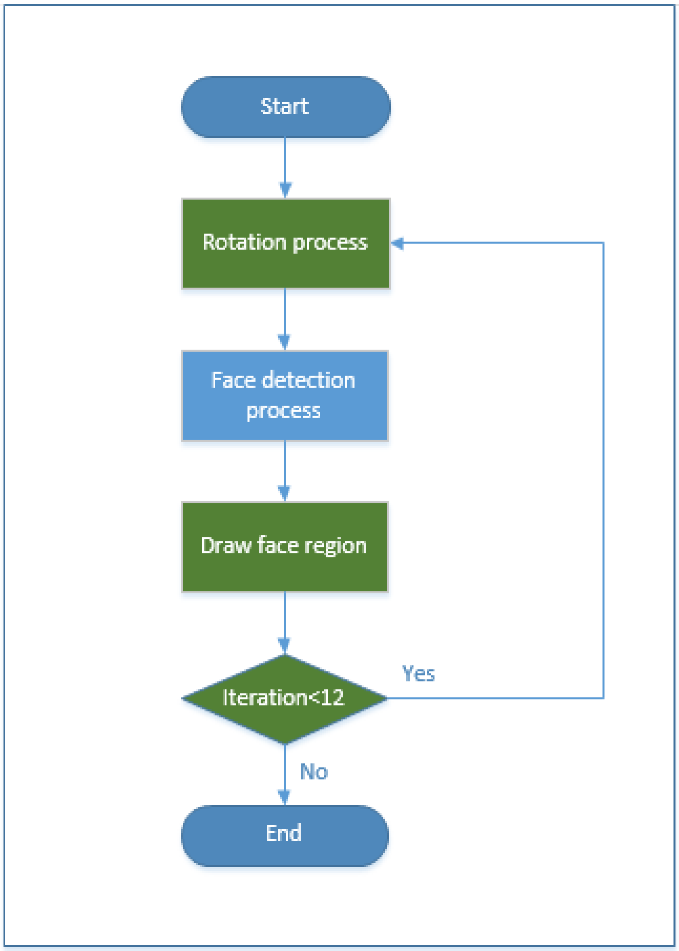


Fig. 5. Proposed V-J Face Detection In Unconstrained Images

IV. Conclusion

The proposed V-J face detection in unconstrained images method has several better achievements to meet the objectives. The accuracy performance by rotating image file with 30° each step within 360° before performing V-J face detection which meets the pattern recognition methodology. V-J face detection provides standard couple of solution with rotation. The accuracy performance is increased by providing flexibility and prior knowledge to any face detection as pre-processing. The proposed V-J face detection in unconstrained images is significant better accuracy than previous method, which are demonstrated the results by the number of unconstrained images. For future works, the interpolation can be combined with rotated face to enhance the rotation accuracy. Besides that, the SIFT can be combined with convolutional neural network to find the eye region for accuracy of face detection.

Acknowledgment

This paper is part of PhD work in the Faculty of Information and Communication Technology, UTeM and funded by MyPhD scholarship under the Ministry of Higher Education, Malaysia.

References

[1] Hatem, H., Beiji, Z., and Majeed, R., 2015. A Survey of Feature Base Methods for Human Face Detection. International Journal of Control and Automation, 8 (5), pp.61–78.

[2] Hu, W.-C., Yang, C.-Y., and Huang, D.-Y., 2011. Feature-based Face Detection against Skin-color Like Backgrounds with Varying Illumination. Journal of Information Hiding and Multimedia Signal Processing.

[3] Mekami, H. and Benabderrahmane, S., 2010. Towards A New Approach for Real Time Face Detection and Normalization. 2010 International Conference on Machine and Web Intelligence, pp.455–459.

[4] See, Y.C., Noor, N.M., and Lai, A.C., 2013. Hybrid Face Detection with Skin Segmentation and Edge Detection. 2013 IEEE International Conference on Signal and Image Processing Applications, pp.406–411.

[5] Patwary, A.A.A.G.M.N., 2012. Multiple Face Detection Algorithm Using Colour Skin Modelling. Image Processing, IET, 6 (8), pp.1093–1101.

[6] Liu, Z., Sha, J., and Yang, P., 2013. Multi-face Detection Based on Improved Gaussian Distribution. 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, pp.54–57.

[7] Zahir, N.B., Samad, R., and Mustafa, M., 2013. Initial Experimental Results of Real-Time Variant Pose Face Detection and Tracking System. 2013 IEEE International Conference on Signal and Image Processing Applications, pp.264–268.

[8] Moravec, H., 1980. Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. Tech Report CMU-RI-TR-3, Carnegie-Mellon University, Robotics Institute.

[9] Harris, C. and Stephens, M., 1988. A Combined Corner and Edge Detector. In Fourth Alvey Vision Conference, pp.147–152.

[10] Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60, pp.91–110.

[11] Lindeberg, T., 1998. Feature Detection with Automatic Scale Selection. International Journal of Computer Vision, 30 (2), pp.77–116.

[12] Mikolajczyk, K. and Schmid, C., 2001. Indexing Based on Scale Invariant Interest Points. Proceedings. Eighth IEEE International Conference on Computer Vision, 1, pp.525–531.

[13] Mikolajczyk, K. and Schmid, C., 2004. Scale & Affine Invariant Interest Point Detectors. International Journal of Computer Vision, 60 (1), pp.63–86.

[14] Mohammad, S., Anvar, H., and Star, A., 2013. Estimating In-Plane Rotation Angle for Face Images from Multi-Poses. Computational Intelligence in Biometrics and Identity Management (CIBIM), 2013 IEEE Workshop, pp.52–57.

[15] Park, K.-Y. and Hwang, S.-Y., 2014. An Improved Haar-Like Feature for Efficient Object Detection. Pattern Recognition Letters, 42, pp.148–153.

[16] Bay, H., Ess, A., Tuytelaars, T., and Gool, L. Van, 2008. SURF : Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), 110 (3), pp.346–359.

[17] Pan, H., Zhu, Y., and Xia, L., 2013. Efficient and Accurate Face Detection Using Heterogeneous Feature Descriptors and Feature Selection. Computer Vision and Image Understanding, 117 (1), pp.12–28.

[18] Martinez-Gonzalez, A.N. and Ayala-Ramirez, V., 2011. Real Time Face Detection Using Neural Networks. 2011 10th Mexican International Conference on Artificial Intelligence, pp.144–149.

[19] Wu, J., Cui, Z., Sheng, V.S., Zhao, P., Su, D., and Gong, S., 2013. A Comparative Study of SIFT and its Variants. Measurement Science Review, 13 (3), pp.122–131.

[20] Jung, J., Lee, H., Member, S., Lee, J.H., and Park, D., 2010. A Novel Template Matching Scheme for Fast Full-Search Boosted by An Integral Image. Signal Processing Letters, IEEE, 17 (1), pp.107–110.

[21] Dong, J., Sun, Y., and Zhao, Y., 2008. Design Pattern Detection by Template Matching. Proceedings of the 2008 ACM symposium on Applied computing, pp.765–769.

[22] Dahal, B., Alsadoon, A., Prasad, P.W.C., and Elchouemi, A., 2016. Incorporating Skin Color for Improved Face Detection and Tracking

System. Image Analysis and Interpretation (SSIAI), 2016 IEEE Southwest Symposium on, pp.173–176.

[23] Singh, A., 2016. Face detection and Eyes extraction using Sobel Edge Detection and Morphological Operations. Anamika Singh Manminder Singh Birmohan Singh, pp.295–300.

[24] Kanmani.N and Babu.T, M., 2016. Minimal Support Based Multi-View Face Ddetection. Emerging Trends in Engineering, Technology and Science (ICETETS), pp.1–3.

[25] Lin, S. and Su, F., 2016. FCFD : Teach The Machine To Accomplish Face Detection Step By Step. Image Processing (ICIP), 2016 IEEE International Conference on, pp.1–5.

[26] Misra, O., 2016. An Approach to Face Detection and Alignment Using Hough Transformation with Convolution. 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall), pp.1–5.

[27] Sarkar, S., Patel, V.M., and Chellappa, R., 2016. Deep Feature-based Face Detection on Mobile Devices. IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), pp.1–8.

[28] Comaschi, F., Stuijk, S., Basten, T., and Corporaal, H., 2016. Robust Online Face Tracking-By-Detection. Multimedia and Expo (ICME), 2016 IEEE International Conference on, pp.1–6.

[29] Gregory P . Meyer , Steven Alfano, M.N.. Do, 2016. Improving Face Detection With Depth. Icassp 2016, pp.1288–1292.

[30] Lebedev, A., Pavlov, V., Khryashchev, V., and Stepanova, O., 2016. Face detection algorithm based on a cascade of ensembles of decision trees. 2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), 1, pp.161–166.

[31] Liuliu, W. and Mingyang, L., 2016. Multi-pose Face Detection Research Based on Adaboost. Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp.409–412.

[32] Mohanty, R. and Raghunadh, M. V, 2016. A New Approach to Face Detection based on YCgCr Color Model and Improved AdaBoost Algorithm. Communication and Signal Processing (ICCSP), 2016 International Conference on, pp.1392–1396.

[33] Tathe, S. V, 2016. Human Face Detection and Recognition in Videos. Advances in Computing, Communications and Informatics (ICACCI), pp.2200–2205.

[34] Zhang, X. and Haar, A., 2016. An Improved Adaboost Face Detection Algorithm Based on the Different Sample Weights. Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on, pp.4–7.

[35] Gupta, M.V., Sharma, D., and Sharma, M.D., 2014. A Study of Various Face Detection Methods. Methods, 3 (5), pp.3–6.

[36] Viola, P. and Jones, M., 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition., 1, pp.511–518.

[37] Li, B. and Yang, A., 2010. Rotated Face Detection Using AdaBoost. 2010 2nd International Conference on Information Engineering and Computer Science (ICIECS), pp.13–16.

[38] Viola, P., Way, O.M., and Jones, M.J., 2004. Robust Real-Time Face Detection. 2013 20th IEEE International Conference on Image Processing (ICIP), 57 (2), pp.137–154.

[39] Yan, J., Zhang, X., Lei, Z., and Li, S.Z., 2013. Face detection by Structural Models. IMAVIS, pp.1–10.

[40] Ban, Y., Kim, S.-K., Kim, S., Toh, K.-A., and Lee, S., 2014. Face Detection Based on Skin Color Likelihood. Pattern Recognition, 47 (4), pp.1573–1585.

[41] Mar, D., Hrka, T., and Ribari, S., 2016. Two-stage Cascade Model for Unconstrained Face Detection. Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on, pp.1–4.

[42] Tajima, Y., Ito, K., Aoki, T., Hosoi, T., Nagashima, S., and Kobayashi, K., 2013. Performance improvement of face recognition algorithms using occluded-region detection. 2013 International Conference on Biometrics (ICB), pp.1–8.

[43] Zakaria, Z. and Suandi, S. a., 2011. Face Detection Using Combination of Neural Network and Adaboost. TENCON 2011 - 2011 IEEE Region 10 Conference, pp.335–338.

[44] Lee, Y., Han, D.K., and Ko, H., 2013. Reinforced AdaBoost Learning for Object Detection with Local Pattern Representations. The Scientific World Journal, 2013.

[45] Xiang, Y., Wu, Y., and Peng, J., 2013. An Improved AdaBoost Face Detection Algorithm Based on the Weighting Parameters of Weak Classifier. 2013 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp.347–350.

[46] Dhingra, T., 2016. Face Detection Through LSE Devoid of Re-initiation via RD. Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp.227–232.

[47] Marčetić, D. and Ribarić, S., 2016. Deformable Part-based Robust Face Detection under Occlusion by Using Face Decomposition into Face Components. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on, pp.1365–1370.

[48] Do, T.T. and Doan, K.N., 2009. Boosted of Haar-like Features and Local Binary Pattern Based Face Detection. RIVF ’09. International Conference on Computing and Communication Technologies, (1992), pp.1–8.

[49] Jain, A.K., Duin, R.P.W., Mao, J., and Member, S., 2000. Statistical Pattern Recognition : A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (1), pp.4–37.

[50] Soon, T.K., Basari, A.S. H., and Hussin, B., 2015. Image Duplication and Rotation Algorithms for Storage Utilization: A Review. MEJSR Middle-East Journal of Scientific Research, 23 (10), pp.2454–2458.

[51] Sridhar, S., 2011. Digital Image Processing. New Delhi: Oxford University Press.

A Correlation based Approach to Differentiate between an Event and Noise in Internet of Things

Dina ElMenshawy[1], Waleed Helmy[2]
Information Systems Department, Faculty of Computers and Information
Cairo University
Egypt

***Abstract*—Internet of Things (IoT) is considered a huge enhancement in the field of information technology. IoT is the integration of physical devices which are embedded with electronics, software, sensors, and connectivity that allow them to interact and exchange data. IoT is still in its beginning so it faces a lot of obstacles ranging from data management to security concerns. Regarding data management, sensors generate huge amounts of data that need to be handled efficiently to have successful employment of IoT applications. Detection of data anomalies is a great challenge that faces the IoT environment because, the notion of anomaly in IoT is domain dependent. Also, the IoT environment is susceptible to a high noise rate. Actually, there are two main sources of anomalies, namely: an event and noise. An event refers to a certain incident which occurred at a specific time, whereas noise denotes an error. Both event and noise are considered anomalies as they deviate from the remaining data points, but actually they have two different interpretations. To the best of our knowledge, no research exists addressing the question of how to differentiate between an event and noise in IoT. As a result, in this paper, an algorithm is proposed to differentiate between an event and noise in the IoT environment. At first, anomalies are detected using exponential moving average technique, then the proposed algorithm is applied to differentiate between an event and noise. The algorithm uses the sensors' values and correlation existence between sensors to detect whether the anomaly is an event or noise. Moreover, the algorithm was applied on a real traffic dataset of size 5000 records to evaluate its effectiveness and the experiments showed promising results.**

***Keywords*—Anomaly detection; event; IoT; noise**

I. Introduction

Internet of Things (IoT) is the consolidation of physical objects that are coupled with electronics, software, sensors, and network connectivity, which permit them to capture and transfer data [1]. In IoT, a thing denotes a physical object that contains sensors to interact with the real world through a network to attain specific functions. Things can comprise smart phones, tablets, washing machines, refrigerators, etc. IoT is a network system which connects different communication devices with the internet to establish rapid, reliable, and real time information interchange that assists in intelligent management [2]. The objects capture data about the surrounding environment to monitor certain phenomena such as temperature and humidity. Consequently, objects can be tracked remotely allowing for the communication between the physical and virtual worlds.

The popularity of the IoT notion relies mainly on current technologies: internet, mobile technologies, cloud computing, communication protocols, and embedded sensors to capture the data [3]. In IoT, data is generated by things, so real world objects are considered the core components of the IoT paradigm. Every object has a distinctive identity and can access the network to integrate between both the physical and digital worlds to provide enhanced services to people. IoT can provide device to device, device to people and device to environment information transfer through the integration of information space and physical space [4], [5].

IoT architecture is composed of three levels as shown in Fig. 1.

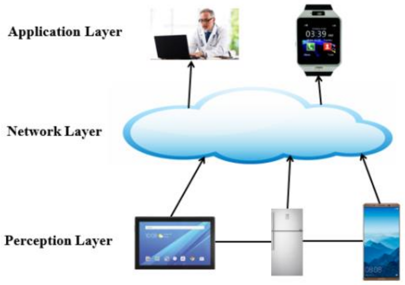


Fig. 1. IoT Architecture (Adapted From [6]).

The topmost layer is the application layer which represents the application service support system. The intermediate layer is the network layer which contains the communication network infrastructure. The bottom layer is the perception layer which comprises the sensor based devices and environmental objects. The captured data from this layer is transferred to the network layer for further processing and analysis [6], [7].

IoT applications generate enormous amounts of data which are characterized by the 5V model

1) *Volume*: huge quantities of generated data.
2) *Variety*: different data types such as structured, semi-structured, and unstructured data.
3) *Velocity*: immense speed of data production and processing.
4) *Veracity*: accuracy and trustiness of the generated data.
5) *Value*: benefits yield from using the data [8].

IoT has numerous applications in different fields such as healthcare, business, smart homes, etc. IoT applications became extensively used in people’s daily lives to make their lives more comfortable [3]. IoT applications are categorized into three main areas:

1) Personal: such as smart homes, telemedicine, and wearables.

2) Social: such as smart grid, smart lighting, and waste management.

3) Business: such as smart farming and smart retail [1], [2].

IoT faces a lot of challenges varying from data management to security threats. Regarding data management, sensors generate enormous amounts of data with various formats so data fusion techniques are required to combine the data. In addition, the IoT environment is vulnerable to a high noise rate since it mainly relies on sensors which possibly be of low power and poor quality [8]. IoT is still in its infancy so it faces a lot of difficulties to have successful employment of different applications. One great challenge is the detection of data anomalies emerging from sensors’ data.

II. Motivation

Sensors generate enormous amounts of data that need to be handled efficiently. IoT applications mainly depend on data generated from these sensors, as a result, anomalies can substantially minimize the effectiveness of IoT applications and consequently may lead to inaccurate decisions. Anomaly detection is beneficial because anomalies are doubtful of not being generated by the same methods as the other data points.

The discovery of data anomalies in IoT is a sophisticated task because it is difficult to determine the normal pattern of data as data in the IoT environment is domain dependent [8]. Moreover, multiple sensors continuously generate data to monitor a certain phenomenon so the generated data have various formats.

Actually there are two main causes of data anomalies, namely: an event and noise. An event refers to a specific incident which took place at a certain time interval, whereas noise is just an error, usually because of: poor quality sensors, environmental effects or communication problems. Both event and noise are considered anomalies in terms of having a great deviation from normal data points, but actually they have two different interpretations.

Event detection in IoT is essential since late discovery of certain events such as a fire can cause huge problems. On the other hand, a noise is just considered an error resulting from sensors.

To the best of our knowledge, no work exists answering the question of how to differentiate between an event and noise in IoT. As result, an approach is needed to differentiate between an event and noise since both are considered abnormal points, i.e anomalies so, in this paper, an algorithm is proposed to differentiate between an event and noise. The main contributions are

1) Proposing a novel algorithm for differentiating between an event and noise based on both sensors’ values and correlation existence between sensors in the IoT environment.

2) Utilizing the factor of correlation existence between the sensors.

3) Applying the proposed algorithm on a real dataset to evaluate its effectiveness.

The rest of the paper is organized as follows: Section 3 presents the literature work of anomaly detection in IoT. Section 4 presents the categories of anomalies. Section 5 presents the proposed algorithm and experiments. Section 6 presents the conclusion and future work.

III. Related Work

Since the IoT paradigm is still in its beginning, little work investigated the detection of anomalies in this new environment. In [9], the paper presented an approach for detecting data anomalies through utilizing expert knowledge. The proposed approach made use of the possible expected attacks for discovering anomalies through a set of predefined constraints on the data. In [10], a real world simulation prototype was proposed that used IoT smart objects to detect behavioral based anomalies across a simulated smart home. The proposed technique used immunity inspired algorithms to discriminate between normal and abnormal behavioral patterns.

In [11], an unsupervised anomaly detection approach using light switches was presented. The proposed algorithm used a statistical based algorithm using expectation maximization to construct the mixture models. In the proposed technique, an anomaly was correlated with a probability. In [12], a correlation based anomaly algorithm was presented as a predictive maintenance method for compact electric generators. Correlations between sensors were determined by using statistical analysis. Anomalies were detected through analyzing sensors’ data and correlation coefficients between sensors.

In [13], a new notion of urban heartbeat which was constructed from sensors’ data in the surrounding environment was proposed. Urban Heartbeat collected the contextual information about patterns which occur regularly in the environment. Techniques were developed to find couplings between sensors. Next, quasi periodic patterns were determined from the data. After that, unexpected events which deviate significantly from the normal behavior were discovered.

In [14], air pollution elements were used to discover the unhealthy or anomalous locations in a smart environment. Anomalies were discovered through examining the air quality index, which is a numerical measure used to find out the anomalous locations which goes beyond a specific threshold. Neural networks, neuro fuzzy method, and support vector machines for binary and multi class problems were applied to identify anomalous locations from a pollution database.

IV. Categories of Anomalies

An anomaly/outlier is a data point that greatly differs from the remaining data points, as though it was produced by another approach [15]. There are three main types of anomalies, described as follows:

1) Global/Point anomaly: in a certain dataset, a data point is a global anomaly if it differs substantially from the remaining data points [15]. Global anomalies are considered the easiest type of anomalies to discover and most anomaly detection techniques focus on detecting them.

2) Contextual anomaly: in a particular dataset, a data point is considered a contextual anomaly if it noticeably differs in the defined context [16]. Contextual anomalies are also named as conditional anomalies because they rely on a specific context. As a result, to discover contextual anomalies, the context has to be determined as a core component of the problem definition. In contextual anomaly detection, the attributes of the data points in consideration are categorized into two types:

- Contextual attributes: these features determine the object's context. Context can refer to a time interval or location.
- Behavioral attributes: these attributes refer to the object's characteristics, and are used to determine whether a data point is an anomaly in the context which it exists [15].

3) Collective anomaly: in a certain data set, a subset of data points creates a collective anomaly if the points as a whole vary greatly from the whole dataset. The individual data points may not be anomalies [16].

In this paper, we will focus on detecting global anomalies.

V. Proposed Approach

In this section, the proposed algorithm along with the experiments will be presented. The process of differentiating between an event and noise consists of two main phases:

- The *first* phase detects the anomalies.
- The *second* phase decides whether each anomaly is an event or noise based on the conditions specified in the proposed algorithm.

The process of differentiating between an event and noise is depicted in Fig. 2

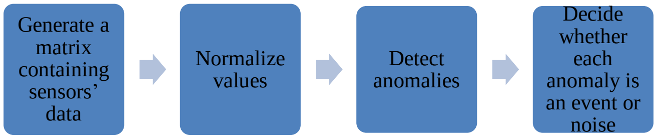

Generate a matrix containing sensors' data
Normalize values
Detect anomalies
Decide whether each anomaly is an event or noise

Fig. 2. Process of Detecting Anomaly's Type.

At first, a matrix is generated to include the sensors' data. Then, the data values are normalized. After that, anomalies are detected. At last, the anomaly is either defined as an event or noise. The exact steps of the algorithm will be illustrated in the following subsections.

A. Anomaly Detection

Sensor' data are usually time series data so techniques that fit time series data should be used to detect anomalies. As a result, in this paper, the technique used for anomaly detection is Exponential Moving Average (EMA), also known as an exponentially weighted moving average (EWMA). Exponential moving average is a technique for smoothing time series data using the exponential window function [17].

In the simple moving average, the previous observations are weighted equally, whereas in exponential moving average, exponential functions are used to assign exponentially decreasing weights over time and the weighting for each older data point decreases exponentially [18], that's why EMA is commonly used in analysis of time series data. The advantage of EMA is that it keeps little record of previous data since it focuses on most recent observations, as the most recent data should be given more weight.

In our proposed approach, EMA analyzed whether the value of the attribute being investigated in a given timestamp exceeds a certain threshold. EMA was chosen because it gives more weight to recent observations rather than older ones, so this will help in determining the trend of data and differentiating between an event and noise.

Luminol [19] which is a light weight python library for time series data analysis, was utilized in the experiments. It supports anomaly detection using EMA. The anomaly score was calculated, then the score was compared to a certain threshold to decide whether the data point is an anomaly or not.

Usually, it is recommended to set the threshold based on the statistical principle which states that: to consider a value as an anomaly, it either exceeds $\mu+3\sigma$ or goes below $\mu-3\sigma$ where μ is the mean value and σ is the standard deviation of the attribute under observation [20], so in our experiments, we used this principle to determine the threshold value.

B. Differentiation between an Event and Noise

The following paragraphs will present the algorithm and experiments in details.

1) Data preprocessing:- At first, data need to be preprocessed so min-max normalization was applied on the dataset. Normalization was done through sklearn.preprocessing.MinMaxScaler [21], [22]. Scikit-learn (sklearn) [23] is a free software machine learning library for the Python programming language, and MinMaxScaler is a preprocessing module which scales each value such that it is in the range between zero and one.

2) Proposed algorithm:- Data from sensors can be represented by a data matrix produced by every sensor at each timestamp, denoted as s_{ti}, where s_{ti} refers to the measured value of attribute i at a timestamp t, described as follows in (1)

$$S_{ti} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{21} & S_{22} & \dots & S_{2n} \\ S_{31} & S_{32} & \dots & S_{3n} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \qquad (1)$$

Where t denotes the timestamp, whereas i refers to the sensor number and n denotes the number of sensors.

At any given timestamp, a sensor can be correlated with any other sensor in the surrounding environment. The sensors' values can be either positively correlated or negatively correlated. To determine whether sensors are correlated or not, a correlation matrix of zeros and ones was constructed to define the correlation between sensors. The correlation value was either zero or one, zero refers to absence of correlation, whereas one refers to presence of correlation (either positively or negatively) between the sensors.

To know if two attributes are correlated or not, we should check the correlation matrix. For example, the average speed and flow of cars in a certain road are negatively correlated because, when the flow (number) of cars increases at a certain timestamp, the average speed of cars decreases. The correlation matrix between these two attributes will be as follows in (2)

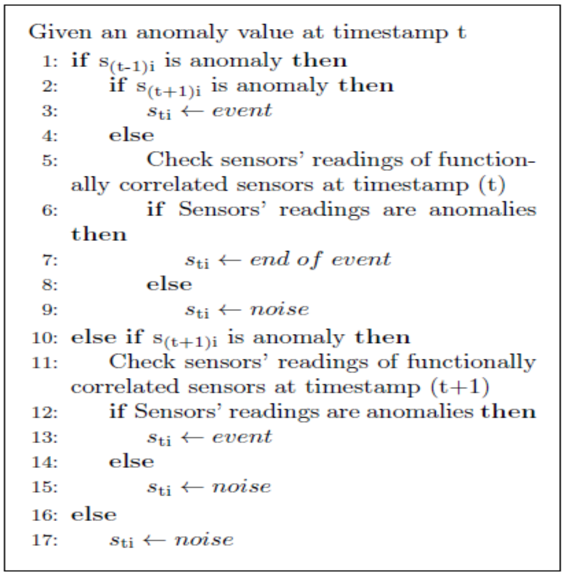
$$\begin{array}{cc} & \begin{array}{cc}\text{Flow} & \text{Speed}\end{array} \\ \begin{array}{c}\text{Flow} \\ \text{Speed}\end{array} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array} \tag{2}$$

To use the correlation notion efficiently, only the functionally correlated sensors should be examined since they usually measure the same phenomena. As a result, the values of the functionally correlated sensors can be used in differentiating between an event and noise.

At first, anomalies should be detected using EMA, then the proposed algorithm is applied to differentiate between an event and noise. It detects whether this anomaly is an event or noise. The proposed algorithm is depicted in Fig. 3.

To better illustrate the algorithm, there are four different cases

- s_{ti}, $s_{(t-1)i}$ and $s_{(t+1)i}$ are anomalies.
- s_{ti} is an anomaly whereas $s_{(t-1)i}$ and $s_{(t+1)i}$ are not anomalies.
- s_{ti} and $s_{(t-1)i}$ are anomalies whereas $s_{(t+1)i}$ is not anomaly.
- s_{ti} and $s_{(t+1)i}$ are anomalies whereas $s_{(t-1)i}$ is not anomaly.



```
Given an anomaly value at timestamp t
if $s_{(t-1)i}$ is anomaly then
    if $s_{(t+1)i}$ is anomaly then
        $s_{ti} \leftarrow event$
    else
        Check sensors' readings of functionally correlated sensors at timestamp (t)
        if Sensors' readings are anomalies then
            $s_{ti} \leftarrow end\ of\ event$
        else
            $s_{ti} \leftarrow noise$
else if $s_{(t+1)i}$ is anomaly then
    Check sensors' readings of functionally correlated sensors at timestamp (t+1)
    if Sensors' readings are anomalies then
        $s_{ti} \leftarrow event$
    else
        $s_{ti} \leftarrow noise$
else
    $s_{ti} \leftarrow noise$
```

Fig. 3. Proposed Algorithm.

The two main contributions of the proposed algorithm are

1) Utilizing the following timestamp: Most existing anomaly detection algorithms use previous timestamps to discover anomalies, whereas the proposed algorithm used the following timestamp besides the previous timestamp to differentiate between an event and noise. The idea behind using the following timestamp is to wait for more time so that more accurate decisions can be taken since events usually last for a time interval.

2) Using correlation existence between sensors:- The whole dataset was scanned at once to detect anomalies using EMA then, the proposed algorithm was applied to determine whether each anomaly point is an event or noise depending on the specified conditions in the algorithm.

3) Dataset used:- In order to evaluate the performance of the proposed algorithm, it was applied on a real traffic dataset and the accuracy of detecting events and noise was measured. The dataset consisted of 5000 records with 3 attributes and the proportion of anomalies in the dataset was 5%. The dataset presented real time traffic data from the Twin Cities Metro area in Minnesota, collected by the Minnesota Department of Transportation. The Minnesota Department of Transportation captured traffic data on the freeway system throughout the Twin Cities Metro area [24].

The dataset contains occupancy, speed, and flow data for every detector in the Twin Cities Metro area and was collected every 30 seconds. Speed refers to the average value of the speed measurements of individual vehicles over time, whereas flow denotes the number of vehicles passing in a specific point at a certain timestamp. Flow and speed were used, whereas occupancy was not included in the experiments, since occupancy is similar to flow as it represents the percent of time the detection zone of a sensor is occupied by vehicles. These two attributes were selected because they are correlated, i.e., when the flow increases, the speed decreases and vice versa as they exhibit negative correlation.

4) Performance evaluation:- The dataset was used to evaluate the performance of the proposed algorithm. The prediction accuracy of detecting both events and noise was computed. The accuracy was measured as in (3) and (4):

Prediction accuracy of detecting events =

$$\frac{\text{number of correct predictions}}{\text{number of events}} \tag{3}$$

Prediction accuracy of detecting noise =

$$\frac{\text{number of correpredictions}}{\text{number of noise values}} \tag{4}$$

Given that most of the available real data have no class labels, so anomaly labels (both an event and noise) were artificially added to the dataset in order to measure the prediction accuracy of the proposed algorithm.

The prediction accuracy of detecting events and noise is shown in Fig. 4.

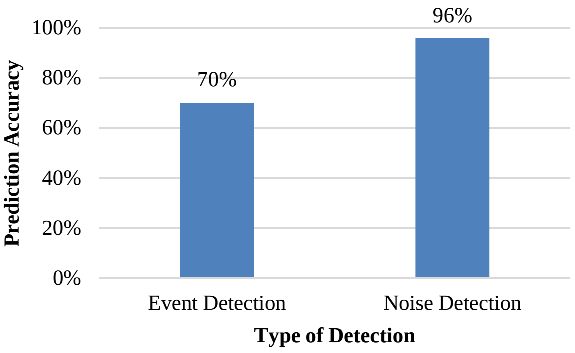


Fig. 4. Prediction Accuracy of Event and Noise Detection.

The proposed algorithm gave promising results especially in detecting noise. The accuracy is higher in detecting noise rather than events, maybe because it is more difficult to detect events, since events' detection involves several factors such as the sensors' values of both the preceding and following timestamps, the values of the functionally correlated attributes, and the nature of event. On the other hand, the noise is just an error resulting from the sensors.

VI. Conclusion and Future Work

IoT is a new paradigm that recently gained popularity. IoT is the integration of physical objects that are attached with software, sensors, and network connectivity, which allow them to capture and transmit data. The IoT paradigm faces numerous challenges ranging from data management to security threats. A substantial challenge is the detection of data anomalies from sensors' data. An anomaly is a data point that greatly varies from the rest of data points. There are two main causes of data anomalies which are: an event and noise. An event denotes an incident which happened at a certain time, whereas noise is just an error. An approach is needed to distinguish between an event and noise since both are considered anomalies so, in this paper, an algorithm was proposed to differentiate between an event and noise in IoT. Also, the effectiveness of the algorithm was tested through experiments.

In future work, we will explore how to enhance the accuracy of the algorithm in detecting events. Also, the algorithm will be applied on other datasets in different domains.

References

[1] S. Ray, Y. Jin, and A. Raychowdhury, "The changing computing paradigm with internet of things: a tutorial introduction," IEEE Design and Test, vol. 33, no. 2, pp. 76–96, April 2016.

[2] S. Elbouanani, M. A. E. Kiram, and O. Achbarou, "Introduction to the internet of things security: standardization and research challenges," in *Proc. IAS,* Marrakech, 2015, pp. 32–37.

[3] S. Kraijak and P. Tuwanut, "A survey on internet of things architecture, protocols, possible applications, security, privacy, real-world implementation and future trends," in *Proc. ICCT,* Hangzhou, 2015, pp. 26–31.

[4] Z. Yue, W. Sun, P. Li, M. U. Rehman, and X. Yang, "Internet of things: architecture, technology and key problems in implementation," in *Proc. CISP,* Shenyang, 2015, pp. 1298–1302.

[5] S. Nalbandian, "A survey on internet of things: applications and challenges," in *Proc. ICTCK,* Mashhad, 2015, pp. 165–169.

[6] D. Rose, Enchanted Objects: Design, Human Desire, and the Internet of Things, 1st ed., New York: Simon & Schuster, 2014.

[7] W. Z. Khan, H. M. Zangoti, M. Y. Aalsalem, M. Zahid, and Q. Arshad, "Mobile RFID in internet of things: security attacks, privacy risks, and countermeasures," in *Proc. ICRAMET,* Jakarta, 2016, pp. 36–41.

[8] I. Butun, B. Kantarci, and M. Erol-Kantarci, "Anomaly detection and privacy preservation in cloud-centric internet of things," in *Proc. ICCW,* London, 2015, pp. 2610–2615.

[9] V. A. Desnitsky, I. V. Kotenko, and S. B. Nogin, "Detection of anomalies in data for monitoring of security components in the internet of things," in *Proc. SCM,* St. Petersburg, 2015, pp. 189–192.

[10] B. Arrington, L. Barnett, R. Rufus, and A. Esterline, "Behavioral modeling intrusion detection system (BMIDS) using internet of things (IoT) behavior-based anomaly detection via immunity-inspired algorithms," *in Proc. ICCCN,* Hawaii, 2016, pp. 1–6.

[11] C.-W. Ho, C.-T. Chou, Y.-C. Chien, and C.-F. Lee, "Unsupervised anomaly detection using light switches for smart nursing homes," in *Proc. DASC,* Auckland, 2016, pp. 803–810.

[12] P. Zhao, M. Kurihara, J. Tanaka, T. Noda, S. Chikuma, and T. Suzuki, "Advanced correlation-based anomaly detection method for predictive maintenance," in *Proc. ICPHM,* Seattle, 2017, pp. 78–83.

[13] S. A. Hasnain and R. Jafari, "Urban heartbeat: From modelling to applications," in *Proc. SMARTCOMP,* Hong Kong, 2017, pp. 1–8.

[14] R. Jain and H. Shah, "An anomaly detection in smart cities modeled as wireless sensor network," in *Proc. IConSIP,* Nanded, 2016, pp. 1–5.

[15] J. Han, M. Kamber, and J. Pei, "Outlier detection," in Data Mining: Concepts and Techniques, 3rd ed., Netherlands: Elsevier, 2012, pp. 544–548.

[16] D. Hand, H. Mannila, and P. Smyth, "Measurement and data," in Principles of Data Mining, Cambridge: The MIT Press, 1st ed., 2001, pp. 35–36.

[17] D. R. Anderson, D. J. Sweeney, and T. A. Williams, "Descriptive statistics: numerical measures," in Essentials of Modern Business Statistics, 3rd ed., Boston: Cengage Learning, 2012, pp.147–148.

[18] "Exponential smoothing," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Exponential_smoothing. [Accessed: 1-Novmeber-2018].

[19] Linkedin, "linkedin/luminol," *GitHub*. [Online]. Available: https://github.com/linkedin/luminol. [Accessed: 1-November-2018].

[20] H.-P. Kriegel, P. Kröger, and A. Zimek, "Outlier detection techniques," in *Proc. SDM,* Columbus, 2010, pp. 1–73.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in python," *JMLR,* vol. 12, pp. 2825–2830, October 2011.

[22] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *Proc. ECML PKDD,* Prague, 2013, pp. 108–122.

[23] "Scikit-learn," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Scikit-learn. [Accessed: 10-November-2018].

[24] "Mn/DOT Traveler Information," *Minnesota Department of Transportation*. [Online]. Available: http://www.dot.state.mn.us/tmc/trafficinfo/developers.html. [Accessed: 11-November-2018].

An Improvement of Performance Handover in Worldwide Interoperability for Microwave Access using Software Defined Network

Fatima Laassiri[1, a], Mohamed Moughit[2, b], Noureddine Idboufker[3, c]
[a, b]IR2M Laboratory, FST, Univ Hassan UH1- Settat, Morocco
[b]EEA&TI Laboratory, FST, Univ Hassan , Mohammedia, Morocco
[b]National Schools of Applied Sciences Khouribga, Univ Hassan 1, UH1- Settat, Morocco
[c]National School of Applied Sciences, Univ Cadi Ayyad Marrakech, Morocco

***Abstract*—The WiMAX network designates in common language of a set of standards and techniques of the world of Wireless Metropolitan Area Networks (WMAN). The standard IEEE 802.16 or WiMAX allows the wireless connection of companies or individuals over long distances at high speed. WiMAX provides an appropriate response for some rural or hard-to-reach areas, which today lack access to Broadband Internet for cost reasons. This technology aims to introduce a complementary solution to the Digital Subscriber Line (DSL) and cable networks on the one hand, and to interconnect WiFi hotspots, on the other hand WiMAX is mainly based on a star topology although mesh topology is possible. Communication can be done in Line of Sight (LOS) or not (NLOS).Software Defined Network (SDN) is a new network paradigm used to simplify network management. It reduces the complexity of network technology.The following article aims to expose a simulation implemented under Omnet4.6++, to improve Handover performance and QoS (End-to-end delay, latency, jitter, MoS and lost packet), by implemented an algorithm in SDN controller. The simulation is tested in WIMAX architecture, and results have been collected from two scenarios with and without SDN controller to proof that this algorithm is more preferment to guarantee a better QOS in Handover.**

***Keywords*—WiMAX; SDN; QoS; handover; openflow; OMNeT 4.6 ++**

I. Introduction

WiMAX (Worldwide Interoperability for Microwave Access) [1] was standardized in 2002 under the name 802.16. Its goal was to cover high-speed white areas [2]

WiMAX technology is synonymous with better transfer rates, lower latency, better availability and efficiency, but not compatible with existing 2G and 3G systems, a major pitfall. In addition, switching from a 3G network to a 4G network is often more difficult, and sometimes problematic.

It is distinguished by the vast extent of its Internet coverage, because it is a very broad band. Besides, there are several types of high-speed Internet coverage (eg 3G, 4G, LTE, satellite and WiMAX). It offers another aspect of how conventional remote systems have been described, due to the effects of innovation on the public and its impact on the environment.

The mechanical points of interest of WiMAX are central points of unprecedented WiMAX radio change. It works in the field of security and quality. It has an open access base focused on IP access. This innovation can be used for various applications. It is a scalable remote correspondence system capable of providing high-speed remote access with high data rate of Fourth-Generation (4G) over a long separation in a point-to-multipoint and visible or an unobservable path condition.

This article presents an improvement the HO[3] and the QoS, " End-to-end delay, latency, jitter, the number of lost packets, and the MOS ", by the implementation a new algorithm that it allows the Software Defined Network (SDN)[4] controller to better improve the performance of all WiMAX network architectures using OMNeT4.6 ++.

II. State of the Art

WiMAX Wireless Microwave Metro Networks that are useful in broadband access, as a central innovation in the IEEE 802.16 reference group, advancing in 4G. With the current presentation of portability management systems in the IEEE 802.16e standard, it is currently competing with current and future ages of remote advances to provide ubiquitous recording arrangements. Nevertheless, the establishment of a decent versatile structure depends to a large extent on the ability to make quick and consistent transfers regardless of the situation of the building being sent. Since the IEEE has characterized the mobile WiMAX MAC layer transfer (IEEE 802.16e) administration structure, the WiMAX Forum Network Working Group (NWG) [5] is working on improving upper layers, or the path to commercialization of an undeniable WiMAX versatility structure, which it is tasked to investigate the difficulties. This focuses on potential research issues related to the transfer into the current and future WiMAX portability structure. An examination of these issues in the MAC, Network and Cross Layer situations are presented alongside the exchange of distinctive answers to these difficulties.

According to research done by the "Mojtaba Seyedzadegan" on the overview of WiMAX, its architecture deals with the supply of data. It is entirely based on IEEE

802.16. It supports two things. One is the alternative broadband cable and the other is DSL.

Subscriber Station (SS)[6] gives the link to WiMAX. Many websites are used to provide the tool used for this method. But this is not a complete view of the tools available as certified that they are set in mobile internet devices and various private labelled tools, laptops.

It assembles Wi-Fi[7] activities because of the same wireless network. It also gives network connectivity to businesses and at home without the help of external devices. For this, it uses WiMAX carriers. It is scaled a few kilometers. It varies at the scale of a city. It is given with a subscriber unit. It helps the customer to connect to the Internet and other accesses. Physical Layer It's worth mentioning that both LTE and WiMAX use Orthogonal Frequency Division Multiple Access (OFDMA)[8] in the downlink, but they differ in the uplink. WiMAX continues to use OFDMA, while LTE's[9] approach is more advanced. Using OFDMA is power inefficient, but it's tolerable in the downlink because the power amplifier is placed at the base station (or at the e-Node-B in 3GPP terminology). At the base station, power is available, and the many mobile terminals share the extra complexity. However, in the uplink, the transmissions start from mobile devices, which are battery powered. The mobile devices are also constrained because they must be low cost to enable mass deployment. 3GPP[10] specifications thus propose a reduced Peak to Average Power Ratio (PAPR) transmission scheme for the uplink signal. This scheme is called Single Carrier Frequency Division Multiple Access (SCFDMA). This makes it easier for the mobile terminal to maintain a highly efficient signal transmission using its power amplifier. The LTE uplink signal achieves this property and saves power without degrading system flexibility or performance. [11]

III. Standard of 802.16 WiMAX

The next table (Table 1) presents the standard of 802.16 WiMAX [12]

TABLE I. CURRENT AND FEATURE STANDARD OF 802.16 WIMAX

Standard	Description
802.16.2-2004	Recommended practice
802.16-2009	Air interface for fixed and mobile
802.16h-2010	Iproved Coexistence mechanism
p802.16n	Higher Reliability Networks (In progress)
p802.16k-2007	Bridging of 802.16
802.16j-2009	Multi hop relay
802.16m-2011	Advanced air interface with high data rate
p802.16p	Support machine to machine Apllication(In progress)

IV. Problem and Solution

In recent years, the SDN [13] plays an important role because of its flexibility and ease of transport. WiMAX is the promising 4G network to meet the needs of customers. It is a telecommunication technology, which provides software-defined data for several distances from a point-to-point link to all cell-type accesses, and it allows the connection between mobile and fixed networks. The coverage area of WiMAX is highly compared to other technologies, among different technologies; it offers good support and good stability.

WiMax refers to certain types of Wireless Local Area Networks (WLANs) [14], which use 802.16 specifications. It is easily used by companies.

Due to the growing demand, users have ubiquitous access to wireless services, which has led to the deployment of forced use of this wireless access technology such as WiMax. It offers a level of quality, within range, but the problem is that the number of devices is increasing which reduces Handover's travel time performance and QoS in terms of end-to-end delay, latency, jitter, number of lost packets, and MOS, for all these reasons, This work proposes as a solution to implement the SDN technology for WiMax to better optimize their performance through the creation of a new algorithm under SDN controller.

V. Advantages of WiMAX

WiMAX is typically used as an alternative to dedicated links and Internet access of all kinds for the following applications

- It is the best solution for suburban and urban areas, where there is a wired technology problem, for large cities, this technology can also be implemented to meet all high-speed requirements;
- possibility of reusing a frequency dedicated to a BTS to increase the capacity of the system, also to support hundreds of users;
- frequency allocation is done on a sectoral basis when the number of users increases;
- networks with high transmission speeds for voice and data;
- connect to the Internet peripheral neighborhoods or suburban cities;
- inter-site private networks for companies;
- security and surveillance that may include video over IP applications;
- regional wireless networks with data and voice applications for industry and transport;
- wireless communications integrating VoIP;
- temporary deployments: Religious events, construction sites, relief infrastructure on a natural disaster;
- that it is a wireless technology with a signal range of a few hundred meters and a maximum bit rate of about 11Mbits/s, WiMAX has a technology that manages bandwidth. Thus, a user that he performs an operation, that it requires a lot of resources (High-quality video conference for example) will have a large bandwidth;
- low WiMAX allows faster deployment without the need for heavy civil engineering work;

- it allows high-speed wireless internet connectivity over long distances;
- it can serve multiple clients at a time;
- one signal despite the obstacles;
- private inter-site networks for companies;
- perspective of nomadism;

VI. Disadvantages of WiMAX

- Debit is shared between users of the same central antenna;
- obligation to serve WiMAX base stations through a collection network (optical fiber, radio link, etc.);
- requirement to have a licence: Only licensees are able to deploy WiMAX networks;
- need to have a high point: To ensure the best possible coverage, the transmitter must be placed on a high point (pylon, water tower, etc.);
- it must first obtain a licence from a public authority;
- to have optimal distances and speeds, transmitters and receivers must be in “line of sight”;
- it will be able to cross only small obstacles like a tree or a house but, the signal is unable to pass through hills or large buildings;
- methods for the implementation of 4G with and without SDN

VII. Methods and Implementation of WIMAX with and without SDN using OMNET 4.6 ++

Software Defined Metropolitan An Network (SDMAN)[13], it is a standard network that is designed to provide broadband access for a large area. There is more flexibility and service. It can be used in both the licensed frequency bands (10-66GHZ) and the unlicensed band (11GHZ). Therefore, it is the best technology for the system designed as long as SDN whereas the good performance and the effective cost. So, that he faces different types of problems in mobile communication. A frequency higher than 10 GHz is required and if visibility is reduced, a frequency below 10 GHz is essential.

It allows a broadband access service, which helps the customer to take advantage of low-cost Internet options. In general, it is a software-defined technology that operates on a frequency between 2 and 66 GHz. In addition, it provides data rates up to 75 Mb/s. Thus, it becomes the backbone of the many software defined communication services. In order to increase the applications of WIMAX.

This work expresses an improvement of the QoS "end-to-end delay, latency, jitter, the number of lost packets, and the MOS " through the intecration of a new algorithm that it allows to chang SDN controller policy for handle handover in the WiMAX network, by two scenarios, the first (Figure 1) expresses WiMAX without SDN and the second (Figure 2) is based on the implemented of SDN program, using OMNeT4.6++.

Scenario 2 (Figure 1): WiMAX without SDN: Is implemented with 10 base stations and with a frequency of 2.4 GHZ.

Scenario 1 (Figure 2): WiMAX with SDN: Is implemented with 10 base stations with a frequency of 2.4 MHZ, an SDN controller, and an OpenFlow switch.

You will find attached the parameters of WiMAX used: Frequency: 2400 MHz.

BSAntennaHeigh

BasicEnergyLowBatteryThreshold: 0.1

Mobile Rx height: 1.5 meter

Effective radius: 4.34Km

Forest/Trees, Offset-loss: 10 dB

Open, Offset-loss: 17 dB

Buildings, Offset-loss: -4 dB

DL Carrier bandwidth: 9x3.3MHzusingFR 1/3

DL Carrier bandwidth: 3 x 10 MHz using FFR 1/1

Bearers for DL and UL: 1

CINR requirement: 4.9 dB

Services for DL and UL: 1

UL Carrier bandwidth: 1x3.3MHz

Suburban, Offset-loss: -3 dB

Noise Figure : 7 dB

Horizontal Beam width of CPE: 360

Noise Figure : 4 dB

Receiver Sensitivity: -109.7300 dB

Control Activity: 20 %

BS antenna: Kathrein80010

BS antenna gain: 18 dB

Horizontal beam width of BS antenna: 60°

Vertical beam width of BS antenna: 6°

Feeder loss: 2 dB

Prediction resolution: 10 m

Intra-Site fading correlation coefficient: 0.8

Inter-Site fading correlation coefficient: 0.5

WiMAX lies in it’s simplicity of implementation. It will take only two antennas to connect two remote networks, where it would have been miles of optical fiber wired.

The WiMAX client must have a receiver (a built-in chip or a CPE: Customer Premise Equipment) and be within the scope of a transmitter. The transmission between the customer and his hot spot WiMAX is said in "No Line Of Sight" (NLOS), that it is to say that the customer is not in direct view with the antenna. Indeed, buildings or vegetation found in cities "force" the signal to be diverted through the use of OFDM frequency modulation.

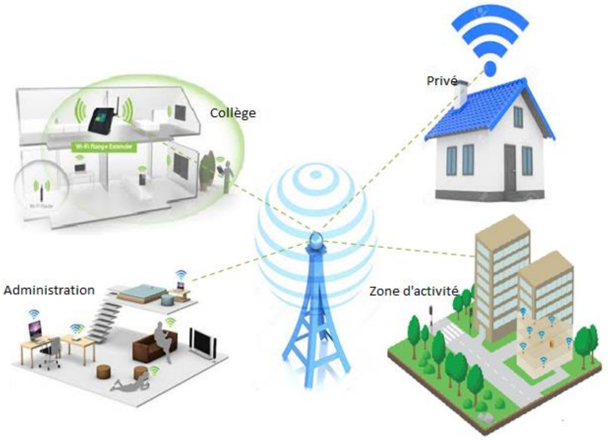


Fig. 1. WiMAX without SDN.

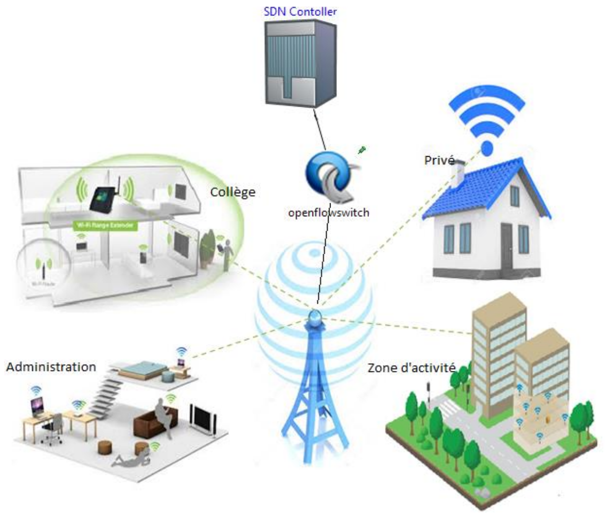


Fig. 2. WiMAX with SDN.

In a network, the collection consists of connecting the access points (WiFi or DSLAM hot spots) thus ensuring the connection with the Internet. This mechanism is called the backhauling of hotspots. Unlike the service, the collection is done in "Line Of Sight" (LOS), thanks to WiMAX transmitters placed high enough.

VIII. Results and Discussion of Simulation in QoS Criteria under WiMAX with and without SDN using OMNeT 4.6 ++

This section presents the QoS performance results for WiMAX using SDN, such as end to end delay, latency, jitter, lost packet, and MOS.

A. End to end Delay under WiMAX with and without SDN

Figure 3 shows that the end to end delay in the WiMAX scenario without SDN with a higher value (26 ms) compared to the SDN based scenario that, it has a reliable delay (0.6 ms), that it expresses the implemented of SDN to WiMAX, it has a positive impact.

B. Jitter under WiMAX with and without SDN

The jitter under figure 4, where the WiMAX scenario, that it is based on SDN is about 10 ms, that it is lower than WiMAX without SDN, that it has the value of 11 ms, which results that the connection of an SDN network for WiMAX is successful.

C. Latency under WiMAX with and without SDN

The results of figure 5 shows that the WiMAX network latency with SDN is less (14 ms) than that of the WiMAX approach without SDN with a value of 16 ms, which explains that the value added by the latter is beneficial

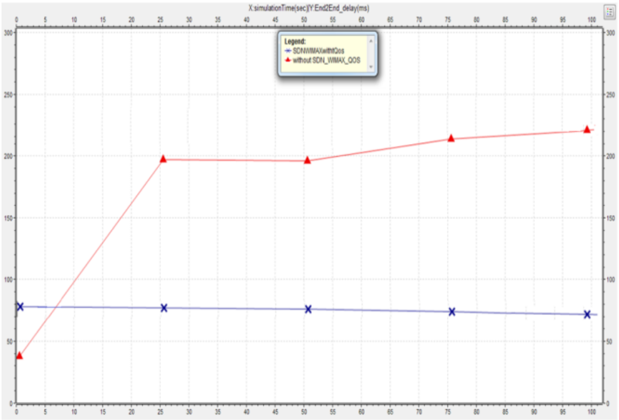


Fig. 3. End to end Delay under WiMAX with and without SDN.

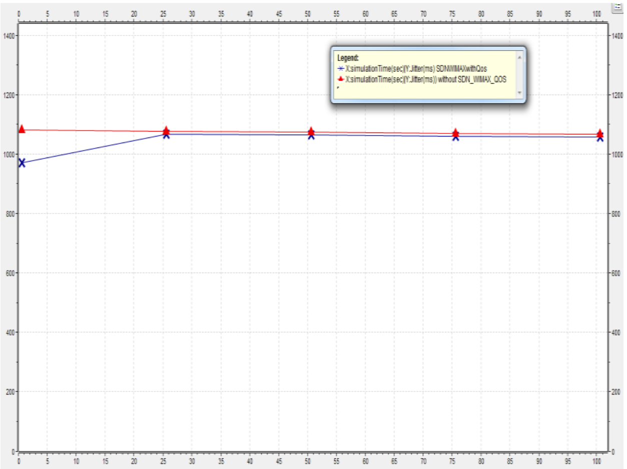


Fig. 4. Jitter under WiMAX with and without SDN.

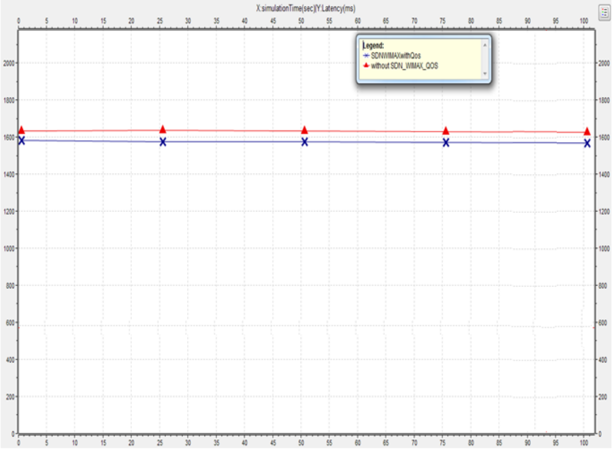


Fig. 5. Latency under WiMAX with and without SDN.

D. Packets Lost under WiMAX with and without SDN

The number of packets lost in the WiMAX approach without SDN is 7%, which is higher compared to the WiMAX approach with SDN that it is about 5%. This shows that the impact of adding SDN for the WiMAX network is totally favorable, as shown in figure 6.

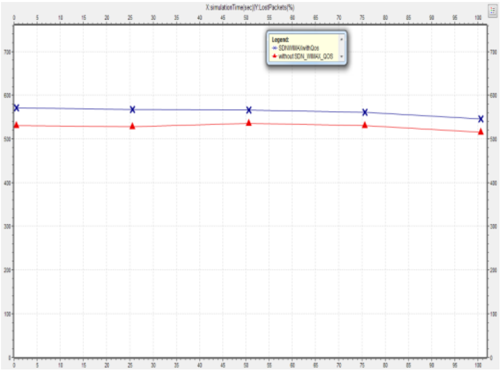


Fig. 6. Packets lost under WiMAX with and without SDN.

E. MOS under WiMAX with and without SDN

Figure 7 shows that the MOS offered by the WiMAX approach without SDN is 1, whereas the WiMAX based approach with SDN is about 2.3, which presents an indicator of the increases in the quality of the WiMAX voice transmission.

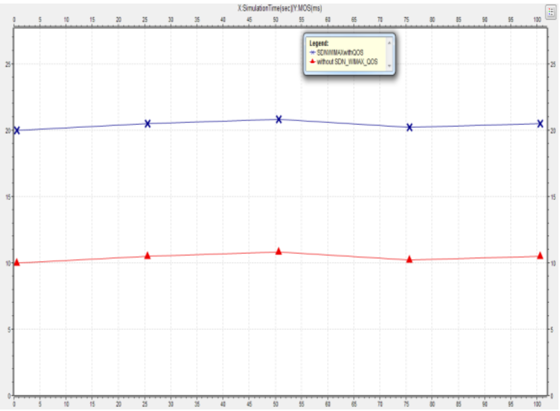


Fig. 7. MOS under WiMAX with and without SDN.

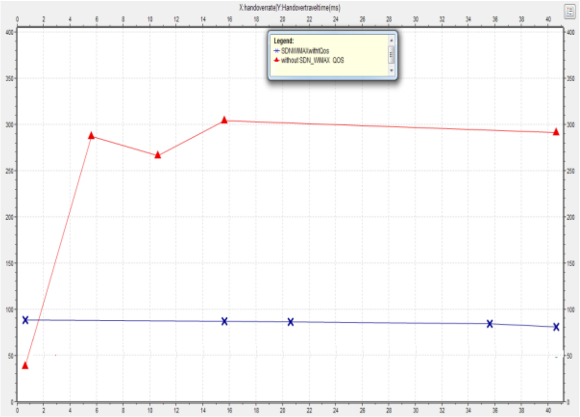


Fig. 8. Handover under WiMAX with and without SDN.

II. Results and Discussion of Simulation in Handover Criteria under WiMAX with and without SDN

Figure 8 shows through two scenarios that the transfer time of the nodes in a WiMAX network with SDN is stable during all communication with a low time (0.7 ms), whereas without SDN, it has the value of 30 ms, then SDN makes it possible to minimize the time of movement from one node to another by the centralization at the controller level.

IX. Conclusion

SDN controller is the best way that allows developers to change Network control with programming efficient algorithms. In this article we proof that our algorithm in comparison with existing SDN controller is very suitable for WIMAX network to more guarantee QOS in Handover.

References

[1] Hung-Yu Wei, S. GangulyR, Iz mailov, ‘‘Interference-aware IEEE 802.16 WiMax mesh networks’’, IEEE 61st Vehicular Technology Conference.

[2] http://4glte.over-blog.com/article-le-wimax-ou-le-lte-3-5-ghz-86329750.html, Jun 2018.

[3] Allan Borges Pontes, Diego dos Passos Silva, Jose Jailton, ‘‘Handover management in integrated WLAN and mobile WiMAX networks’’, IEEE Wireless Communications : Volume: 15, Issue: 5, October 2008.

[4] Fatima LAASSIRI, Mohamed MOUGHIT, Noureddine IDBOUFKER, ‘‘Evaluation of the QoS Parameters in Different SDN Architecture using Omnet 4.6++’’, IEEE, March 2018.

[5] Ronny Yongho Kim, Jin Sam Kwak, Kamran Etemad, ‘‘WiMAX femtocell: requirements, challenges, and solutions’’, IEEE Communications Magazine, Volume: 47, Issue: 9, September 2009.

[6] James A. Hutchison, IV Rotem Cooper, Paul T. Williamson, ‘‘Subscriber station with dynamic multi-mode service acquisition capability’’, App/Pub Number : US10254143 , 2005.

[7] Brian Ferris, Dieter Fox, Neil Lawrence†, ‘‘WiFi-SLAM Using Gaussian Process Latent Variable Models’’, IJCAI-07 2480, 2007.

[8] Ala'a Al-Habashna ; Octavia A. Dobre ; Ramachandran Venkatesan ; Dimitrie C. Popescu , ‘‘Second-Order Cyclostationarity of Mobile WiMAX and LTE OFDM Signals and Application to Spectrum Awareness in Cognitive Radio Systems’’, IEEE Journal of Selected Topics in Signal Processing , Volume: 6 , Issue: 1 , Feb. 2012.

[9] David Martín-Sacristán, Jose F. Monserrat, Jorge Cabrejas-Peñuelas, Daniel Calabuig, Salvador Garrigas, Narcís Cardona, ‘‘On the Way towards Fourth-Generation Mobile: 3GPP LTE and LTE-Advanced’’, Springer 03 August 2009.

[10] Douglas N. Knisely, Takahito Yoshizawa Frank Favichia Standardization of femtocells in 3GPP’’, I EEE Communications Magazine (Volume: 47 , Issue: 9 , September 2009.

[11] Zakhia Abichar and J. Morris Chang, Chau-Yun Hsu, “WiMAX vs. LTE: Who Will Lead the Broadband Mobile Internet?’’, 1520-9202/10/$26.00 © 2010 IEEE, IT Pro May/June 2010.

[12] Mohammed Torad, Ahmed El Qassas, Hadia Al Henawi, ‘‘Comparison between LTE and WiMAX based on system level simulation using OPNET modeler (release 16)’’, IEEE, June 2011.

[13] Steven Gringeri; Nabil Bitar; Tiejun J. Xia, ‘‘Extending software defined network principles to include optical transport’’, IEEE Communications Magazine, Volume: 51, Issue: 3, March 2013.

[14] A. Kumar; E. Altman; D. Miorandi; M. Goyal, ‘‘New insights from a fixed point analysis of single cell IEEE 802.11 WLANs’’, Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, 22 August 2005.

Cryptography using Random Rc4 Stream Cipher on SMS for Android-Based Smartphones

Rifki Rifki[1], Anindita Septiarini[2], Heliza Rahmania Hatta[3]
Department of Computer Science, Faculty of Computer Science and Information Technology,
Mulawarman University, Jl. Panajam Kampus Gn. Kelua, Samarinda, Indonesia.

***Abstract*—Messages sent using the default Short Message Service (SMS) application have to pass the SMS Center (SMSC) to record the communication between the sender and recipient. Therefore, the message security is not guaranteed because it may read by irresponsible people. This research proposes the RC4 stream cipher method for security in sending SMS. However, RC4 has any limitation in the Key Scheduling Algorithm (KSA) and Pseudo Random Generation Algorithm (PRGA) phases. Therefore, this research developed RC4 with a random initial state to increase the randomness level of the keystream. This SMS cryptography method applied the processes of encryption against the sent SMS followed by decryption against the received SMS. The performance of the proposed method is evaluated based on the time of encryption and decryption as well as the average correlation value. Based on the time, it shows that the length of the SMS characters sent affects the time of encryption and decryption. Meanwhile, the best correlation value achieved 0.00482.**

Keywords—Cryptography; SMS security; RC4 stream cipher; random initial state; correlation value

I. Introduction

Cryptography is a science to protect data or information from irresponsible people by turning it into a form where the attacker cannot recognize the data or information while in the processes of storing and transmitting [1]. Moreover, it can be applied to communication services through wireless systems for the communication applications of cellular and wireless [2]. It consists of two-phase, namely encryption and decryption. The encryption is implemented to make data unreadable, invisible or incomprehensible during transmission or storage. While the opposite of it is decryption to reverse the encryption data become an original text [3]. Nowadays, there are various types of smartphones are widely used by the public, one of them is an Android-based smartphone. However, SMS service has not a security method on Android smartphones. SMS is a text messaging service that allows cellular customers to send the text to each other. Global System for Mobile Communication (GSM) uses as a tool for sending SMS messages. SMS message sent by the user, then it was stored by the SMSC to forward to the target mobile device. SMSC uses a store-and-forward technique to store messages to forward it to the target device. If the Home Location Register (HLR) of the target mobile device is active, then SMSC will transfer the SMS message to target mobile device. SMSC receives the verification message that confirms the delivery status of SMS message to the target mobile device [4]. The maximum length of an SMS without the image/graphic is 160 characters using 7 bits or 70 characters using 16 bits of character encoding [5].

Cryptographic methods are divided based on key-based and keyless [6]. Several conventional keyless cryptographic methods have implemented for improving data security such as Caesar ciphers [7], Vigenere ciphers [8], [9], Zigzag ciphers [10], and Playfair cipher [11]. Those methods are more complex and consume a significant amount of power when applied in the resource-constrained devices for the provision of secure communication [12]. Another method that has used is key-based with Symmetric Cryptography. The type of encryption that used is to provide end-to-end security to SMS messages. This method is appropriate for mobile devices because of limited resources, namely limited power/energy, insufficient memory, and less processing power [4]. The examples of symmetric key cipher block cryptography are AES, DES, and 3DES [3].

Several methods have been performed on SMS services such as AES [13,14], Blowfish [5,15], One-Time Pad Cipher [16], MNTRU [17], Certificate-Less Public Key Cryptography (CL-PKC) to Authentication over a GSM System [18]. Several previous works have developed RC4 for WEP [19], combined RC4 with a genetic algorithm [20] and compared RC4 with other methods. RC4 is one of the most popular stream ciphers in symmetric key cryptography since it uses in several security protocols. Moreover, it has the higher speed and the lower complexity than other stream ciphers. The data of statistics show that the RC4 algorithm is used to protect 50% of TLS traffic as the most widely used secure communication protocol on the internet nowadays [21]. RC4 has a secret internal state and works by generating the pseudorandom stream of bits [22]. The internal state of RC4 consists of an S-box array permutation of 256 bytes from the number 0...., N - 1 and two indices i, j $\in \{0,. . . , N - 1\}$. The index i is determined and known to the public, while j and S-box permutations remain confidential [23,24]. The RC4 algorithm consists of the Key Scheduling Algorithm (KSA) used for initializing S-box using variable length key and Pseudo-Random Generation Algorithm (PRGA) to generate keystream bytes.

In the previous researches, RC4 stream cipher compared to AES [25] shows that the performance of RC4 is better than AES which based on the throughput, CPU processing time, memory utilization, encryption time and decryption time. Subsequently, RC4 compared to Blowfish method [17] shows that RC4 has better encryption performance while Blowfish has better decryption performance for small message texts. However, RC4 has better performance in power consumption for communication. Furthermore, the comparison of RC4 with RSA [23] shows that the algorithm of RC4 better than RSA based on the presented experimental and analytical results of both algorithms evaluated. RC4 has more excellence in execution speed and throughput compared to several other cryptographic methods such as VMPC, HC-128, HC-256, Salsa20 and Grain [24]. Nevertheless, RC4 has a limitation in the KSA and PRGA phases due to the initialization process which produces sequential numbers (0,1,2, ..., 255) that may provide the opportunities for hackers [26]. The development of RC4 with a random initial state will increase the randomness level of the keystream produced by RC4 [23,24].

The development of cryptography in SMS service security is an important and challenging issue. It caused the hackers may steal the contents of the original message of the SMS sent. This research proposed a cryptography method using the RC4 stream cipher on SMS for Android-based smartphones to overcome this issue. The contribution of this research is the use of the initial random state to increase the randomness level of the keystream. This method is expected to increase the level SMS service security.

II. Research Method

This research aims to implement cryptography on SMS for Android-based Smartphones using the RC4 stream cipher method with a random initial state to increase the randomness level of the keystream. The proposed method consists of two main stages, namely encryption, and decryption. The illustration of the cryptography system on sending the message via SMS is shown in Fig. 1. Based on Fig. 1, the initial process of this system is the sender and receiver as the user must apply the process of login. Afterward, in the encryption process, the sender should send the message (plaintext) and the key simultaneously. The message is sent as ciphertext as the implementation result of the RC4 method. Subsequently, the information of the sender's identity, key, and keystream are saved in the server. Meanwhile, in the decryption process, the receiver who has successfully login receives the ciphertext, key, and keystream from the server according to the message. Messages can be decrypted into plaintext based on the key similarity during message encryption.

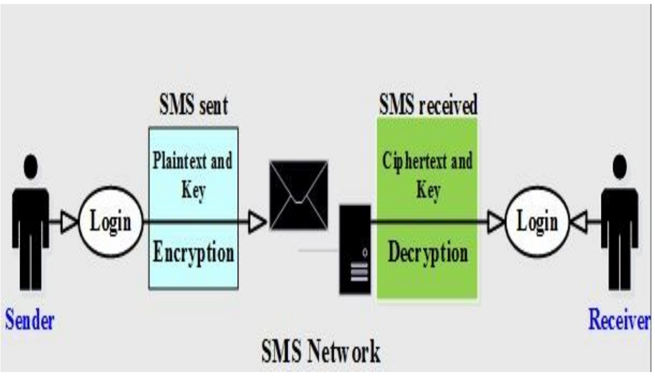


Fig. 1. The Illustration of the Cryptography System.

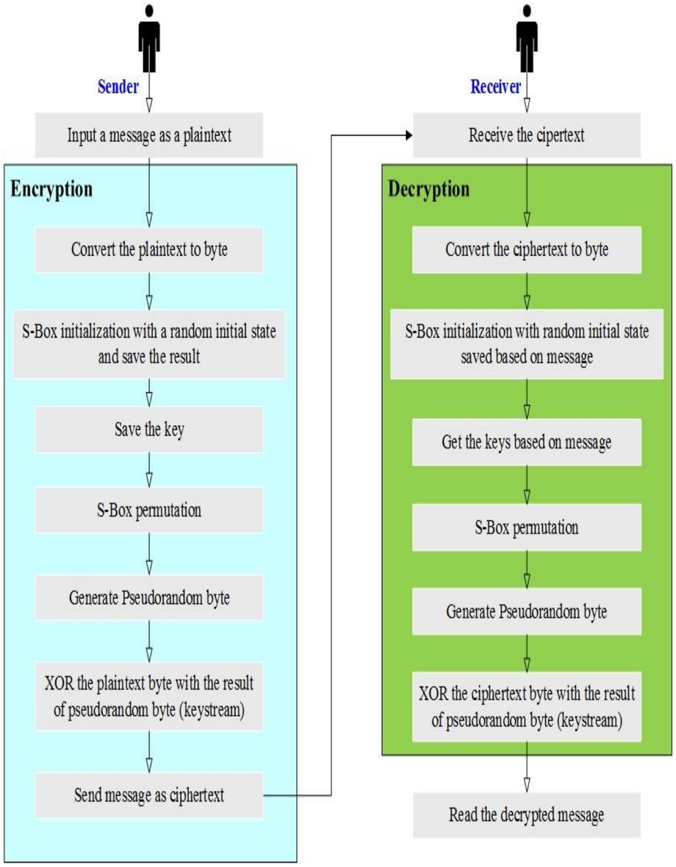

Sender
Receiver
Input a message as a plaintext
Receive the cipertext
Encryption
Convert the plaintext to byte
S-Box initialization with a random initial state and save the result
Save the key
S-Box permutation
Generate Pseudorandom byte
XOR the plaintext byte with the result of pseudorandom byte (keystream)
Send message as ciphertext
Decryption
Convert the ciphertext to byte
S-Box initialization with random initial state saved based on message
Get the keys based on message
S-Box permutation
Generate Pseudorandom byte
XOR the ciphertext byte with the result of pseudorandom byte (keystream)
Read the decrypted message

Fig. 2. The Stage Diagram of the Proposed RSA Method.

There are several similar processes in the stage of encryption and decryption, namely, (1) convert the plaintext/ciphertext to byte, (2) S-Box initialization with a random initial state, (3) S-Box permutation, and (4) generate pseudorandom byte to obtains the keystream. The keystream is used to implement XOR operation between the plaintext and ciphertext in a byte. The difference in both stages is in the process of saving the random initial state results and the key in the encryption process. In the decryption process, the results of the random initial state and keys select from the database based on the message to carry out the subsequent process. The detail process of the proposed method is depicted in Fig. 2.

A. *Encryption*

In this work, encryption is applied to encode the messages sent with the aim only the authorized people whose can access the messages. KSA phase of this work, RC4 allow producing the similar state even though two different keys used and a similar keystream output generated. This case is known as a key collision or related key pairs [27]. It is caused by the initialization process which produces the numbers of 0, 1, 2, .., 255, sequentially which opens opportunities for hackers. The proposed method of the RC4 stream cipher with a random initial state will increase the randomness level of the keystream. Development in KSA phase produces N values from 0 to N-1 without duplication by a pseudo-random number generator which distributes as an additional secret key. The steps of the encryption process in this work are as follows:

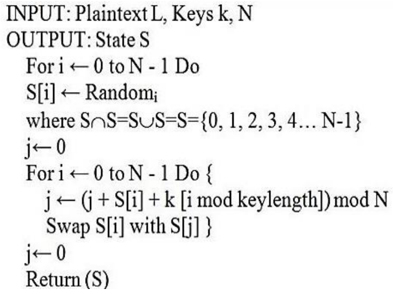
1) Get the ASCII values from the messages sent as the plaintext then they are converted to bytes.

2) In KSA phase, the initialization of the S-Box array with a random initial state followed by saving the key and the S-Box permutation. This step is implemented to produce random values between 0 and 255 without duplication.

3) Initialize the keys array then save it.

4) S-Box permutation is performed against the values in the S array by exchanging the contents of the S [i] with S [j].

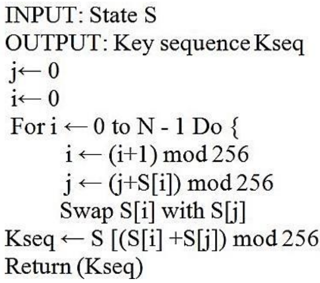
The Pseudocode of this step is as follows:



```
INPUT: Plaintext L, Keys k, N
OUTPUT: State S
  For i ← 0 to N - 1 Do
  S[i] ← Random$_i$
  where S∩S=S∪S=S={0, 1, 2, 3, 4… N-1}
  j← 0
  For i ← 0 to N - 1 Do {
    j ← (j + S[i] + k [i mod keylength]) mod N
    Swap S[i] with S[j] }
  j← 0
  Return (S)
```

Input in KSA phase is Plaintext L, Keys k, and N where the message length of plaintext L is the initial key length in bytes, N is the size of the array S, and i and j are indexed pointers. The output of this phase is the array S.

5) In the PRGA phase, retrieving the values of S [i] and S [j] aims to sum up those values in the form of modulo 256. This phase obtains a keystream. The Pseudocode of this step is as follows:



```
INPUT: State S
OUTPUT: Key sequence Kseq
j← 0
i← 0
For i ← 0 to N - 1 Do {
        i ← (i+1) mod 256
        j ← (j+S[i]) mod 256
       Swap S[i] with S[j]
Kseq ← S [(S[i] +S[j]) mod 256]
Return (Kseq)
```

Input in PRGA phase states S where N is the size of the array or state S, and i and j are indexed pointers. The output of this phase is array byte of Kseq using for XOR-ing with plaintext for obtaining ciphertext.

6) Performing XOR-ing keystream, plaintext bytes, obtain ciphertext.

7) Send the SMS message as ciphertext.

B. *Decryption*

The input of this stage is ciphertext. Decryption aims to reproduce the plaintext, which performed by decoding the ciphertext. The steps of the decryption process are as follows:

1) Get the ASCII values from the received message as the ciphertext then they are converted to bytes.

2) The initialization of the S-Box array in KSA has applied the similar step as in encryption based on the saved key.

3) Get the key based on the message.

4) The processes of S-Box permutation and generate Pseudorandom byte are performed similarly as in encryption.

5) Performing XOR-ing keystream, ciphertext bytes, to obtain the plaintext.

6) Receive the decrypted message as plaintext.

TABLE I. THE SPECIFICATION OF THE VARIOUS SMARTPHONES

Android Version	**Smartphone specification**	
	Processor	***RAM***
Version 4.4 (KitKat)	1.6 GHz	2 GB
Version 5.0 (Lollipop)	2.2 GHz	3 GB
Version 6.0 (Marshmallow)	1.8 Ghz	3 GB
Version 7.0 (Nougat)	2.4 Ghz	4 GB

III. RESULT AND ANALYSIS

This research used 225 message SMS with variations in character length as experimental data. Those data were sent using four smartphones with different specification. The specifications of the smartphone are summarized in Table 1.

The proposed method was evaluated based on the time of encryption and decryption as well as the correlation value between plaintext and ciphertext. The correlation value indicates the quality of encrypted data. This value lies between -1 and 1. The correlation value is defined as follows [28]:

$$|r| = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{\left[n\sum(x^2) - (\sum x)^2\right]\left[n\sum(y^2) - (\sum y)^2\right]}} \qquad (1)$$

Where *r* is the correlation value, *x* is the ASCII code value of plaintext and *y* is the ASCII code value of ciphertext. The correlation value should be close to 0 for a good method.

The proposed method aims to encode the SMS message sent as plaintext to ciphertext using the key with the RC4 with a random initial state. Several examples of the encryption result in the form of ciphertext obtaining based on the plaintext and key is shown in Table 2. The evaluation of this proposed method divides into two ways based on the time of encryption and decryption, and the value of the correlation between plaintext and ciphertext.

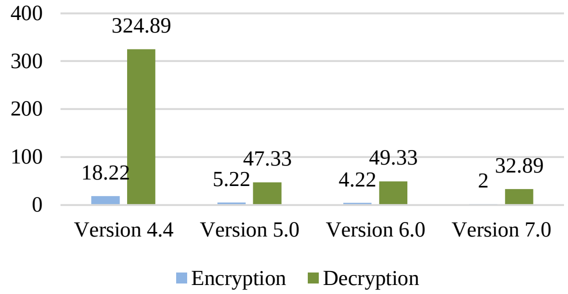


Fig. 3. The Evaluation Results based on the Time of Encryption and Decryption (in Millisecond) using Smartphones with the Different Android Version: Kitkat (Version. 4.4), Lollipop (Version. 5.0), Marshmallow (Version 6.0) and Nougat (Version. 7.0).

A. Evaluation based on the Time of Encryption and Decryption

This evaluation applied by comparing the time of encryption and decryption in four types of smartphones. Those smartphones built in the different Android version, namely version 4.4 (KitKat), version 5.0 (Lollipop), version 6.0 (Marshmallow) and version 7.0 (Nougat). The specification details of those smartphones are presented in Table 1. The performance of this proposed method based on the time evaluation against the SMS message and key with the various length of the character as shown in Table 2. Furthermore, the performance comparison of this proposed method based on the time of encryption and decryption in four types of smartphones is depicted in Fig. 3.

TABLE II. THE RESULT OF ENCRYPTION USING THE PROPOSED METHOD

Key	**Plaintext**	**Ciphertext**
first	first test	hZÚ9Áú·Ÿ
application user interface	Android device	Èå³åÖù'=£OÑ¤¬
SMS cryptographic system for safe data user	RC4 STREAM CIPHER	]oÂ âd‡ • ÏkÀ…L
Encryption	plaintext and ciphertext generated by system	ECMĐW)ãºuE÷Ff· ,pÎ XÂ˜Ìwt ±ˆŸÒt™>@ Ù]Å¸
medium SMS and medium key	RC4 has better performance and efficiency	fVDL]zwµãl¿(8ÍÏ¯…l^mÁAÄ 4 aþËpŠ?GT
Michael Jackson - earth song (cover by me)	what about sunrise what about rain what about all	-‰?;W¦•&)õðõlÚ‰qšçÊÏÔŽ ¹bX¢Ì h,]aŸ□ èá“›O¹D
Love	you say you love me, I say you crazy. We're nothing more than friends	âñf˜L!«:Ä□ /×kÃÂô&ÏÖ‰W.5Ìñï¿ ›ïc□ ›T“m1¼I"– 76móâ…´²OÎðÓøvCFs£ñu
Department of computer science	RC4 has become the most popular stream cipher in the history	Ü¡$Æ¸ç=†□ Ô‡ úîL□ SKW2å\½”ðy\• ÛOÖj× ®ÔÑK óß□ ©‘ä~ïyœ«—½ø
end to end encryption method for safe data	Key Scheduling Algorithm and Pseudo Random Generation Algorithm	E©úÖ=)Í§À0]àÛs¿‹ÞT6ÙVÿ¼gïl× 0\(å‰ƒƒ},J@4÷‘ÁU?w‡ž -!,§ •¿

Based on the experiment result as shown in Fig. 3, the time of encryption and decryption is influenced by the smartphone specification and the number of characters of the SMS message. Related to the smartphone specification, the time of encryption and decryption of Version 5.0 is faster than Version 6.0 due to the processor speed of Version 5.0 higher than Version 6.0 are 2.2 GHz and 1.8 GHz, respectively. In addition, the specification of RAM and the number of applications that are running on the smartphone may affect the time of encryption and decryption. Moreover, the increasing number of characters in an SMS message cause the time of the encryption and decryption to become longer.

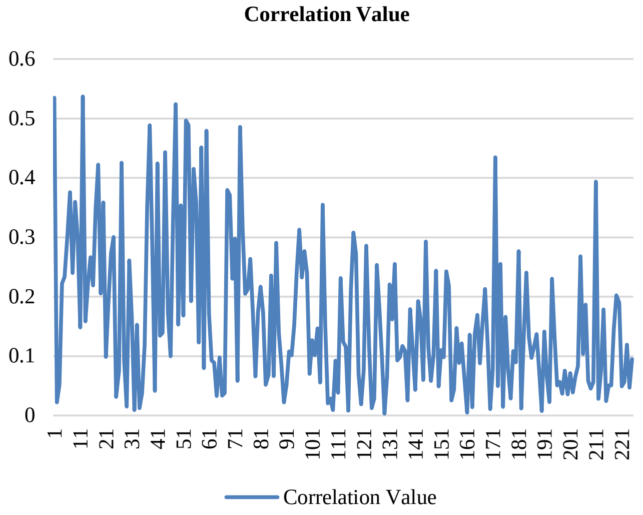


Fig. 4. The Evaluation Results of the Quality of Encrypted Data based on the Correlation Values of the Plaintext and Ciphertext.

B. Evaluation based on the Correlation Values

This research also evaluated the quality of encrypted data which obtains by the proposed method based on the correlation values. Low correlation value (close to 0) indicates that the encryption system is becoming more secure [26]. The correlation values are computed based on the ASCII values of plaintext and ciphertext using Eq. (1). All correlation values which generated from 225 examples of experimental data are summarized in Fig. 4. Based on the result is presented in Fig. 4, the correlation values of the best, the worst and the average achieve of 0.00337, 0.53716 and 0.10188, respectively. These results indicate that the increasing number of characters in the SMS message and the key, the correlation value is getting closer to 0. In this research, the correlation value is getting closer to 0 if the number of keys and plaintext character more than 30 and 25, respectively. Otherwise, the correlation value tends to closer to 1. The resulting correlation value only influenced by the number of characters from the plaintext and the key used but not affected by the smartphone specification used.

Furthermore, this research also calculated correlation values based on the plaintext and ciphertext which produced by other methods, namely Vigenere and Playfair. A summary of the performance comparisons between those methods and the proposed method is presented in Table 3. Table 3 shows that the proposed method produces the lowest correlation value of 0.10188. It indicates that the quality of the encryption data of the proposed method yields the best result than the other methods.

TABLE III. PERFORMANCE CORRELATION VALUE OF PREVIOUS METHODS AND PROPOSED METHOD

Method	**Correlation value**
Vigenere cipher	0.81229
Playfair cipher	0.21345
Proposed method	**0.10188**

IV. Conclusion

This paper developed the RC4 method with a random initial state. The random initial state is needed to increase the randomness of the keystream so that this method is safer than RC4 without a random initial state. The proposed method evaluated using 225 data. Based on the evaluation result, the time of encryption and decryption is influenced by the characters, number of the SMS message and the key as well as the smartphone specification. Meanwhile, the correlation value is only affected by the characters number of the SMS message and the key. The correlation value of the proposed method shows an improvement compared to the method of Vigenere and Playfair. For future works, other cryptographic methods are still possible to be developed to reduce correlation values.

Acknowledgment

The authors would like to thank KEMENRISTEK DIKTI for Bidikmisi Scholarship. We also would like to thank the Department of Computer Science, Faculty of Computer Science and Information Technology, Mulawarman University for the facilities during this research.

References

[1] A.A. Zaidan, A. Majeed, and B.B. Zaidan, “High Securing Cover-File of Hidden Data Using Statistical Technique and AES Encryption Algorithm,” Int J Comput Inf Eng, vol. 3, pp. 463–74, 2009.

[2] I. Memon, I. Hussain, R. Akhtar, and G. Chen, “Enhanced privacy and authentication: An efficient and secure anonymous communication for location based service using asymmetric cryptography scheme," Wireless Personal Communications, vol. 84, pp.1487–1508, 2015.

[3] H.O. Alanazi, B.B. Zaidan, A.A. Zaidan, H.A. Jalab, M. Shabbir, and Y. Al-Nabhani, “New Comparative Study Between DES, 3DES and AES within Nine Factors,” J Comput, vol. 2, pp. 152–7, March 2010.

[4] M.W. Khan, “SMS Security in Mobile Devices : A Survey,” Int J Adv Netw Appl, vol. 5, pp. 1873–82, 2013.

[5] H.M. El-Bakry, A.E. Taki_El_Deen, and A. H. El tengy, “Implementation of a Hybrid Encryption Scheme for SMS/Multimedia Messages on Android,” Int J Comput Appl, vol. 85, pp. 1–5, January 2014.

[6] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, "A Survey on the Cryptographic Encryption Algorithms," International Journal of Advanced Computer Science and Applications, vol. 8, pp. 333–344, 2017.

[7] A. Jain, and A. Patil, “Enhancing the Security of Caesar Cipher Substitution Method using a Randomized Approach for more Secure Communication,” Int J Comput Appl, vol. 129, pp. 975–8887, 2015.

[8] S.K. Mandal, and A.R. Deepti, “A Cryptosystem Based On Vigenere Cipher By Using Mulitlevel Encryption Scheme,” Int J Comput Sci Inf Technol, vol. 7, pp. 2096–2099, 2016.

[9] H. Hamdani, H. Ismanto, A.Q. Munir, B. Rahmani, A. Syafrianto, D. Suprihanto, and A. Septiarini, “The Proposed Development of Prototype with Secret Messages Model in Whatsapp Chat,” Int. J. Electr. Comput. Eng., vol. 8, pp. 3841-3849, 2018.

[10] Mu. Annalakshmi, and A. Padmapriya, “Zigzag Ciphers: A Novel Transposition Method,” IJCA Proceedings on International Conference on Computing and information Technology, IC2IT(2):8-12, December 2013.

[11] S. Bhattacharyya, N. Chand, and S. Chakraborty, ”A Modified Encryption Technique using Playfair Cipher 10 by 9 Matrix with Six Iteration Steps,” Int J Adv Res Comput Eng Technol, vol. 3, pp. 307–12, 2014.

[12] K. Sharma, M Ghose, and D. Kumar, “A comparative study of various security approaches used in wireless sensor networks,” Int J Adv Sci Technol, vol. 17, pp. 31–44, 2010.

[13] B. Patil, “SMS Security Using RC4 & AES,” Indian J Sci Res, vol. 11, pp. 34–8, 2015.

[14] R. Rayarikar, S. Upadhyay, and P. Pimpale, “SMS Encryption using AES Algorithm on Android,” Int J Comput Appl, vol. 50, pp. 12–17, 2012.

[15] A. Kaur, and R. Dhadwal, “Performance Comparison of Symmetric Algorithms for SMS Communication,” Int J Adv Res Comput Commun Eng, vol. 4, pp. 62–64, 2015.

[16] M. Iqbal, M. A. S. Pane, and A. P. U Siahaan, “SMS Encryption Using One-Time Pad Cipher,” IOSR J Comput Eng, vol. 18, pp. 54–58, 2016.

[17] S. Jha, and U. Dutta, “Review on SMS Encryption using MNTRU Algorithms on Android,” Int J Comput Sci Inf Technol, vol. 6, pp. 3855–3858, 2015.

[18] I. Memon, M. R. Mohammed, R. Akhtar, H. Memon, M. H. Memon, and R. A. Shaikh, “Design and implementation to authentication over a GSM system using certificate-less public key cryptography (CL-PKC)," Wireless personal communications, vol 79, pp. 661–686, 2014.

[19] L. Stošić, and M. Bogdanović, "RC4 stream cipher and possible attacks on WEP," International Journal of Advanced Computer Science and Applications, vol. 3, pp. 110–114, 2012.

[20] B. Ferriman, and C. Obimbo, "Solving for the RC4 stream cipher state register using a genetic algorithm," International Journal of Advanced Computer Science and Applications, vol. 5, pp. 216–223, 2014.

[21] E. Taqieddin, O. Abu-Rjei, K. Mhaidat, and R. Bani-Hani, “Efficient FPGA Implementation of the RC4 Stream Cipher using Block RAM and Pipelining,” Procedia Comput Sci, vol. 63, pp. 8–15, 2015.

[22] J. Chen, and A. Miyaji, “Novel strategies for searching RC4 key collisions,” Comput Math with Appl, vol. 66, pp. 81–90, 2013.

[23] A. A. Okedola, and Y. N. Asafe, “RSA and RC4 Cryptosystem Performance Evaluation Using Image and Text File,” Int J Sci Eng Res, vol. 6, pp. 289–294, 2015.

[24] S.O. Sharif, and S.P. Mansoor, “Performance Analysis Of Stream And Block Cipher Algorithms,” Int. Conf. Adv. Comput. Theory Eng., vol. 1, pp. 522–525, 2010.

[25] N. Singhal, and J. P. S. Raina, “Comparative Analysis of AES and RC4 Algorithms for Better Utilization,” Int J Comput Trends Technol, vol. 2, pp. 177–181, 2011.

[26] M. M. Hammood, K. Yoshigoe, and A. M. Sagheer, “RC4 Stream Cipher with a Random Initial State,” Lect. Notes Electr. Eng., vol. 253 LNEE, pp. 407–415, 2013.

[27] P. Jindal, and B. Singh, “RC4 encryption - A literature survey,” Procedia Comput Sci, vol. 46, pp. 697–705, 2015.

[28] E. Setyaningsih, C. Iswahyudi, and N. Widyastuti, “Image Encryption on Mobile Phone using Super Encryption Algorithm,” TELKOMNIKA, vol. 10, pp. 837–845, 2012.

Improvement of the Handover and Quality of Service on Software Defined Wireless Networks

Fatima Laassiri[1,a], Mohamed Moughit[2,b]
[a,b]IR2M Laboratory, FST, Univ Hassan UH1- Settat, Morocco
[b]EEA&TI Laboratory, FST, Univ Hassan , Mohammedia, Morocco, [b]National Schools of Applied Sciences Khouribga, Univ Hassan 1, UH1- Settat, Morocco

Noureddine Idboufker[3]
National School of Applied Sciences,
Univ Cadi Ayyad Marrakech,
Morocco

***Abstract*—The Wireless Fidelity (WiFi) is the business name given to the 802.11b and 802.11g IEEE standard by the WiFi Alliance, formerly known as Weca industry with more than 200 member companies dedicated to supporting the growth of wireless LANs. This standard is currently one of the most used standards in the world. The theoretical data rates of 802.11b are 11 Mb/s and 54 Mb/s for 802.11g. This article presents Handover's improvement performance and quality of service (QoS) parameters and they are: end-to-end delay, latency, jitter, lost packets, and Mean Opinion Score (MoS), under networks Wi-Fi with the help of the OMNeT 4.6 ++, by implementation of a new algorithm at the level of the SDN controller that allows handover management without breaking the connection by respecting the priority per class of traffic. The realization of this work is based on the intra-Wi-Fi mobility, that it is adopted by a macro mobility of level 3 and it is MIPv6 as well as it exploited the protocol of Voice over IP that it is SIP, and the implementation of SDN rules on the OpenFlow protocol.**

Keywords—SDN; Wi-Fi; QoS; OpenFlow protocol; handover; SDN controller; OpenFlow switch

I. Introduction

Software Defined Network (SDN) [1] is a new network paradigm; it is used to reduce the complexity of network technology. The following work aims to expose a simulation implemented under OMNeT 4.6++, to assess the performance of Handover and the QoS with two architectures one Wi-Fi [2] without SDN and the second offers the improvement of the implementation the SDN via the OpenFlow protocol [3].

The communication between the two topologies is carried out with the SIP protocol [4], with high quality intra-Wi-Fi mobility that is MIPv6 [5].

Today, we are seeing the evolution of the Internet in a number of users. Among the factors of this evolution is the success of 802.11. The IEEE 802.11 networks are becoming increasingly popular as they allow users to connect to the Internet at an affordable price with relatively large bandwidth and the ability to roam without being disconnected. In addition, nowadays IEEE 802.11 wireless network cards are deployed in the majority of technologies such as PDAs and laptops relatively important and also the ability to move without being disconnected. In addition, nowadays IEEE 802.11 wireless network cards are deployed in the majority of technologies such as PDAs and laptops.

In parallel, multimedia communication techniques have also evolved with the new compression and coding algorithms. Thus, many multimedia applications become accessible from wireless networks. But they still present obstacles to deployment. The major problems of these networks are the lost rate and the delay variation knowing that multimedia applications are very demanding. An obvious solution to optimize bandwidth utilization and improve video quality is to transmit multi-point video to a set of users.

But the use of standard Multipoint has three main problems. The first is the impossibility of adapting the collision windows according to the state of the network. The second is the impossibility of adapting the physical rate according to the state of the transmission medium, sod the packets are transmitted at a fixed physical rate. The third is the impossibility of retransmitting lost packets at the MAC layer.

A new approach has recently been proposed to remedy its problems. It consists of the election of a receiver called leader to ensure the acknowledgment of received packets. Thus, the transmitter can adapt to the physical rate and retransmit the lost packets.

II. Problem and Solutions

Wi-Fi refers to certain types of Wireless Local Area Networks (WLANs) [6] and uses specifications that fall under the 802.11 standard [7]. It is widely used by companies, because of the growing demand, users have ubiquitous access to wireless services, what it led to the deployment of forced use of this wireless access technology such as Wi-Fi, it offers a level of quality within reach, but the problem is that the number of devices is increasing which reduces the performance of travel time Handover and QoS in terms of end-to-end delay, latency, jitter, the number of lost packets, and MoS, for all these reasons, this work proposes as a solution to implemented SDN technology for Wi-Fi to better optimize performance.

III. State of the Art

Before the appearance of SDNs, they are defined by the Open Networking Foundation (ONF) [8] as we know them today, several ideas and works have been proposed before, including network programming, separation of control and data designs. This segment is a brief overview of this work, that he can be considered as ancestors of SDN. The first idea of programming networks was developed in 1996 under the name of “Active Network” (AN) [9]. These networks infusion programs among the packet data. When a network node receives packets, it extracts and executes the programs from the data of the packet and therefore triggers standard activities of transmission, adjustment or concealment of the packet. With this approach, new network administration and routing mechanisms can be implemented without altering transmission equipment. Several studies have been conducted on NAs, especially on smart packets [10]. Since packet can carry malicious programs, and an elective called "Programmable Networks" (PN) [11] was proposed in 1999. The PNs inject programs inside the nodes of the network. These nodes run the programs only after a signalling and verification stage, to enhance security. ANs and PNs have sought to introduce programmability into networks through packets and programmable switches. These approaches did not reduce the complexity of the network infrastructure. In 1998, the Internet Engineering Task Force (IETF) working group could propose a General Protocol for Managing Switches (GSMP) [12], also another project called 4D was launched in 2005 [13] to separate the routing decisions and the protocols that govern the connections between network devices. Dispersal and discovery designs collect information from the network and send it to the decision plan, which has a global view of the network, to control the transmission of traffic flowing through the data model. The beginning of the SDN networks started with the Ethane project [14], launched in 2006 at the University of Stanford. In fact, that it defines a new engineering for business networks. Ethane goal was to have a centralized controller to manage the rules (Arrangements) and security in the network. Ethane uses two components: A controller to decide whether a packet should be forwarded, and an Ethane switch consisting of a table and a match string between the two. He was a source of inspiration for a networking operating system called Nox [15], and for a new idea called today Software Defined Network (SDN). Noting that Ethane's researchers are behind Nox and SDN [16].

IV. IEEE 802.11 Standards

IEEE 802.11ac is a wireless standard for the Wi-Fi family, standardized by IEEE; it allows a high-speed wireless connection to a local area network and uses only a frequency band between 5 and 6 GHz, with variations depending on the country. This frequency band is commonly named: “band of 5 GHz” [17].

The aggregated channels allow, under ideal radio conditions, a theoretical throughput of up to 1.3 Gbit/s and the throughput of 910 Mbit/s (using four channels occupying an 80 MHz sub-band), up to 7 Gbit/s overall throughput [18].

V. Openflow Protocol

The OpenFlow protocol is defined by the ONF, this non-profit consortium dedicated to the development and standardization of SDN. It uses the TCP protocol via port 6633[19]. The communication uses a secure channel based on TLS.

OpenFlow protocol supports three types of messages [20]:

Controller to switch messages: These messages are sent only by the controller to the switches, they perform the switch configuration functions, they exchange information about the capabilities of the switch and they also manage the flow tables.

Symmetrical messages: These messages are sent back and forth signalling the connection problems of the switch controller.

Asynchronous messages: These messages are sent by the switch to the controller to announce changes in the network and switch status. [20]

A. Layers of SDN

Fig. 1 offers the general architecture of the SDN layers and the OpenFlow protocol.

Infrastructure layer [21]: For an OpenFlow protocol implementation function of a network element, the part of the equipment that it provides an API and an interface to the controller.

Control layer[22] : Responsible for making decisions about how packets should be transmitted by one or more network devices, and push those decisions to network devices for execution.

Northbound SDN Interfaces (NBI) [23]: NBI interfaces of SDN are interfaces between SDN applications and controllers, they typically provide abstract network views, and they allow direct expression of network behaviour and requirements.

Applications SDN [24]: This layer is a program; it communicates the necessary behaviours and resources with the SDN controller through application programming interfaces (APIs). In addition, applications can build an abstract view of the network by gathering information from the controller for decision-making purposes.

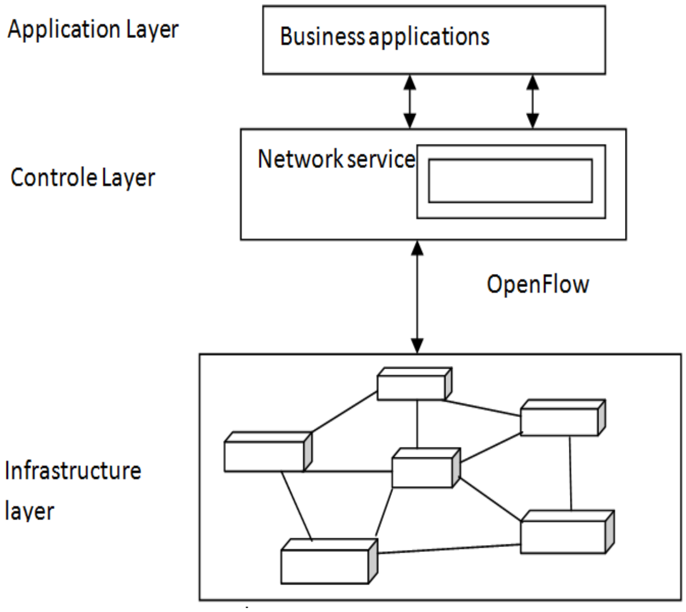


Fig. 1. SDN Network Architecture.

VI. Methods for Implementing the Wi-Fi with and without SDN under OMNET 4.6++

This section describes two simulation topologies implemented under OMNeT 4.6++, with the first presenting a Wi-Fi network without SDN (Fig. 2) and the same with the addition a new algorithm at the Controller SDN, that it is based on the OpenFlow protocol (Fig. 3).

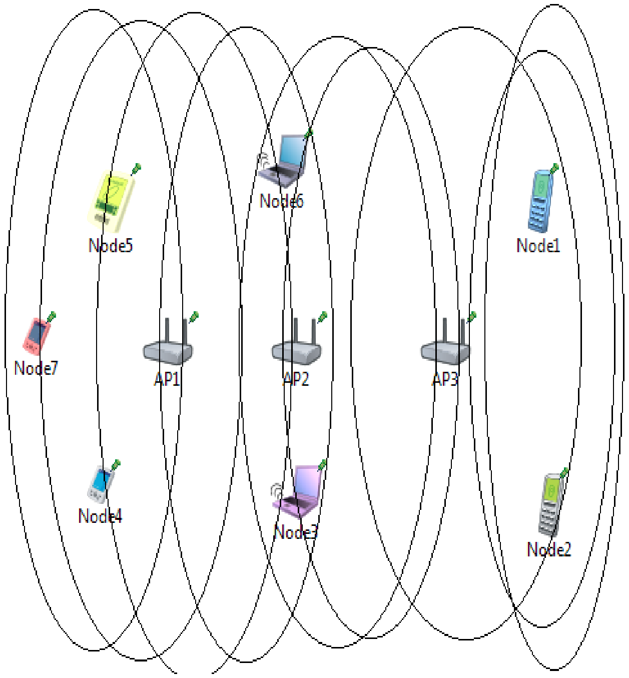


Fig. 2. Scenario 1: Wi-Fi Network Architecture without SDN.

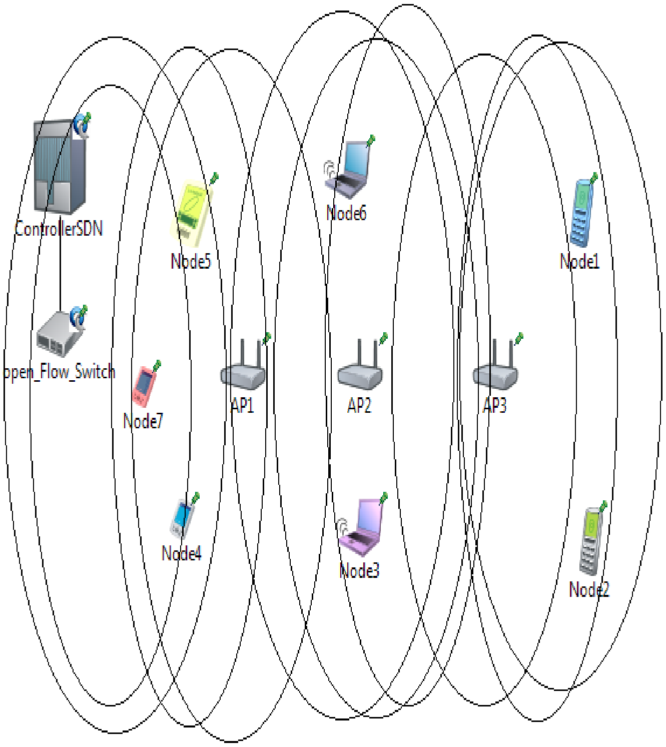


Fig. 3. Scenario 2: Wi-Fi Network Architecture with SDN.

For the realization of this work, the mobility used intra Wi-Fi, it occurs when a user device moves from one node to another, with a manipulation of a level 3 Macro Mobility, taking advantage of the MIPv6 protocol for mobile nodes to move randomly across the Internet, while continuing to receive their datagrams at a fixed address.

Each access point has at least two network interfaces: An 802.11 interface communicates with 802.11 clients and the second Ethernet-type interface connects to the main wired network. Packets received from the customer will be transmitted from one side to the other... Transmitting a package or not. It is determined by consulting the flow tables. That it consists of a set of rules, maintained by the controller via the interface OpenFlow. If the access point enabled by OpenFlow finds no corresponding rule in the flow table of a packet, it will ask the controller to process the packet.

The OpenFlow controller schedule coordinates the transmission between access points. The responsibility of scheduling algorithm is running in the controller and reducing the occurrence of conflicts and retransmission.

The OpenFlow interface is simply started, with the abstraction of a single table of rules it could match packets on a dozen header fields (MAC addresses, IP addresses, protocols, TCP / UDP Port numbers, etc.). In the last five years, the specification has become more and more folded, with many of the header fields and multiple stages of the rule tables, to allow the switches to expose more of their capabilities to the controller.

The OpenFlow specification allows future switches to support exile mechanisms to parse the corresponding packets and header fields, allowing controller applications to take advantage of these capabilities through a common open interface.[25]

A. *Components of the SDN Architecture with and without SDN*

The Wi-Fi architecture with and without SDN consists of the following:

Station (STA): Client device in an 802.11 (Wi-Fi) that it has chip and antenna such as a computer, a laptop or a smartphone. The term STA is sometimes used for the access point (AP), in which case an STA is a device communicating via the 802.11 protocol.

An access point (AP): Is a device of a wireless LAN, or WLAN, usually in an office or in a large building. It connects to a wired router, switch, or Ethernet cable hub and delivers a Wi-Fi signal to a dedicated zone.

SDN Controllers: To provide a layer of abstraction of the network and present it as a system. It allows to quickly implement a change on the network by translating a global demand (for example: Prioritize application X) in a sequence of operations on network devices (OpenFlow Additions States). Orders are given to the controller by an application via a so-called API “Northbound” or north. Controller software vendors publish the API documentation to allow applications to interface. The controller communicates with the equipment via one or more APIs called "Southbound" or south.

OpenFlow is positioned as a south API acting directly on the data plane. Other APIs can act on the management plan or control. A controller can even speak directly in CLI with a device to activate a feature. [26]

OpenFlow switch: It is a physical switching device that contains a number of ports and queues, it is based on the OpenFlow protocol.

VII. Results and Discussion of Simulation in Quality of Service Criteria (QoS)

This section presents Handover and QoS performance evaluation such as end-to-end delay, latency, jitter, lost packet, and MoS. Under Wi-Fi without SDN and with the latter's increment to determine the impact of implementation the new algorithm in SDN controller.

A generic access point supports multiple wireless radios and multiple Ethernet ports. The Ethernet MAC type of the relay unit is wireless card type.

By default, the access point is stationary (Mobility Stationary), but it can also be configured by parameters.

A. End-to-end Delay under Wi-Fi without and with SDN

Fig. 4 shows the end-to-end delay results in the Wi-Fi without SDN scenario with a higher value of (35 ms) compared to the SDN-based scenario which has a reliable delay of (22 ms), which explains why adding the latter is beneficial.

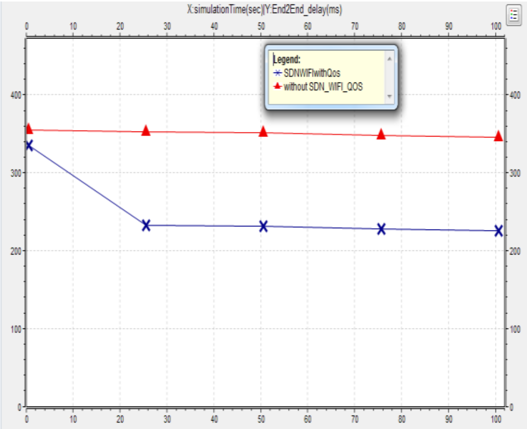


Fig. 4. End-to-end Delay in Wi-Fi SDN-Free and SDN-Based Scenarios.

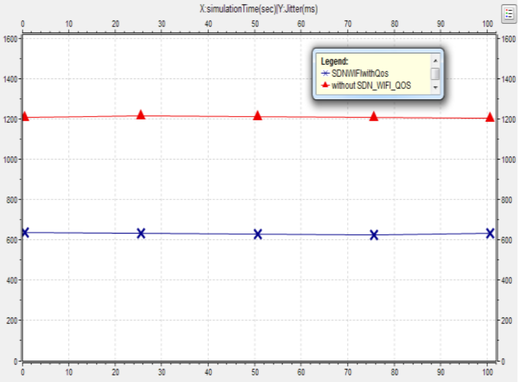


Fig. 5. Jitter in Wi-Fi SDN-Free and SDN-Based Scenarios.

B. Jitter under Wi-Fi without and with SDN

The jitter under Fig. 5 watch the Wi-Fi scenario based on SDN is about 60 ms, it is half of Wi-Fi without SDN; it has the value of 12 ms, which results that the addition of a SDN network for Wi-Fi is fruitful.

C. Latency under Wi-Fi without and with SDN

The results of Fig. 6 shows that the Wi-Fi network latency with SDN is lower (55 ms) than that of the Wi-Fi approach without SDN with a value of 125 ms, which justifies that SDN adds a positive appreciation for Wi-Fi.

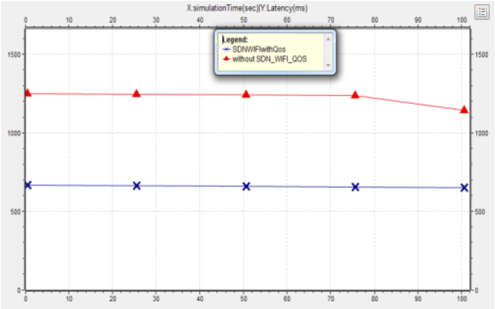


Fig. 6. Latency in Wi-Fi SDN-Free and SDN-Based Scenarios.

D. Packets lost under Wi-Fi without and with SDN.

The number of packets lost in the Wi-Fi approach without SDN is 75%, it is higher compared to the Wi-Fi approach with SDN, which is about 62%. This shows the impact of adding the SDN approach to the Wi-Fi network, as shown in Fig. 7.

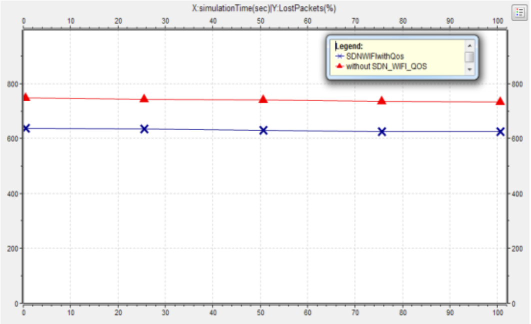


Fig. 7. Figlost Packets in Wi-Fi SDN-Free and SDN-Based Scenarios.

E. MoS under Wi-Fi without and with SDN.

Fig. 8 shows that MOS offered by the Wi-Fi approach without SDN is 2.2 while the approach based on Wi-Fi with SDN is about 3.2; it presents an indicator of increases in quality of voice transmission.

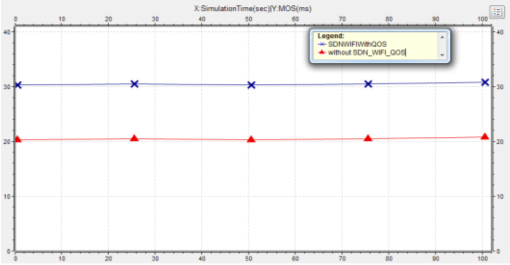


Fig. 8. Fig8. MOS in Wi-Fi SDN-free and SDN-based scenarios.

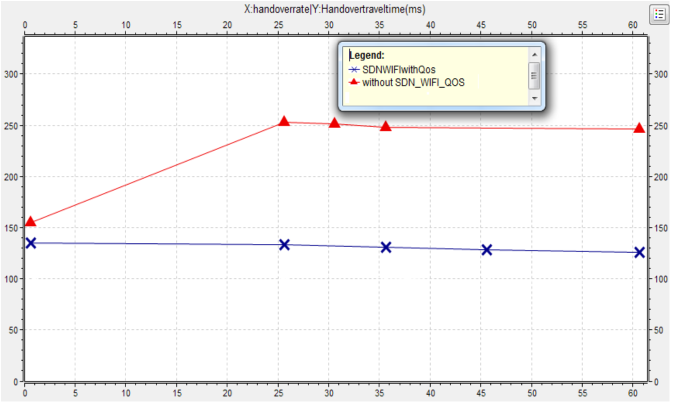


Fig. 9. Handover in Wi-Fi SDN-Free and SDN-Based Scenarios.

VIII. Handover under Wi-Fi without and with SDN

Fig. 9 presents two scenarios to describe node transfer signalling in a Wi-Fi network with SDN, it is stable during all the communication with a minimum of time (13ms), it is inferior to Wi-Fi without SDN (25 ms), so SDN can benefit from the services of another cell instead of the old one. This allows the mobile station the ability to continue its ongoing communication with a minimum of interruption, knowing that the two cells involved are managed by one or more networks.

IX. Conclusion

This article presents the improvement the Handover parameters and the QoS, under OMNeT 4.6++, that it is offered by the WiFi network by adding a new algorithm SDN, for Wi-Fi, which it is more preferable to Wi-Fi without SDN.

This work is considered a part of another article that determines a contribution of 4G with SDN in terms of HO and QoS.

References

[1] Fatima LAASSIRI, Mohamed MOUGHIT, Noureddine IDBOUFKER, ‘‘Improvement of Multiprotocol Label Switching Network's Performance using Software Defined Network Approach’’, IJSER, June -2018.

[2] Jin-Shyan Lee, Yu-Wei Su, Chung-Chou Shen, ‘‘A Comparative Study of Wireless Protocols: Bluetooth, UWB, ZigBee, and Wi-Fi’’, IEEE, 03 March 2008.

[3] Laizhong Cui, F. Richard Yu, and Qiao Yan, When Big Data Meets, ’’Software-Defined Networking: SDN for Big Data and Big Data for SDN’’, IEEE Network. January/February 2016.

[4] Henry Sinnreich, Alan B.Johnston, ‘‘Internet Communications Using SIP, Delivering VOIP and Multimedia Services with Session Initiation Protocol’’, Second Edition, books.google.com, 2012.

[5] A. Patel, K. Leung, M. Khalil, H. Akhtar, K. Chowdhury, RFC 4283, ‘‘Mobile Node Identifier Option for Mobile IPv6 (MIPv6)’’, November 2005.

[6] Jin-Shyan Lee, Yu-Wei Su, Chung-Chou Shen, ‘‘A Comparative Study of Wireless Protocols: Bluetooth, UWB, ZigBee, and Wi-Fi’’, IEEE,03 March 2008.

[7] Shailandra Kaushik, An overview of Technical aspect for WiFi Networks Technology, IJECSE, Volume1, Number 1.

[8] Anita Nikolich, ‘‘SDN Research Challenges and Opportunities’’, CODASPY '16: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, March 2016.

[9] D. Tennenhouse and D. Wetherall, “Towards an active network architecture”, ACM Computer Communication Review, 26(2), pp. 5–17, 1996.

[10] B. Schwartz, A. W. Jackson, W. T. Strayer, W. Zhou, R. D. Rockwell, and C. Partridge, “Smart Packets for active networks”, IEEE Second Conference on Open Architectures and Network Programming Proceedings OPENARCH ’99, IEEE xplore, pp. 90–97, 1999.

[11] Bruno Astuto A. Nunes, Marc Mendonca, Xuan-Nam Nguyen “A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks”, Volume: 16, Issue: 3, IEEE, February 2014.

[12] Alberto Bemporad, Antonio Bicchi, Giorgio Buttazzo, “Hybrid Systems: Computation and Control”, 10th International Workshop, Proceedings, HSCC 2007, Pisa, Italy, April 3-5, 2007.

[13] A. Greenberg, G. Hjalmtysson, D.A. Maltz, A. Myers, J. Rexford, G. Xie, H. Yan, J. Zhan et H. Zhang, “A Clean Slate 4D Approach to Network Control and Management”, ACM SIGCOMM Computer Communication Review, 35(5), pp. 41–54, 2005.

[14] M. Casado, M. J. Freedman, J. Pettit, J. Luo, N. McKeown et S. Shenker, “Ethane: Taking Control of the Enterprise”, Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 1–12, 2007.

[15] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown et S. Shenker, “NOX: Towards an Operating System for Networks”, ACM SIGCOMM Computer Communication Review, 38(3), pp. 105–110, 2008.

[16] Fouad BENAMRANE, Etude des Performances des Architectures du Plan de Contrôle des Réseaux ‘Software-Defined Networks, January 2017.

[17] Rohde-schwar, ‘‘White Paper 802.11ac Technology Introduction’’, page 6 – 2012.

[18] https://fr.wikipedia.org/wiki/IEEE_802.11ac, January 2017.

[19] Doug Marshke, Jeff Doyle and Pete Moyer, “Software Defined Networking (SDN): Anatomy of OpenFlow Volume I’’, “SDN Essentials”, June 2017.

[20] Marcial P Fernandez, “Comparing OpenFlow Controller Paradigms Scalability: Reactive and Proactive’’, IEEE 27th International Conference on Advanced Information Networking and Applications, 2013.

[21] Wendell Odom, SDN Termilogy from layered models, https://www.sdnskills.com/learn/ sdn-terms-01/, May 13, 2015.

[22] Evangelos Haleplidis, Stefano Salsano, “Overview of RFC7426: SDN Layers and Architecture Terminology’’, IEEE Softwarization, September 2017.

[23] Open Networking Foundation, “SDN Architecture Overview’’, Version 1.0 December 12, 2013.

[24] https://www.sdxcentral.com/sdn/definitions/inside-sdn-architecture/, January 2018.

[25] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, “Walker, Public Review for Programming Protocol-Independent Packet Processors’’, ACM SIGCOMM Computer Communication Review, Volume 44, Number 3, July 2014.

[26] Jérôme Durand, ‘‘Le SDN pour les nuls’’, Cisco Systems, p8, JRES 2015.

A Mapping Approach for Fully Virtual Data Integration System Processes

Ali Z. El Qutaany[1]
PhD Student, Faculty of Computers and Information,
Cairo University
Cairo, Egypt

Osman M. Hegazi[2]
Professor, Faculty of Computers and Information,
Cairo University
Cairo, Egypt

Ali H. El Bastawissy[3]
Professor, Faculty of Computer Science,
MSA University
Cairo, Egypt

***Abstract*—Nowadays, organizations cannot satisfy their information needs from one data source. Moreover, multiple data sources across the organization fuels the need for data integration. Data integration system's users pose queries in terms of an integrated schema and expect accurate, unambiguous, and complete answers. So the data integration system is not limited to, getting the answers to the queries from the sources, but also it is extended to detect and resolve the data quality problems appeared due to the integration process. The most crucial component in any data integration system is the mappings constructed between the data sources and the integrated schema. In this paper a new mapping approach is proposed to map not only the elements of the integrated schema as done by the existing approaches, but also it maps other elements required in detecting and resolving the duplicates. It provides a means to facilitate future extensibility and changes to both the sources and the integrated schema. The proposed approach provides a linkage between the fundamental components required to provide accurate and unambiguous answers to the users' queries from the integration system.**

Keywords—Data integration; inconsistency detection; inconsistency resolution; mapping; virtual data integration

I. Introduction

Data integration refers to the problem of combining data residing at autonomous, homogenous/heterogeneous sources, and providing users with a unified global schema [1]. Data integration system I is formalized in terms of a triple (GS, S, M) [2], where; GS is the integrated schema to represent the participating data sources or the data integration requirements based on predetermined business objectives, it is also called mediated schema between the users and the data sources, S is the "data Sources" participating in the integration process, and M is to map GS to S. There are two radically different integration methods: virtualization and materialization. Virtualization leaves the data where it is, as it is, and dynamically retrieves, merges and transforms it on request. Materialization does the integration up front, creating a new dataset of requests to run against. The authors of this research are interested in virtualization. Two main concepts constitute the architecture of a virtual data integration system: wrappers and mediators. Wrapper wraps and models the source using a source schema while the mediator maintains a global schema and mappings between the global and source schemas [3]. Users are posing their queries to the integrated system in terms of the global schema and expecting to receive accurate, complete and unambiguous answers. To ensure users' expectations; the integration system should perform three main processes; Data Integration (DI) process, including getting the raw answers from the sources, Inconsistency Detection (ID) process, and Inconsistency Resolution (IR) process. The three main processes can be detailed as follows.

Data Integration (DI) Process: In this process, the GS is constructed, the S is marked, and M is built. Users pose queries in terms of the GS, and the data integration system converts these queries using M into a set of subqueries over S. Each data source answers the subquery with the help of its wrapper(s). Data sources were created in heterogeneous environments; thus data quality problems [4] appear in the collected answers from the sources. These problems occur because the sources often contain redundant data in different representations. Even if, the sources are clean, accurate and the data representations are unified across all the participating sources; some data quality problems appear due to the integration process. One of these problems is mutual inconsistencies which need efforts to be detected and resolved as functions of the successive processes to the integration process. The collected answers to each user's query should be sent as an input to the inconsistency detection process.

Inconsistency Detection Process: This process is called "Duplicate Records Detection" or "Entity Matching", and due to the duplicates; inconsistencies appear, so this process also called "Inconsistencies Detection". In this process; duplicates are detected [5, 6, 7, 8, 9, 10] in preparation to remove the ambiguities in the generated answers and to fuse inconsistencies before passing the answers to the user. Detected duplicates are marked in the answer set, and passed as an input to the successive process to resolve the inconsistencies.

Inconsistency Resolution Process: In this process; detected inconsistencies are resolved [11, 12, 13, 14, 15, 16, 17, 18, 31] before passing the generated answers to the users. In the literature; there are 3 different strategies [19] to deal with the inconsistencies, some researchers ignore the conflicts resolving process at all, this strategy called "conflict ignorance", others are avoiding [20, 21] dealing with conflicts by defining a pre-determined decision to be taken in case of conflicts called "conflict avoidance", and the rest [22, 23, 24, 25] are trying to resolve the inconsistencies once detected called “conflict resolution”.

Obviously, one of the main tasks in the design of a data integration system is to establish the mapping M between S and GS, such mapping should be suitably taken into consideration in formalizing a data integration system to serve all of its processes not only the DI process. Basically, there are two mapping approaches [1] to define M: Global-as-View (GAV) and Local-as-View (LAV). However, both approaches have their limitations. To overcome these limitations, another mapping approach is introduced to combine the best of GAV and LAV called Both-as-View BAV. Other derivatives of these approaches, such as Global-Local-as-View GLAV, and Both-Global-Local-as-View provide alternatives for more flexible and scalable data integration but still has a set of limitations. GAV, LAV, and BAV have a common limitation, which is; while defining M, they are not considering the data integration successive processes, they only used for the integration and query answering process, and they are also facing a lot of issues when no shared identifier is used for the integrated real world object from different sources. In this paper, a mapping approach is proposed not only to define the mappings between GS and S, but also to prepare parameters assisting in performing the after integration processes; i.e. the inconsistency detection and resolution processes, and provide means to facilitate future changes, extensibility, flexibility, and scalability of the integrated system, and to work with the non-federated and heterogeneous data sources as well as the federated and homogenous ones. The rest of the paper is organized as follows, GAV, LAV, and BAV will be detailed in section II showing their principles, advantages, and limitations, while section III introduces the proposed approach, section and finally section V concludes the work and states the future work.

II. Related Work

One of the most important aspects in the design of a data integration system is the specification of the correspondence between GS and S. It is exactly this correspondence that will determine how the users’ queries posed to the integration system are answered. Three basic approaches for specifying such mapping in a data integration system have been proposed in the literature: LAV, GAV, and BAV. Some derivations are also examined to avoid drawbacks noticed in both GAV and LAV, e.g. BGLAV and GLAV. In this section; the basic approaches are investigated, showing their principles, pros and cons. Then the common limitations faced in the approaches are listed, and a demonstration example is shown to be used throughout the full paper.

A. Mapping Approaches

Global as View (GAV) approach: Mappings in data integration systems based on **GAV** as shown in Figure 1 (a) associates each global relation symbols with views over local relation symbols. In GAV based mapping integration systems, the same GS relation may have more than one mapping assertions over S in case of the unavailability of global relation elements in all data sources. Query processing and simple query reformulation is the most important advantage of GAV. GAV is effective whenever the data integration system is based on a set of stable (do not change too much) sources, but it does not support scalability for the data integration system as changes in GS and/or local schemes derive the designer to revise and alter the mappings. GS in the systems based on GAV approach; can only contain available elements in S at the design time. Finally it does not prepare parameters for the successive data integration processes as it only considers the data integration and query answering process. IBIS [26], Multiplex, Fusionplex and Autoplex [21] are GAV data integration systems examples.

Local as View (LAV) approach: The mapping in data integration systems based on the LAV as shown in figure 1 (b) associates local relation symbols with a view over global relation symbols. LAV approach favors the extensibility of the system where adding a new source simply means enriching the mapping with new assertions, without other changes, so it is effective whenever the data integration system is based on a global schema that is stable and well-established in the organization. But query reformulation has exponential time complexity with respect to query and source schema definitions. GS in the systems based on LAV approach; can only contain available elements in S at the design time. Like GAV; LAV does not prepare parameters for the successive data integration processes as it only considers the data integration and query answering process. Information Manifold [27], System described in [22] are LAV data integration systems examples.

Both-as-View (BAV) approach: BAV as shown in figure 1 (c) is an alternative point of view that is neither GAV nor LAV as it uses source to-target mappings based on a predefined conceptual target schema, which is specified ontologically and independently of any of the sources. In BAV for each pair (vS, vG) incrementally modify vS / vG using primitive schema transformations to match vG /vS. BAV [28] is easier to maintain than both GAV and LAV, and query reformulation reduces to rule unfolding [1]. GS can only contain available elements in the sources at the design time. And like both LAV and GAV; BAV does not prepare parameters for the successive data integration processes as it only considers the data integration and query answering process. Clio [29] is a BAV data integration systems example.

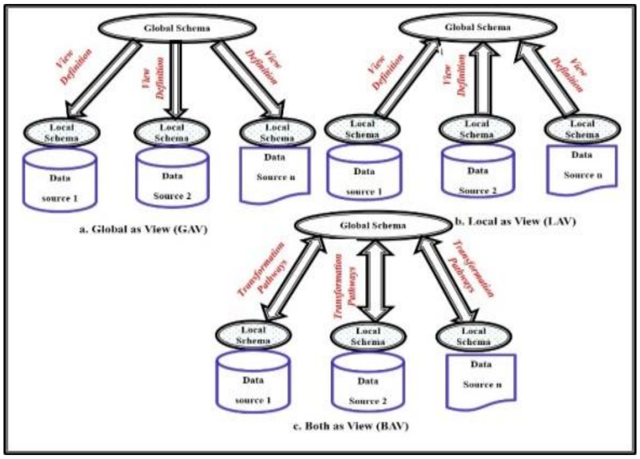


Fig. 1. Mapping Approaches GAV, LAV and BAV.

B. Common Limitations for all Existing Mapping Approaches

These mapping approaches are used to define the mappings between a global schema GS which was designed to integrate data existing in 8 heterogeneous data sources S built under different platforms, these sources use different identification method for the same real world object, i.e. no common identifier for the integrated object from all sources, but some sources may agree on one identifier and others may agree on another identifier. The 8 sources contain data for around 5,000,000 real world objects. In S; the same object may have records in different sources, but each source does not have duplicates for the same object, the GS contains around 80 relations. Attributes within each GS relation are not mapped to all data sources and none of the sources has all attributes of one GS relation. The issues faced during the implementation:

1) None of the mapping approaches, allows the possibility of adding elements to GS for future extensibility of the business objectives, if they are not existing in sources at the design time,

2) As number of the participating information sources increases [30], as the mappings construction, the query answering, and adding new information source or modifying existing one becomes more complicated processes.

3) More than one mapping assertion built for each global schema relation, as not all data sources provide the same attributes and the same number of attributes for the global schema relation.

4) None of the mapping approaches considering the data integration successive processes. As they do not consider mapping the parameters which may help in the detection and resolution processes, e.g. source qualifications.

5) When two of the participating information sources share an identifier for the real world object; then some duplicates are prevented by the mapping assertions definition. In this case, the accurate and most recent information is **not** always presented in the chosen source of providing information in the mappings.

6) Changes in the data sources and/or the global schema require extensive efforts to keep the mappings consistent.

These limitations in the existing mapping approaches become challenges for the proposed approach.

C. Demonstration Example

To explain and approve the limitations and drawbacks of the existing approaches, let's start by hypothetically demonstrate the following schemes, later the same hypothesis will be also used to highlight the advantages of the proposed mapping approach.

Example 1. Suppose we have 5 data sources, representing oil and gas wells data with their semantics and a global schema which is designed and uses notations and naming independently from the sources.

GS: Well (WellAPI, WellName, Latitude, Longitude, FieldName, County, CompIntervalID) – GS is designed to integrate USA wells.

Data sources S – heterogeneous data sources; as the real world object (Well Object) represented in the 5 sources does not has the same identification key across all sources, sources are partially agree on the well object identifier.

S1: WellDetails (WellAPI, WellName, Latitude, Longitude, MeasuredDepthFt, HorizontalWell, Country) – Contains data about wells from different countries.

S2: Well (APINo, WellName1, Lat, Long) – Contains data about wells from GOM (Gulf of Mexico).

S3: USAWellData (WellName, WellSuffix, Latitude, Longitude, FieldName, County, WellMD, HWFlag) – Contains data about wells from USA.

S4: GulfArabiaOilWells (Name, TopLatitude, TopLongitude, PrimaryField, Country, MD, HWFlag) – contains data about wells from Gulf of Arabia countries. This is irrelevant source to the integration objective.

S5: NorthDakotaWells (Name, APICompSTR, SurfaceLat, SurfaceLong, FieldLocation, Field) – contains data about wells from only North Dakota state (USA).

Detectors (this term will be explained and used while exploring the proposed approach, these detectors can be automatically detected or defined by domain experts. Here in this paper, they defined by domain expert): Detectors for the well object in S1 are {WellAPI} and {Latitude, Longitude, MeasuredDepthFt, HorizontalWell}, S2 uses {APINo}, S3 uses {Latitude, Longitude, WellMD, HWFlag}, and finally S5 uses {Substring (APICompSTR, 0, CharIndex (' ')-1)} as WellAPINumber. WellAPI from S1 is equivalent to both APINo from S2 and Substring (APICompSTR, 0, CharIndex (' ')-1) from S5. WellMD from S3 and MeasuredDepthFt from S1 are equivalent and HorizontalWell from S1 and HWFlag from S3 are equivalent.

III. Proposed Mapping Approach

Not all of the participating sources in the data integration process are federated as they do not use the same identifier for the real world object. A new term called *detector* is invented to be used in this case. *Detector* is an identifier for the real world object in its origin and it may **not** be shared between all the sources mapped to the GS relation Ri. Real world objects indicated in some sources may agree on a set of detectors while others may agree on another set. Detectors may be one or many for the real world object in its data source. A detector may be single or composite. As Ri will be mapped to data coming from different sources, so the union of these detectors constructs the detectors of Ri although the attributes of these detectors may not be appearing in Ri, by default if the sources are sharing the same identifier then the detector of Ri will be the shared identifier between all sources. One detector or many may be existing per the GS relation. None of the detectors can be considered as an identifier for the GS relation as it will contain nulls for the records extracted from the source(s) which do not agree on these detectors and then violates the entity integrity constraints. Each detector identifies only the objects which extracted from the sources agreed on such detector(s). All the detectors will be used in the duplicate detection process in a hierarchy based, by starting with the first detector and ending with the last detector. Duplicate record detection is out of scope in this paper, but mapping of the detectors for each GS relation is considered. Figure 2 shows how detectors are collected from the sources and processed to construct the GS relation Ri detectors. In example 1 there is no unified identifier for the **well** object in all the data sources, so each data source is required to provide its detectors for the well object as shown in the example. The union of these detectors will construct the detectors of the GS relation. As in the example; S1, S2, and S5 agreed on the detector {WellAPI} and S1, and S3 agreed on the detector {Latitude, Longitude, WellMD, HWFlag}. The GS relation **Well** have two detectors {WellAPI} and {Latitude, Longitude, WellMD, HWFlag}, these detectors will be used during the duplicate record detection process. Inconsistency resolution is required before passing the results to the user and after the duplicate record detection process. Source preference [22] is one of the fusion policies known in the inconsistency resolution, which fuse the conflicting data based on the preferred source, but to apply such policy, you should have the source name in the result set passed to the inconsistency resolution process. In order to accomplish this; source name will be considered in the mapping construction process with the detector sets even if they are not considered in the GS design. The source qualifications, e.g. Timestamp, Cost, Availability… used for the inconsistency resolution process may also be extracted and mapped during the mapping construction process. Here a mapping approach is proposed which is unlike all of the existing mapping approaches, it does not assume the homogeneity between all of the participating data sources, as it works for federated and non-federated data sources. The proposed approach provides means to facilitate the process of defragmenting the results from the data sources, add a new data source(s), remove an existing data source(s), and modifying data source(s). The detectors and source name element defined in this mapping approach may not be part of the elements required in the GS relations for business objectives, but they will be mapped only for performing the data integration successive processes; entity matching and resolution.

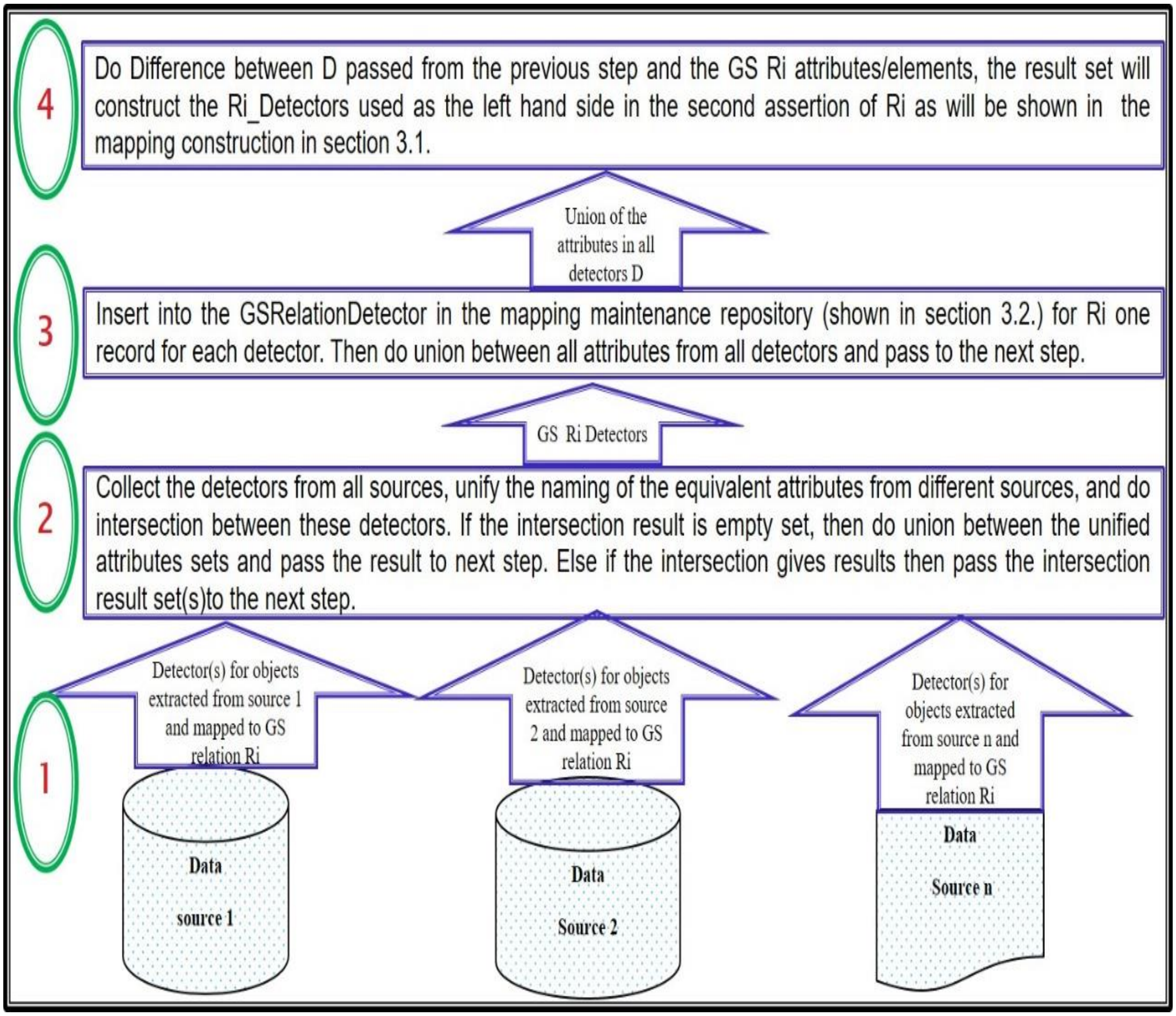

4
Do Difference between D passed from the previous step and the GS Ri attributes/elements, the result set will construct the Ri_Detectors used as the left hand side in the second assertion of Ri as will be shown in the mapping construction in section 3.1.
Union of the attributes in all detectors D
3
Insert into the GSRelationDetector in the mapping maintenance repository (shown in section 3.2.) for Ri one record for each detector. Then do union between all attributes from all detectors and pass to the next step.
GS Ri Detectors
2
Collect the detectors from all sources, unify the naming of the equivalent attributes from different sources, and do intersection between these detectors. If the intersection result is empty set, then do union between the unified attributes sets and pass the result to next step. Else if the intersection gives results then pass the intersection result set(s)to the next step.
1
Detector(s) for objects extracted from source 1 and mapped to GS relation Ri
Detector(s) for objects extracted from source 2 and mapped to GS relation Ri
Detector(s) for objects extracted from source n and mapped to GS relation Ri
Data source 1
Data Source 2
Data Source n

Fig. 2. Construction Process for Detectors of GS Relation Ri.

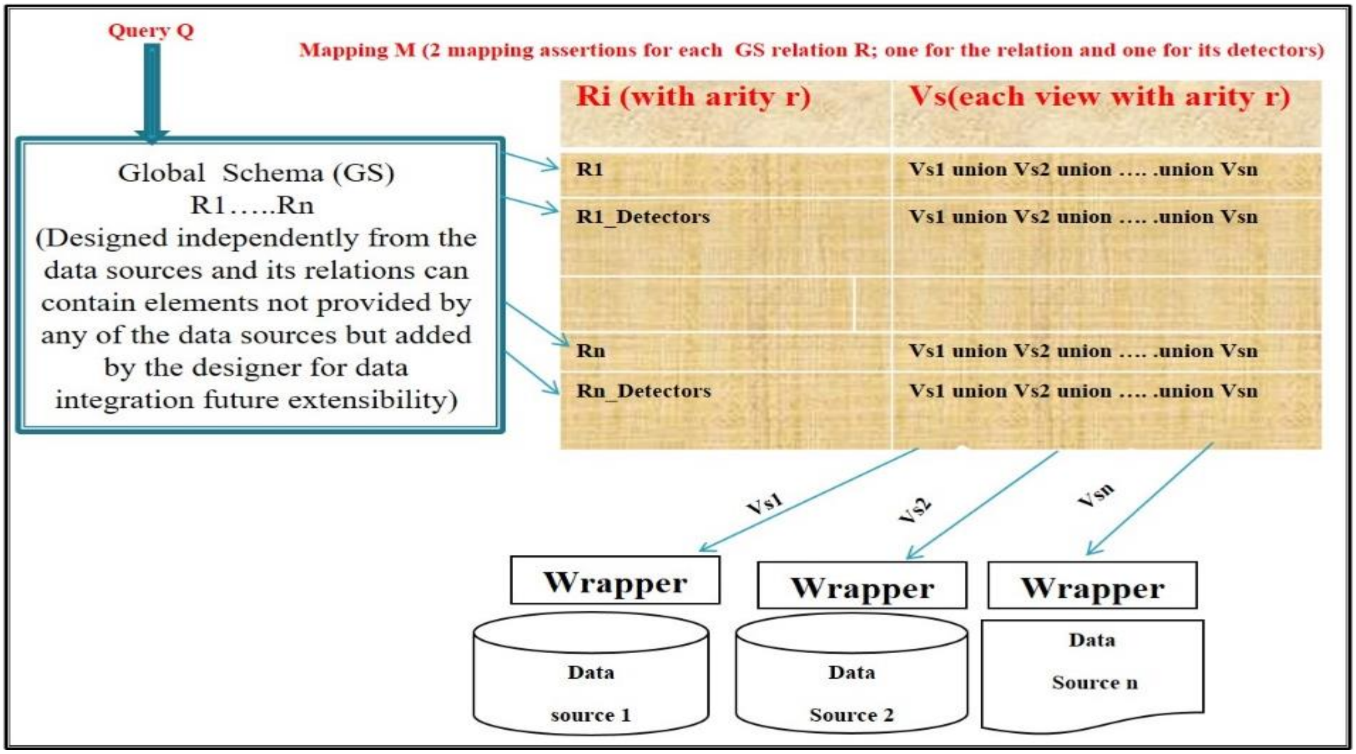

Query Q
Mapping M (2 mapping assertions for each GS relation R; one for the relation and one for its detectors)
Ri (with arity r)
Vs(each view with arity r)
R1
Vs1 union Vs2 union …. .union Vsn
R1_Detectors
Vs1 union Vs2 union …. .union Vsn
Rn
Vs1 union Vs2 union …. .union Vsn
Rn_Detectors
Vs1 union Vs2 union …. .union Vsn
Global Schema (GS)
R1…..Rn
(Designed independently from the data sources and its relations can contain elements not provided by any of the data sources but added by the designer for data integration future extensibility)
Vs1
Vs2
Vsn
Wrapper
Wrapper
Wrapper
Data source 1
Data Source 2
Data Source n

Fig. 3. The Mapping Assertions Construction Process using the Proposed Approach.

A. *Principles of the Proposed Approach*

1) GS designed independently of the sources, and can contain relations and elements which may not be present in the available sources, but added for future scalability and extensibility of the integration system objectives.

2) The mappings between GS constructs and S constructs are built as shown in figure 3, where each GS relation Ri has two assertions; one assertion to map Ri elements in the form Ri ➔ views Vs over all data sources linked by union, such that a single view per each source appears in the union of local views to map such source to Ri. View V over source s has the same arity as Ri, such that each attribute appeared on Ri and does not have correspondence with attribute from s is replaced with **Null** and aliased with the corresponding attribute from Ri to facilitate modifying of both the data sources and the GS relations. The second assertion will be constructed to map the Ri detectors and the source name element with the sources participating in Ri mapping assertion, even if they are not present in the GS for business objectives.

First mapping assertion for the GS relation will be used for the traditional query answering, and the second mapping assertion is used for the successive data integration processes.

Appearance of a specific data source in the mapping assertions follows a specific ordering, where the ordering of the view vS over sourcei is predetermined and stored in MappingHelper table in a standalone repository, shown in the next subsection B. This repository will aid in adding or removing data source (s).

Users pose their queries in terms of the GS relations.

A query Q on the global relations must be translated to a set of subqueries over the data sources.

As an example; the mapping assertions for the GS relation **WELL** in example 1 with the sources will look like:

Assertion-1: Well (WellAPI, WellName, Latitude, Longitude, FieldName, County, CompIntervalID) ➔ Select WellAPI, WellName, Latitude, Longitude, Null as FieldName, Null as County, Null as CompIntervalID from S1.WellDetails Where Country = ‘USA’ **Union** Select APINo, WellName1, Lat, Long, Null as FieldName, Null as County, Null as CompIntervalID from S2.Well **Union** Select Substring (APICompSTR, 0, CHARINDEX (APICompSTR, ‘ ’)-1) as WellAPI, Name, SurfaceLat, SurfaceLong, FieldLocation+ ‘-’+ Field as FieldName, Null as County, Substring (APICompSTR, CHARINDEX (APICompSTR, ‘ ’)+1, length (APICompSTR)-1) as CompIntervalID from S5.NorthDakotaWells **Union** Select Null as WellAPI, WellSuffix +’ –‘+ WellName as WellName , Latitude, Longitude, FieldName, County, Null as CompIntervalID from S3.USAWellData

Assertion-2: Well_Detectors (WellMD, HorizontalWellFlag, SourceName)➔ Select MeasuredDepthFt, HorizontalWell, ‘S1’ as SourceName From S1.WellDetails Where Country = ‘USA’ **Union** Select Null as WellMD, Null as HorizontalWellFlag, ‘S2’ as SourceName from S2.Well **Union** Select Null as WellMD, Null as HorizontalWellFlag, ‘S5’ as SourceName from S5.NorthDakotaWells **Union** Select WellMD, HWFlag, ‘S3’ as SourceName from S3.USAWellData.

Well_Detectors only contains three attributes and it was supposed to contain 6 attributes; 5 for detectors (WellAPI Latitude, Longitude, WellMD, and HorizontalWellFlag) and another attribute for SourceName. But as the Well relation contains 3 attributes from these 6, so the difference process between Well_Detectors and Well gives {WellMD, HorizontalWellFlag, SourceName} which is used in the **Well_Detectors**. In the first assertion (Well Assertion), all the local attributes may be aliased with the GS corresponding attributes names, even if they provided from the sources, to facilitate the process of query answering afterwards. In example 1 if the designer needs to add two elements WellType and WellStatus to the GS relation Well, at the design time, although they do not have correspondence with any of the data sources, this is possible in the proposed approach; it becomes as easy as; just adding them to the GS relation, modifying the mapping assertion Well, and enriching each view over the sources with two elements Null as WellType, Null as WellStatus.

B. Mapping Maintenance Helper Repository

This repository contains 2 tables; one called GSRelationDetector, and it has the detectors of each GS relation, it takes the form GSRelationDetector (GSRelationName, Detector), and it is used to help in the query answering to prepare the answer for the duplicate record detection and resolution processes. The second called MappingHelper and it takes the form MappingHelper (GSRelationName, SourceName, SourceIndex), where the SourceIndex is the order of this data source's view within the mapping assertion for the corresponding GS relation.

Mapping Helper table helps in adding, removing, modifying data sources and/or GS relations.

TABLE I. MAPPINGHELPER FOR EXAMPLE 1

GSRelation	**Properties**	
	SourceName	***SourceIndex***
Well	S1	1
Well	S2	2
Well	S3	4
Well	S	3

TABLE II. GSRELATIONDETECTOR FOR EXAMPLE 1

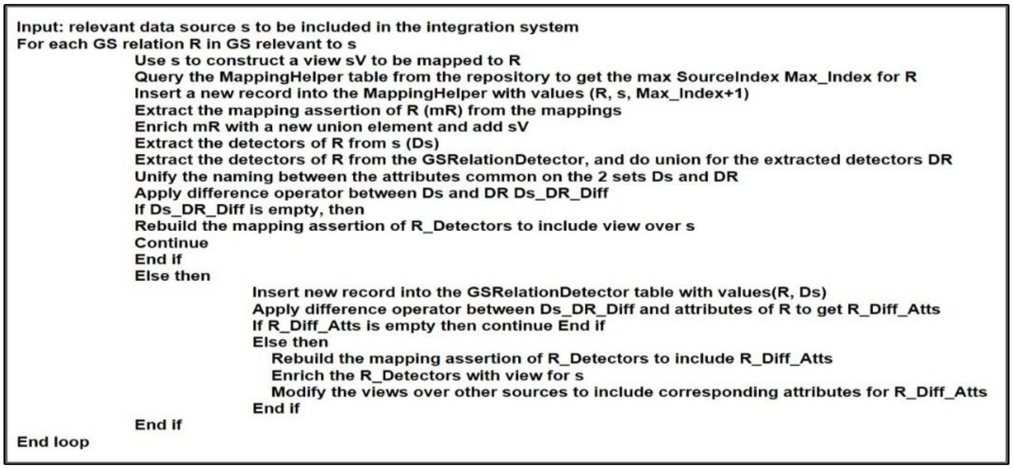
GSRelation	**DetectorSet**	
	Detector	***Index***
Well	WellAPINumber	1
Well	Latitude, Longitude, WellMd, HorizontalWellFlag	2

In example 1, the MappingHelper table takes the form shown in table 1. These ordering was used in the previous section A. To build the mappings. And table 2 shows the GSRelationDetector for example 1.

C. Data Sources Management in the Proposed Approach

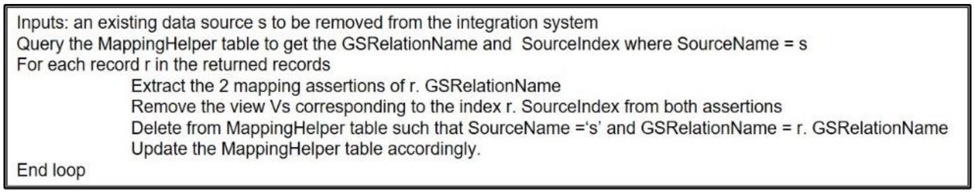
In this section, the operations applied in the data sources are shown, such operations are:

1) Addition and removal of a data source: One of the features in the proposed mapping approach is the way of adding and removing a data source(s) to and from the integration system. Figure 4 shows an algorithm for the addition process, and Figure 5 shows an algorithm to be used to remove a data source. The same 2 algorithms can be used when adding a relation to a data source or removing a relation from a data source.



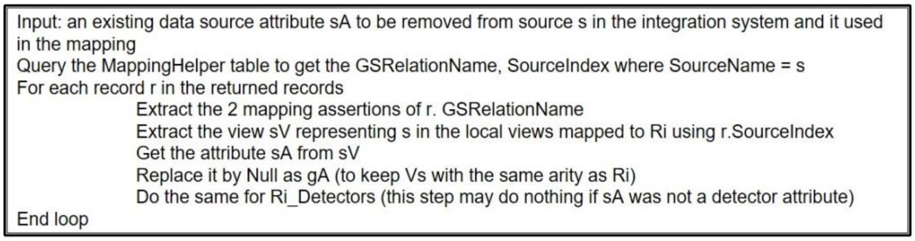
```
Input: relevant data source s to be included in the integration system
For each GS relation R in GS relevant to s
        Use s to construct a view sV to be mapped to R
        Query the MappingHelper table from the repository to get the max SourceIndex Max_Index for R
        Insert a new record into the MappingHelper with values (R, s, Max_Index+1)
        Extract the mapping assertion of R (mR) from the mappings
        Enrich mR with a new union element and add sV
        Extract the detectors of R from s (Ds)
        Extract the detectors of R from the GSRelationDetector, and do union for the extracted detectors DR
        Unify the naming between the attributes common on the 2 sets Ds and DR
        Apply difference operator between Ds and DR Ds_DR_Diff
        If Ds_DR_Diff is empty, then
        Rebuild the mapping assertion of R_Detectors to include view over s
        Continue
        End if
        Else then
                Insert new record into the GSRelationDetector table with values(R, Ds)
                Apply difference operator between Ds_DR_Diff and attributes of R to get R_Diff_Atts
                If R_Diff_Atts is empty then continue End if
                Else then
                  Rebuild the mapping assertion of R_Detectors to include R_Diff_Atts
                  Enrich the R_Detectors with view for s
                  Modify the views over other sources to include corresponding attributes for R_Diff_Atts
                End if
        End if
End loop
```

Fig. 4. Algorithm for Addition of a New Data Source to the Integration System using the Proposed Approach.



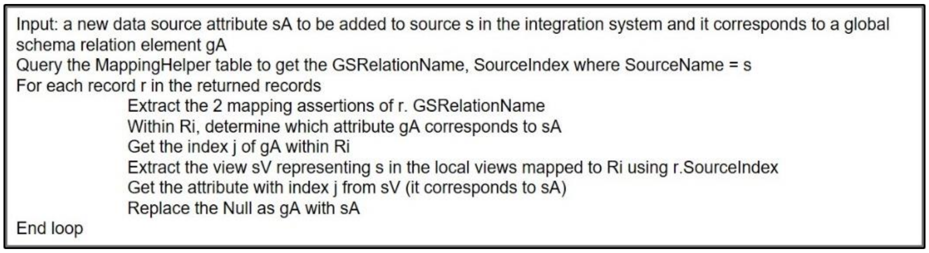
```
Inputs: an existing data source s to be removed from the integration system
Query the MappingHelper table to get the GSRelationName and SourceIndex where SourceName = s
For each record r in the returned records
        Extract the 2 mapping assertions of r. GSRelationName
        Remove the view Vs corresponding to the index r. SourceIndex from both assertions
        Delete from MappingHelper table such that SourceName ='s' and GSRelationName = r. GSRelationName
        Update the MappingHelper table accordingly.
End loop
```

Fig. 5. Algorithm for Removal of an Existing Data Source using the Proposed Approach.



```
Input: an existing data source attribute sA to be removed from source s in the integration system and it used
in the mapping
Query the MappingHelper table to get the GSRelationName, SourceIndex where SourceName = s
For each record r in the returned records
            Extract the 2 mapping assertions of r. GSRelationName
            Extract the view sV representing s in the local views mapped to Ri using r.SourceIndex
            Get the attribute sA from sV
            Replace it by Null as gA (to keep Vs with the same arity as Ri)
            Do the same for Ri_Detectors (this step may do nothing if sA was not a detector attribute)
End loop
```

Fig. 6. Algorithm for Addition of a New Attribute to a Data Source.



```
Input: a new data source attribute sA to be added to source s in the integration system and it corresponds to a global
schema relation element gA
Query the MappingHelper table to get the GSRelationName, SourceIndex where SourceName = s
For each record r in the returned records
            Extract the 2 mapping assertions of r. GSRelationName
            Within Ri, determine which attribute gA corresponds to sA
            Get the index j of gA within Ri
            Extract the view sV representing s in the local views mapped to Ri using r.SourceIndex
            Get the attribute with index j from sV (it corresponds to sA)
            Replace the Null as gA with sA
End loop
```

Fig. 7. Algorithm for Removal of an Attribute from a Data Source.

2) Addition/removal of an element to a data source: In the proposed mapping approach; the views built over the data sources have the same arity as the GS relation in the mapping assertion. Thus adding and removing attributes to and from a data source become an easy process. Figure 6 presents an algorithm to remove an old attribute from a data source and Figure 7 presents an algorithm to show how a new attribute can be added to a data source. In Figure 6, if sA is a detector and does not exist in Ri and Ri_Detectors, it will be added to the Ri_Detectors as the last element, and added to the view representing s in the detectors assertion and finally add a new field to the other sources' views to represent this attribute, this new field will take the form *Null as* gA, where gA is the GS relation element corresponding to sA.

D. GS Management in the Proposed Approach

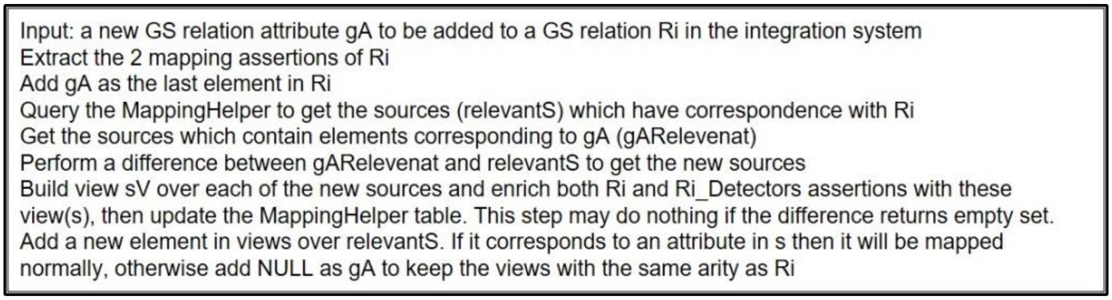
In this section the operations done over the GS are detailed, such operations are

1) Addition and removal of a GS relation: To remove a GS relation Ri from an integration system; search in the mapping assertions for Ri and Ri_Detectors and remove them. If a new GS relation Ri needed to be added to the GS:

a) Construct the views sV over the relevant sources, perform a union over all the constructed views, fill in the MappingHelper table with the order of the sources appearing in the union, and build the mapping assertion.

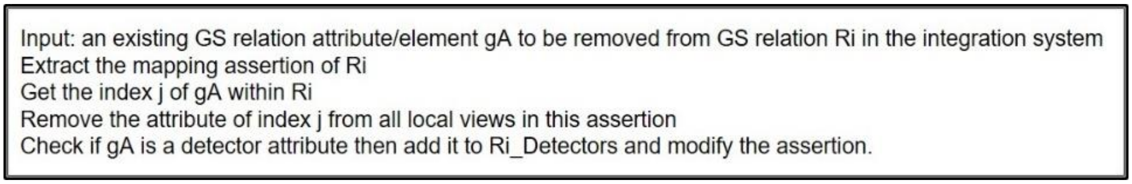
b) Collect the detectors as shown in Figure 2, fill in the GSRelationDetector with Ri detectors and construct another mapping assertion for Ri_Detectors.

2) Addition and removal of an element with a GS relation: Figure 8 presents an algorithm to add a new attribute gA to a GS relation Ri, while Figure 9 presents an algorithm to remove an attribute gA from a GS relation R.



```
Input: a new GS relation attribute gA to be added to a GS relation Ri in the integration system
Extract the 2 mapping assertions of Ri
Add gA as the last element in Ri
Query the MappingHelper to get the sources (relevantS) which have correspondence with Ri
Get the sources which contain elements corresponding to gA (gARelevenat)
Perform a difference between gARelevenat and relevantS to get the new sources
Build view sV over each of the new sources and enrich both Ri and Ri_Detectors assertions with these
view(s), then update the MappingHelper table. This step may do nothing if the difference returns empty set.
Add a new element in views over relevantS. If it corresponds to an attribute in s then it will be mapped
normally, otherwise add NULL as gA to keep the views with the same arity as Ri
```

Fig. 8. Algorithm for Addition of an Attribute gA to a GS relation Ri.



```
Input: an existing GS relation attribute/element gA to be removed from GS relation Ri in the integration system
Extract the mapping assertion of Ri
Get the index j of gA within Ri
Remove the attribute of index j from all local views in this assertion
Check if gA is a detector attribute then add it to Ri_Detectors and modify the assertion.
```

Fig. 9. Algorithm for Removal of an Attribute gA from a GS Relation Ri.

E. Query Answering in the Proposed Approach

Figure 10 shows the query answering in the proposed approach. A query Q is answered as follows:

1) The query Q is parsed against the GS relations.

2) The queried GS relations are extracted using the query reformulation and unfolding module and asks the mapping helper repository for the detectors of the queried relations. The query Q is reformulated to add the detectors (if they are not in Q), the source name, and the filtering attributes, ask for the 2 mapping assertions for each queried relation which serve the query elements, merge the elements/attributes of the 2 assertions of each GS relation and at the end replace each GS relation with its corresponding merged assertion to construct Q*.

3) The reformulated query Q* is passed to the query translator to prepare a subquery for each data source. The subqueries prepared for the sources are adjusted to include the filtering attributes of Q, such that any of the filtering attributes corresponds to Null value in the view is removed from the filtering clause of the subquery, and if the filtering attribute is one and corresponds to null in any of the source views or have “and” condition with any of the other attributes, then this means the subquery will not return any answers from the source, so it will not be sent to the source from the beginning. This serves as a huge optimization since a whole data source will not be visited in this case.

4) The answers are collected from the sources.

The answers of Q* are sent to the duplicate detection process, to detect the duplicates using the detectors, send the answers with detected duplicates to duplicate resolution to resolve the conflicts, and finally project over the original query attributes to be sent to the user as the final query answer. As an example, using example 1 and the mapping assertions defined in 3.1. If a user poses a query Q *(Select WellAPI, WellName, Latitude from Well where FieldName = ‘CHARLES KRAMER 1608’)*, this Q will be answered as follows:

a) The query is parsed against GS.

b) The query reformulation and unfolding module extracts the queried relation(s) from Q, in this case it will be the relation **Well**. It then asks the mapping maintenance helper repository for the detectors of the Well GS relation and the source name attribute, it will be WellAPI, Latitude, Longitude, WellMD, and HorizontalWellFlag, reformulates Q to be (Select WellAPI, WellName, Latitude, Longitude, WellMD, HorizontalWellFlag, FieldName SourceName from Well where FieldName = ‘CHARLES KRAMER 1608’) after union the query projection part, the query selection part, the detectors, and the SourceName attribute. The query reformulation and unfolding module asks for the mapping assertions of the relation **Well**, this will result in the 2 mapping assertions in section A.

c) The query reformulation and unfolding module merges the 2 mapping assertions to be one assertion to serve Q elements, the merged assertion looks like: Well (WellAPI, WellName, Latitude, Longitude, WellMD, HorizontalWellFlag, FieldName, SourceName) ➔ Select WellAPI, WellName, Latitude, Longitude, MeasuredDepthFt, HorizontalWell, Null as FieldName , ‘S1’ as SourceName From S1.WellDetails Where Country = ‘USA’ **Union** Select APINo, WellName1, Lat, Long, Null as WellMD, Null as HorizontalWellFlag, Null as FieldName , ‘S2’ as SourceName From S2.Well **Union** Select Substring (APICompSTR, 0, CHARINDEX (APICompSTR ,‘ ‘)-1) as WellAPINumber, Name, SurfaceLat, SurfaceLong, , Null as WellMD, Null as HorizontalWellFlag, FieldLocation+ ‘-’+ Field as FieldName, ‘S5’ as SourceName From S5.NorthDakotaWells **Union** Select Null as WellAPI, WellSuffix +’ –‘+ WellName as WellName , Latitude, Longitude, WellMD, HWFlag, FieldName, ‘S3’ as SourceName From S3.USAWellData. Finally, replace **Well** by the new merged assertion to construct Q* and pass Q* to the query translator.

d) The query translator translates the Q* into a set of sub-queries for the data sources, so query on S1 will be Select WellAPI, WellName, Latitude, Longitude, MeasuredDepthFt, HorizontalWell, Null as FieldName , ‘S1’ as SourceName from S1.WellDetails Where Country = ‘USA’ and FieldName = ‘CHARLES KRAMER 1608’. Query on S2 will be Select APINo, WellName1, Lat, Long, Null as WellMD, Null as HorizontalWellFlag, Null as FieldName, ‘S2’ as SourceName from S2.Well where FieldName = ‘CHARLES KRAMER 1608’. Query on S5 will be Select Substring (APICompSTR, 0, CHARINDEX (APICompSTR,‘ ‘) -1) as WellAPI, Name, SurfaceLat, SurfaceLong, Null as WellMD, Null as HorizontalWellFlag, FieldLocation+ ‘-’+ Field as FieldName, ‘S5’ as SourceName from—S5.NorthDakotaWells. where FieldName = ‘CHARLES KRAMER 1608’ Query on S3 will be Select Null as WellAPI, WellSuffix +’ –‘+ WellName as WellName , Latitude, Longitude, WellMD, HWFlag, FieldName, ‘S3’ as SourceName from—S3.USAWellData where FieldName = ‘CHARLES KRAMER 1608’

e) Subqueries over S1and S2 will **not** be sent to the sources as they will not retrieve answers, while sub-queries over S3 and S5 will be sent.

f) The query translator will collect the answers from the sources, do simple union between the collected answers, as all subqueries are with the same arity and the columns headers of the query result will be using the GS corresponding headers. The answers sent to the duplicate detection and resolution modules.

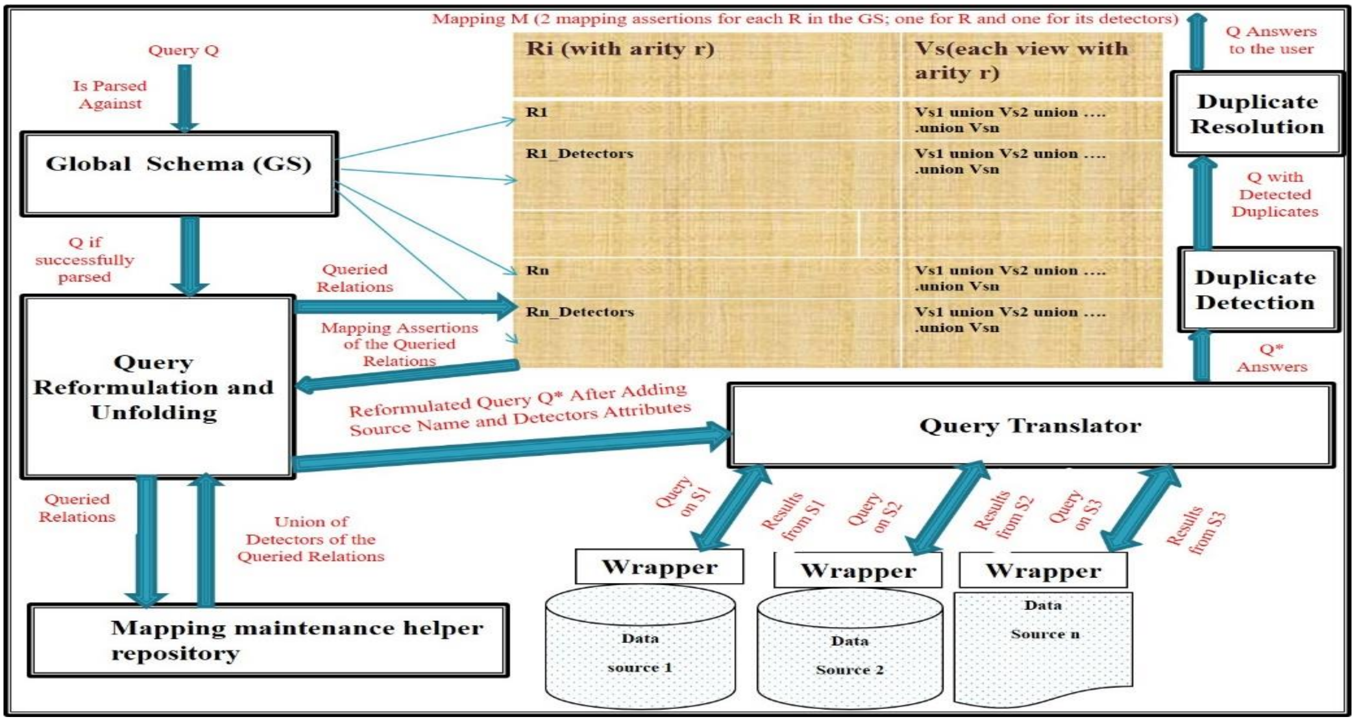

Mapping M (2 mapping assertions for each R in the GS; one for R and one for its detectors)
Query Q
Is Parsed Against
Global Schema (GS)
Ri (with arity r)
Vs(each view with arity r)
R1
Vs1 union Vs2 union …. .union Vsn
R1_Detectors
Vs1 union Vs2 union …. .union Vsn
Rn
Vs1 union Vs2 union …. .union Vsn
Rn_Detectors
Vs1 union Vs2 union …. .union Vsn
Q Answers to the user
Duplicate Resolution
Q with Detected Duplicates
Duplicate Detection
Q* Answers
Q if successfully parsed
Queried Relations
Mapping Assertions of the Queried Relations
Query Reformulation and Unfolding
Reformulated Query Q* After Adding Source Name and Detectors Attributes
Query Translator
Query on S1
Results from S1
Query on S2
Results from S2
Query on S3
Results from S3
Wrapper
Wrapper
Wrapper
Data source 1
Data Source 2
Data Source n
Queried Relations
Union of Detectors of the Queried Relations
Mapping maintenance helper repository

Fig. 10. Query Answering in the Proposed Mapping Approach.

IV. Comparison between the Proposed and the Existing Approaches

In this section, a comparison is performed between the proposed mapping approach and the existing ones, through showing how all the operations are done using the different approaches.

A. *Mapping Assertions Construction*

When using the mapping approaches to define the mapping assertions between the GS relation and the data sources shown in example 1; the mapping assertions using the existing mapping approaches will be done as follows:

1) Mapping assertions in GAV: GAV produces 4 assertions, for example. 1 as below

a) Well (WellName, Latitude, Longitude) ➔ Select WellName, Latitude, Longitude from S1.WellDetails Where Country = 'USA' **Union** Select WellName1, Lat, Long from S2.Well **Union** Select, Name, SurfaceLat, SurfaceLong from S5. NorthDakotaWells

b) Well (WellAPI) ➔ Select WellAPI from S1.WellDetails Where Country = 'USA' Union Select Substring (APICompSTR, 0, CHARINDEX (APICompSTR ,' ')-1) As WellAPI from S5.NorthDakotaWells Where WellAPI not in (Select WellAPI from S1.WellDetails) Union Select APINo from S2.Well Where WellAPI not in (Select WellAPI from S1.WellDetails union Select WellAPI from S5.NorthDakotaWells)

c) Well (FieldName, County) ➔ Select FieldLocation +'-'+Field As FieldName, 'North Dakota' as County from S5.NorthDakotaWells union Select FieldName, County from S3.USAWellData

d) Well (CompIntervalID) ➔ Select Substring (APICompSTR, CHARINDEX (APICompSTR ,' ')+1, Length (APICompSTR) As CompIntervalID from S5.NorthDakotaWells

2) Drawbacks of GAV assertions: drawbacks noticed when using GAV in example 1

a) When a shared identifier found as mapping assertion 2, incomplete answer may be generated due to the mapping assertion construction as when the real world object extracted from S1, it will not be extracted from the other sources S2 and S5. This means the most recent data, the complete data, and/or the accurate data may not be extracted for the resolution process.

b) GAV cannot add elements which have not previously existed in the sources to the GS relation.

c) More than one mapping assertions represent the same GS relation.

d) The source name is not considered in the mappings as it does not appear in the GS relation Well, so the successive processes will not have enough parameters to be done effectively. As the source preferences will not be performed, and the duplicates will be checked between all records, including the ones coming from the same data source, even if there are no duplicates in the data coming from the same data source which is time consuming.

e) In case of no shared identifier defined commonly between all sources for the same real world object, shared identifier may be available partially between some sources, but these identifiers will not be mapped if they are not requested in the GS relations, so they will not be leveraged in the successive processes.

3) Mapping assertions in LAV: LAV produces 4 assertions for, example.1 as below

a) S1.WellDetails (WellAPI, WellName, Latitude, Longitude, Country)➔Select WellAPI, WellName, Latitude, Longitude, ‘USA’ as Country from GS.Well

b) S2.Well (APINo, WellName1, Lat, Long) ➔ Select WellAPI, WellName, Latitude, Longitude from GS.Well where Latitude between 26.01614 and 30.23556 and longitude between - 97.14265 and -86.594202.

c) S3.USAWellData (WellName, WellSuffix, Latitude, Longitude, FieldName, County)➔ Select Substring (WellName, CHARINDEX (WellName,‘-’) + 1, Length (WellName)) as WellName, Substring (WellName, 0, CHARINDEX (WellName ,‘-’)-1) as WellSuffix, Latitude, Longitude, FieldName, County from GS. Well

d) S5.NorthDakotaWells (Name, APICompSTR, SurfaceLat, SurfaceLong, FieldLocation, Field)➔ Select WellName, WellAPI + ‘ ‘+ CompIntervalID as APICompSTR ,Latitude, Longitude, Substring (FieldName, 0 , CHARINDEX (FieldName ,‘-’)-1) as FieldLocation, Substring (FieldName, CHARINDEX (FieldName ,‘-’)+ 1, Length (FieldName)) as Field, from GS. Well where County = ‘North Dakota’

4) Drawbacks of LAV assertions: drawbacks noticed when using LAV in example 1

a) Source views may be mapped to the complete set of objects of the GS view; e.g. assertion1 associates all records of GS relation Well to S1.WellDetails. And this causes, complexities in the query translations and answering.

b) Requires extra information about the source semantics, e.g. mapping assertion 2 requires information about how we can determine the GOM wells, the min/max of latitude and longitude for GOM wells which stored in S2.

c) Drawbacks 2, 3, 4, and 5 noticed in GAV are also noticed here in LAV.

Mapping assertions in BAV: BAV produces 4 assertions for example.1 as below

d) Select WellAPI, WellName, Latitude, Longitude from GS.Well ➔ Select WellAPI, WellName, Latitude, Longitude from S1.WellDetails Where Country = ‘USA’

e) Select WellAPI, WellName, Latitude, Longitude from GS.Well where Latitude between 26.01614 and 30.235566 and longitude between - 97.14265 and -86.59420 ➔Select APINo, WellName1, Lat, Long from S2.Well.

f) Select WellName, Substring (WellName, 0, CHARINDEX (WellName ,‘-’)-1) as WellSuffix, Substring (WellName, CHARINDEX (WellName ,‘-’) + 1, Length (WellName)) as WellName, Latitude, Longitude, FieldName, County from GS. Well➔Select (WellSuffix +’-‘+ WellName) As WellName, WellSuffix, WellName, Latitude, Longitude, FieldName, County from S3.USAWellData

g) Select WellAPI, WellName, Latitude, Longitude, Substring (FieldName, 0 , CHARINDEX (FieldName ,‘-’)-1) As FieldLocation, Substring (FieldName, CHARINDEX (FieldName ,‘-’)+ 1, Length (FieldName)) as Field, FieldName, CompIntervalID, WellAPINumber + ‘ ‘+ CompIntervalID as APICompSTR from GS. Well Where County = ‘North Dakota’ ➔ Select Substring (APICompSTR, 0, CHARINDEX (APICompSTR ,‘ ’)-1) as WellAPI, Name, SurfaceLat, SurfaceLong, FieldLocation , field, (FieldLocation+ ‘-’+ Field) as FieldName, Substring (APICompSTR , CharIndex (APICompSTR ,‘ ’)+1, Length (APICompSTR)) as CompIntervalID, APICompSTR from S5.NorthDakotaWells

5) Drawbacks of BAV assertions: drawbacks noticed when using BAV in example 1 are

a) Drawbacks 2, 3, 4, and 5 noticed in GAV are noticed also here in BAV.

b) Needs extra efforts and time to keep matching between the local and global views.

B. Data Sources Management using the Existing and Proposed Approaches

1) Addition of a data source: A new source S6 with *WellData (API, Name, Field, County, Comp, Country)* will be added to the integration system in example 1

Using GAV: In GAV, adding a new data source leads to revisiting all the mapping assertions to see which one should be altered and may lead to the addition of a new assertion. The addition of the relation WellData will cause:

a) Changes to the mapping assertion 2, 3, and 4 under the GAV mapping assertions shown in section 4.1 to include union with new views Select API from S6.WellData where Country = ‘USA’ and API not in (Select WellAPI from S1.WellDetails union Select WellAPI from S5.NorthDakotaWells **Union** Select APINo from S2.Well), Select Field, County from S6.WellData where Country =’USA’, and Select Comp from S6.WellData where Country =’USA’ respectively.

b) Adding of a new mapping assertion to map the Name element to the GS relation Well. The new mapping assertion will be number 5, and will take this form Well (WellName) ➔ Select Name from S5.NorthDakotaWells Where Country = ‘USA’

Using LAV: In LAV, adding a new data source S6 will only cause adding a new mapping assertion 5, for the LAV mapping assertion Select API, Name, Field, County, Comp, Country from S6.WellData Where Country= ‘USA’➔Select WellAPI, WellName, FieldName, County , CompIntervalID, “USA” as Country from GS.Well

Using BAV: Adding S6 to the integrated system in example 3.1 using BAV approach, will be done by building a view vG over the GS relation Well and a view vS over the WellData relation from S6, and mapping vG to vS.

Using the proposed approach: In the proposed approach, adding a new data source S6 will be performed by adding a new union in the Well assertion with a view over the added source and the Well_Detectors assertion will be modified to include a new detector view of the added source. The two views are Select API, Name, Null as Latitude, Null as Longitude, Null as FieldName, County, Comp from S6.WellData Where Country = 'USA'. And the detectors view will be Select Null as WellMD, Null as HorizontalWellFlag, 'S6' as SourceName from S6.USAWell Where Country = 'USA' Afterwards the MappingHelper table will be updated to have this record ('Well', 'S6', 5)

2) Removal of a data source

If S5 in example 1 will be removed from the integration system built:

Using GAV: The following steps will be required to remove the S5 source:

a) Reconstruct the mapping assertion number 1 under the GAV mapping assertions shown in section 4.1. To remove the view representing S5.

b) Remove the mapping assertion number 4 as it contains CompIntervalID which comes only from S5,

c) Revisit the GS relation **Well** to remove the element CompIntervalID as it only exists in S5.

Using LAV: If S5 is removed from the integration system in example 1 LAV will remove the mapping assertion number 4, and revisit the GS relation Well to remove the CompIntervalID from there.

Using BAV: Removal of S5 in example 1 from the integration system built using BAV will cause to remove the mapping assertion number 4 underneath the BAV assertions, and revise the GS to remove the CompIntervalID from there.

Using the proposed approach: Removing S5 from the integration system build using the proposed approach in example 1 will be performed as follows:

a) Remove the view 3 from both the Well_Detectors and Well mapping assertions.

b) Remove from MappingHelper the records related to S5 and GS relation Well, finally update the MappingHelper data to keep the consistency of the ordering of the sources in the mappings caused by the removal of S5.

3) Removal and the addition of an attribute in a data source

Using GAV: To add API attribute to S3 in GAV, the mapping assertions which will be affected are:

a) The mapping assertion number 2 will be removed as it will not be needed.

b) The mapping assertion1 will be modified to include WellAPI attribute from all sources.

To remove APICompSTR attribute from S5 in GAV, both the mappings and the GS will be affected, where:

c) The mapping assertion number 2 will be revised to remove the view over S5.

d) The mapping assertion number 4 will be removed.

e) The GS will be revised to remove the CompIntervalID attribute from there.

Using BAV and LAV: To add API attribute to S3 in LAV and BAV, only the mapping assertion number 3 will be modified to include the API attribute. And to remove APICompSTR attribute from S5 in LAV and BAV, the mappings and the GS will be affected, where:

f) The mapping assertion number 4 modified to not include two attributes WellAPI and CompIntervalID

g) The GS will be revised to remove the CompIntervalID attribute from there.

Using the proposed approach: To add API attribute to S3 in the proposed approach, only the local view corresponding to S3 will be extracted and modified such that Null as WellAPI will be replaced by API. And to remove APICompSTR attribute from S5 in the proposed approach, only the local view corresponding to S5 will be extracted and modified to replace Substring (APICompSTR, 0, CHARINDEX (APICompSTR ,' ')-1) as WellAPI by Null as WellAPI and Substring (APICompSTR, CHARINDEX (APICompSTR,' ')+1, Length(APICompSTR)) as CompIntervalID by Null as CompIntervalID.

C. GS Management using the Existing and Proposed Approaches

If the attribute WellType intended to be added to the GS relation **Well,** for future usage, and at the same time the attribute WellAPI will be removed from GS relation Well, the existing mapping approaches will refuse the addition process and can handle the removal process as follows:

Using GAV: The removal of the attribute in GAV will cause to modify all the mapping assertions with correspondence to this attribute. E.g. assertion2 will be removed.

Using LAV and BAV: The removal of the attribute in LAV and BAV will cause to modify most of the mapping assertions with correspondence to this attribute. Explicitly mapping assertions number 1, 2, and 4 will be revised.

Using the proposed approach: For the addition, the proposed approach will be considering it, and will modify the mapping assertion of GS relation **Well** to include such attribute as the last attribute in the relation and map it with the sources as usual. If the attribute existing in any source will be mapped normally, else on the other case, the view over such source will have an extra attribute *NULL* as WellType. In example 1 all local views will have *NULL* as *WellType*. Moreover the removal of an attribute will be simpler as no need to revise the GS relation Well, what will be done is parsing only the local views and replace the attribute mapping with a *NULL* as *WellAPI.*

Finally, these are other features provided by the proposed approach:

a) Prepares the environment for the successive process, duplicate detection and resolution.

b) Handles the situation where no shared identifier for the real world object was common between the data sources in the integration system.

c) Ensures correctness and efficiency of collecting the answers from the different data sources as the views already linked with a traditional union operator and all the views have the same arity as the GS relation.

d) Ensures the completeness of the query answers, as it allows all alternatives for the same real world object from all data sources, and does not prevent any source to participate in the mapping construction.

D. Approach Limitations

This approach is limited to mapping of the relational schemes of the available data sources and the GS should be in relational form. Another limitation is; the detectors mapped should be defined prior to the integration system development and should be defined by the domain expert.

E. Complexities in the Proposed Approach Compared to other Approaches

Table 3 shows the mapping assertions complexities for the proposed approach compared to other approaches.

The notations used in the comparison and calculations of the complexities are; **N**: # GS relations, **n**: # relevant information sources, **Ys**: # GS relations a data source s has correspondence with, **R**: # GS relations appear in the user's query, and **Tr**: # Mappings for a given GS relation R, Ai is number of sources used to map GS relation Ri.

TABLE III. COMPARISON BETWEEN THE PROPOSED AND EXISTING APPROACHES

	GAV	**LAV**	**BAV**	**Proposed Approach**
# mapping assertions	∑Ai where i=1 ..N.	∑Ys where s=1, ..n	Min: N, Max = N * n	2 * N
# assertions revised for adding/removing data source	Min: 1, Max = Ai for removing, and max. N for addition.	Ys for removing and max. N for addition	Ys and max. N for addition	2 * Ys
# mapping assertions extracted for answering user query	∑Tr where r=1 …R.	Min: R, Max = n * R	∑Tr where r=1, …R	2 * R
# mapping assertions revised for removing GS relation	Min: 1, Max = n	Min: 1, Max = n	Min: 1, Max = n	2
# mapping assertions revised for adding GS relation	Min: 1, Max = n	Min: 1, Max = n	Min: 1, Max = n	2

V. CONCLUSION AND FUTURE WORK

In this paper a new mapping approach is introduced to avoid most of the noticed limitations in the existing approaches; as it is not only mapping the GS elements with the local schemes, but also mapping the elements required for detecting and resolving the conflicts happened due to the integration process. The proposed approach facilitates the extensibility of the GS, and the sources. The proposed approach provides improvement in adding, removing and updating the global schema GS and the sources S. The proposed approach links the 3 main processes required to answer the user's queries to help in providing complete, and unambiguous answers to those queries. As a future work; formalizing a duplicate detection algorithm to leverage this mapping approach and the detectors defined to detect the duplicates, and use the sources of the data to resolve the duplicate through source preferences.

REFERENCES

[1] M. Lenzerini, "Data integration: A theoretical perspective", Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Madison, Wisconsin, USA, June 2002.

[2] B. Golshan, A. Y.Halevy, M. Mihaila and W. C. Tan. "Data integration: after the teenage years", Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS'17), Chicago, Illinois, USA, 2017.

[3] L. Xu and D. W. Embley, "Combining the best of Global-as-View and Local-as-View for data integration", Proceedings of the 3^{rd} ISTA, Salt Lake city, Utah, USA, 2004.

[4] E. Rahm and H. H. Do. "Data cleaning: problems and current approaches", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 23, 2001, pp. 103-113.

[5] X. Chu, I. F. Ilyas, and P. Koutris, "Distributed data deduplication", Proceedings of the VLDB Endowment, vol. 9, 2016, pp. 864-875.

[6] A. Elmagaramid, P. G. Ipeirotis, and V.S. Verykios, "Duplicate record detection: a survey", IEEE Transactions on Knowledge and Data engineering, vol.19, 2007, pp. 1– 16.

[7] M. Nentwig, M. Hartung, A. Ngomo and E. Rahm, "A survey of current link discovery frameworks", Semantic Web Journal, vol. 8, 2016, pp. 419-436.

[8] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute and V. Raghavendra, "Deep learning for entity matching: a design space exploration", Proceedings of the International Conference on Management of Data SIGMOD'18, TX, USA, June 2018.

[9] Y. Yang, Y. Sun, J. Tang, B. Ma and J. Li, "Entity matching across heterogeneous sources", Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), Sydney, NSW, Australia, August 2015.

[10] A. Gruenheid, X. L. Dong and D. Srivastava, "Incremental record linkage", VLDB Endowment, vol. 7, 2014, pp. 697-708.

[11] E. K. Rezig, E. C. Dragut, M. Ouzzani and A. K. Elmagarmid, "Query-time record linkage and fusion over Web databases", Proceedings of IEEE 31st International Conference on Data Engineering, Seoul, South Korea, April 2015.

[12] E. K. Rezig, E. C. Dragut, M. Ouzzani, A. K. Elmagarmid and W. G. Aref, "ORLF: A flexible framework for online record linkage and fusion", Proceedings of IEEE 32nd International Conference on Data Engineering, Helsinki, Finland, May 2016.

[13] I. F. Ilyas and X. Chu, "Trends in cleaning relational data: consistency and deduplication", Foundations and Trends in Databases Journal, vol. 5, 2015, pp. 281-393.

[14] A. Bronselaer, D.V. Britsom and G.D. Tre, "Pointwise multi-values fusion", Proceedings of the 18th International Conference on Information Fusion, Washington, USA, July 2015.

[15] D. Dubois, W. Liu, J. Ma and H. Prade, "The basic principles of uncertain information fusion. An organized review of merging rules in different representation frameworks", Proceedings of Information Fusion Heidelberg, Germany, July 2016.

[16] A. Bronselaer, D.V. Britsom and G.D. Tre, "Propagation of data fusion", IEEE Tran. on Knowledge and data engineering, vol. 27, 2015, pp. 1330 – 1342.

[17] X. Chen, E. Schallehn and G. Saake, "Cloud-scale entity resolution: current state and open challenges", Open Journal of Big Data (OJBD), vol. 4, pp. 30-51, 2018.

[18] A. Gal, "Tutorial: uncertain entity resolution", VLDB Endowment, vol. 7, 2014, pp. 1711-1712.

[19] J. Bleiholder and F. Neumann, "Conflict handling strategies in an integrated information system", Workshop on Information Integration on the Web (IIWeb), Edinburgh, UK, May 2006.

[20] A. Bilke, J. Bleiholder, C. Bohm, K. Draba, F. Naumann and M. Weis, "Automatic data fusion with HumMer", Proceedings of the 31st VLDB, Trondheim, Norway, September 2005.

[21] A Motro, J Berlin, and P. Anokhin, "Multiplex, Fusionplex and Autoplex: three generations of information integration", ACM SIGMOD Record, vol. 33, 2004, pp. 51-57.

[22] G. D. Giacomo, D. Lembo, M. Lenzerini and R. Rosati, "Tackling inconsistencies in data integration through source preferences", Proceedings of the International Workshop on Information Quality in Information Systems, Paris, France, June 2004.

[23] Y. Katsis, A. Deutsch, Y. Papakonstantinou and V. Vassalos, "Inconsistency resolution in online databases", proceedings of IEEE 26th International Conference on Data Engineering (ICDE), Long Beach, California, USA, March 2004.

[24] P. N. Mendes, H. Muhleisen, and C. Bizer, "Sieve: linked data quality assessment and fusion", 2nd International Workshop on Linked Web Data Management (LWDM 2012) at the 15th International Conference on Extending Database Technology, Berlin, Germany, March 2012.

[25] W. Fan, F. Geerts, N. Tang and W. Yu, "Inferring data currency and consistency for conflict resolution", in Proceedings of the 2013 IEEE International Conference on Data Engineering, Brisbane, Australia, April 2013.

[26] A. Cali, D. Calvanese, G. De Giacomo and M. Lenzerini, "Data integration under integrity constraints", In Proceedings of the 14th International Conference on Advanced Information Systems Engineering, Ontario, Canada, May 2002.

[27] T. Kirk, A. Y. Levy, Y. Sagiv and D. Srivastava, "The information manifold", in Proceedings of the AAAI Spring Symp. On Information Gathering from Heterogeneous, Distributed Enviroments, Cambridge, Massachusetts, United States, November 1995.

[28] P.J. McBrien and A. Poulovassilis, "Data integration by bi-directional schema transformation rules", Proceedings 19th International Conference on Data Engineering, Bangalore, India, March 2003.

[29] R. Fagin, L. M. Haas, M. Hernandez, R. J. Miller, L. Popa and Y. Velegrakis, "Clio: schema mapping creation and data exchange", Conceptual Modeling: Foundations and Applications, Springer-Verlag, Berlin, Heidelberg, 2009.

[30] E. Rahm. "The case for holistic data integration", in Proceedings of East European Conference on Advances in Databases and Information Systems, Prague, Czech Republic, August 2016.

[31] A. Alsarkhi, and J. R. Talburt, "A method for implementing probabilistic entity resolution", IJACSA, vol. 9 (11), November 2018, pp 8-15.

Automation of Combinatorial Interaction Test (CIT) Case Generation and Execution for Requirements based Testing (RBT) of Complex Avionics Systems

P Venkata Sarla[1]
Research Scholar, Bharathiar University,Coimbatore
Scientist-G, Aeronautical Development Agency,
Bangalore, India

Dr. Balakrishnan Ramadoss[2]
Professor, Department of Computer Applications,
National Institute of Technology,
Trichy, India

***Abstract*—In the field of avionics, most of the software systems are either safety critical or mission critical. These systems are developed with high quality standards strictly following the relevant guidelines and procedures. Due to the high criticality of the systems, it is mandatory that the verification and validation of these systems are done with utmost importance and only then any system is cleared for flight trials. The verification and validation activities need to be very exhaustive and hence take a considerable amount of time in the software development lifecycle. This paper describes about the innovative approach towards automation of Combinatorial Interaction Test case generation and execution for Requirements Based Testing of complex avionics systems for achieving test adequacy in a highly time efficient and cost efficient manner.**

Keywords—Avionics; combinatorial interaction testing; requirement specifications; requirements based testing; safety critical; validation; verification

I. INTRODUCTION

Avionics systems are complex real time embedded systems with a very high criticality associated with them. These systems are software intensive and exhaustive verification and validation activities need to be carried out both at system level and software level to ensure error free and safe functioning of the system. Verification of the Software Development Life Cycle (SDLC) deliverables right from requirements engineering phase is essential in order to ensure that defects are discovered early and fixed as doing it at later stages has high impact on cost and effort.

The validation testing of avionics system is done with the Software Under Test (SUT) running on the actual target hardware and all the interfacing subsystems simulated. Implementation of each of the functionality is tested by running a number of test cases on the SUT. The test cases for the Functionality Under Test (FUT) are designed to uncover errors, demonstrate that the inputs are properly accepted by the SUT and the outputs are correctly produced. Validation testing is basically black box testing that examines the aspects of system functionality with little regard for the internal logical structure of the software. The SUT and the simulated systems run in real time during the validation tests.

A. Combinatoral Interaction Testing

Combinatorial Interaction Testing (CIT) can detect failures triggered by interactions of parameters in the SUT with a covering array test suite which tests all the required parameter value combinations. Traditionally testers develop scenarios of how an application will be used, then select inputs that will exercise each of the application features using representative values, normally supplemented with extreme values to test the performance and reliability. The problem with this often ad hoc approach is that unusual combinations will usually be missed, so that a system may pass all tests and may work well under normal circumstances, but may eventually encounter a combination of inputs that it fails to process correctly. By testing all combinations, for a specific interaction strength within the input variables, CIT can help to avoid this type of situation.

B. Requirements based Testing

A general principle of good requirements engineering practice [1] is that requirements should be testable. Requirements Based Testing (RBT), therefore, is a systematic approach to test case design where you consider each requirement and derive a set of tests for it. RBT is done to demonstrate that the system has properly implemented its requirements. By combining methods from requirements engineering and software testing, this testing methodology provides a set of quality assurance activities and management tools that enable getting requirements right from the outset. The RBT process addresses two major issues [2] first, validating that the requirements are correct, complete, unambiguous, and logically consistent; and second, designing a necessary and sufficient (from a black box perspective) set of test cases from those requirements, to ensure that the design and code fully meet the requirements. When designing tests, two issues need to be overcome: reducing the enormous number of potential tests down a reasonable size set and ensuring that the tests got the right answer for the right reason. The RBT process will drive out ambiguity and drive down the level of detail. The overall RBT strategy is to integrate testing throughout the SDLC and focus on the quality of the requirements specification. This leads to early defect detection which has been shown to be much less expensive than finding defects during integration testing or later. The RBT process also has a focus on defect prevention, not just defect detection. The test cases for each of the FUT are designed using the corresponding Software Requirements Specifications (SRS) and Interface Requirement Specifications (IRS). For the FUT, the requirements related to processing of input data and generation of output data are specified in the SRS. The address and format of the input and output parameters are defined in the IRS between the SUT and the interfacing subsystems for the FUT.

C. *Contents of the Pape*

The rest of the paper is structured as follows. Section II discusses literature survey on related work. Section III introduces the new approach of CIT for RBT of complex avionics systems which is explained with a case study detailed in sections IV and V. In Section IV the manual testing approach used for a Mission critical system in a combat aircraft/helicopter is explained followed by the disadvantages of manual testing. In Section V automation of CIT cases generation and enhancement of the manual test rig for automatic execution for RBT of the system is elaborated followed by the advantages of automation.

II. Literature Survey

A. *Automatic Test Data Generation*

In [3] development of Test Case Generation (TCG) algorithm for CIT and idea for considering input constraints and building a unit testing harness from TCG is addressed. In [4] and [5], the authors have used programs from Software-artifact Infrastructure Repository (SIR) as their subjects for examining the effectiveness of CIT on regression testing. In [6], the authors illustrated that adding constraints in CIT of highly configurable systems, reduces the number of feasible system configurations but it is not guaranteed to reduce the size of the CIT sample to achieve coverage of desired strength. In [7] covers discussion on integrated approach for finding covering arrays and application of the same for constructing variable strength arrays. In [8] an approach to automate unit and integrating testing of radio’s control software is described. In [9], the authors have illustrated an automated approach for finding and fixing conformance faults between given software system and its combinatorial model. In [10] automatic generation of test configurations that cover all pair-wise interactions using feature models for testing Software Product Line (SPL) is explained. In [11] the authors have proposed a framework for automated pair-wise testing of SPL, with an objective to generate the minimal set of test configurations that are valid and cover all pair-wise feature interactions.

B. *MC/DC Coverage with CIT*

In [12] automatic test data generation for testing of C programs at white box level for obtaining multiple coverage criteria including MCDC is covered. In [13], the authors have discussed about the extent of statement/branch and MC/DC coverage and the Fault Detection Rate (FDR) that can be achieved by executing CIT cases with strength varying from 2 to 5 on two subjects taken from SIR. They have taken original implementation and number of mutants of the subjects to study the effectiveness of CIT. Though they have generated covering arrays with strength of 5, as executing all the test cases with higher combination strength is a huge effort, the number of test cases that were executed is limited to that with strength 4. But a decrease in the statement/branch/MCDC coverage and FDR was noted by running only a subset of test cases with higher strength when compared to execution of complete set of test cases with lower strength.

But for mission critical and safety critical avionics software systems, it is essential to achieve 100% statement/branch and MC/DC coverage so that the FDR also is high. In f DO-178B guidelines followed for development and certification of avionics systems, RBT is emphasized because this testing strategy is found to be most effective in revealing errors. From the new approach followed by us we have found that the advantages of CIT for RBT are more compared to CIT with higher strength than needed.

III. Introduction to New Approach

We have evolved a new approach of performing CIT for RBT for verification and validation of complex avionics systems involving interactions of varying strengths within the parameters of the functionalities. CIT for RBT with minimum required strength is more effective in uncovering the errors with lesser effort than performing CIT with higher strength than required. For computing the expected output of the CIT cases for each FUT, the corresponding reference models are developed using the corresponding SRS and IRS. Because of this approach, the requirements get refined at initial stages of SDLC saving time of rework if detected later. The required optimal strength for CIT of the FUT is derived from the requirements thus validated and elaborated instead of generating and executing CIT cases with higher strength than needed. This approach which provides the benefits of both CIT and RBT involves the following activities:

1) *Generation of expected output for each test case*

As the FUTs will be computation intensive involving number of parameters, for generating expected output for each set of inputs, reference models are developed.

2) *Generation of optimal, reusable combinatorial interaction test cases for RBT*

Because of complex nature of requirements and the typical constraints on values of input as well as intermediate and output parameters for the systems of this domain, for generation of test data, additional considerations are required as compared to systems of other domains. Hence this activity is automated by enhancing the reference models for the FUTs and integrating with covering array generation tool for CIT suite.

3) *Execution of CIT for RBT*

As the systems are highly interface intensive with a number of other sub systems interfacing through various buses, feeding the inputs to the SUT for a particular FUT is highly cumbersome. Hence the test rig is enhanced and integrated with the CIT suite for automated execution of CIT for RBT.

4) *Generation of test reports*

As the number of test cases is too many, generation and tracking of test results manually is very difficult. Hence the test report generation activity is automated.

As per [15], generally, automation always follows manual testing. Typically, one or more rounds of manual testing already would be performed on the automated test rig. This implies that manual test cases already exist and have been executed at least once. But in our approach, even the first time execution of test cases can be done automatically.

IV. Case Study

To explain automation of test data generation for CIT and automation of execution of the generated test cases, we have taken the case of Mission Management Computer (MMC) software as a case study.

Introduction to the MMC which is the SUT, manual test rig used for black box testing, manual test procedure and the drawbacks of the same are explained in sections A to D below. Further in section V the automation of all the activities for the new approach of CIT for RBT of MMC is explained.

A. *Mission Management Compute*

In a modern combat aircraft/helicopter, the Mission Management Computer (MMC) is a highly complex unified software system. It is the heart of the avionics architecture which is the bus controller for more than 25 subsystems connected to it namely, Weapon Management System, Multi Mode Radar, Laser Designation Pod, Cockpit Controls System Redundancy management system, Data Acquisition System, Fuel System, Engine System, Electrical System, Brake Management System, Hydraulics System, Environment Control System, Recording And Replay Systems, multiple display processors for displaying more than 1000 symbols on MFD (Multi Function Display unit), Head-Up Display (HUD) unit, HMD (Helmet Mounted Display), Communication Systems, Backup Instruments, Vehicle Health Management Systems, Flight Control System. These systems interface with MMC on different buses like 1553B, RS422, Video and discretes.

Most of the mission management functions of the combat aircraft are implemented in MMC, mainly, weapon management for various modes of air-to-ground and air-to-air attack functions, redundancy management of the external interface systems for fault tolerance, Pilot Vehicle Interface (PVI) functions that include processing of cockpit controls, driving various display surfaces and Warning/Caution management, sensor management functions and so on. Thus there are thousands of software requirements for the MMC system to receive and process data from multiple subsystems and transmit the processed data to other subsystems. The MMC software is developed incrementally with a set of new functionalities added during each iteration.

B. *Manual Test Rig for Testing of MMC*

1) *Components of the Manual Test Rig*

In order to facilitate testing of various functionalities implemented in the SUT, the test facility has the following components as shown in Fig–1.

- Desk top PCs for Simulated Interface Models (SIMs) of the subsystems interfacing with the MMC on different data buses same as in the target aircraft.
- Provision of connecting MMC (SUT).
- Power Supply unit for SUT and the SIMs.
- Patch panel for feeding /tapping various signals/data.

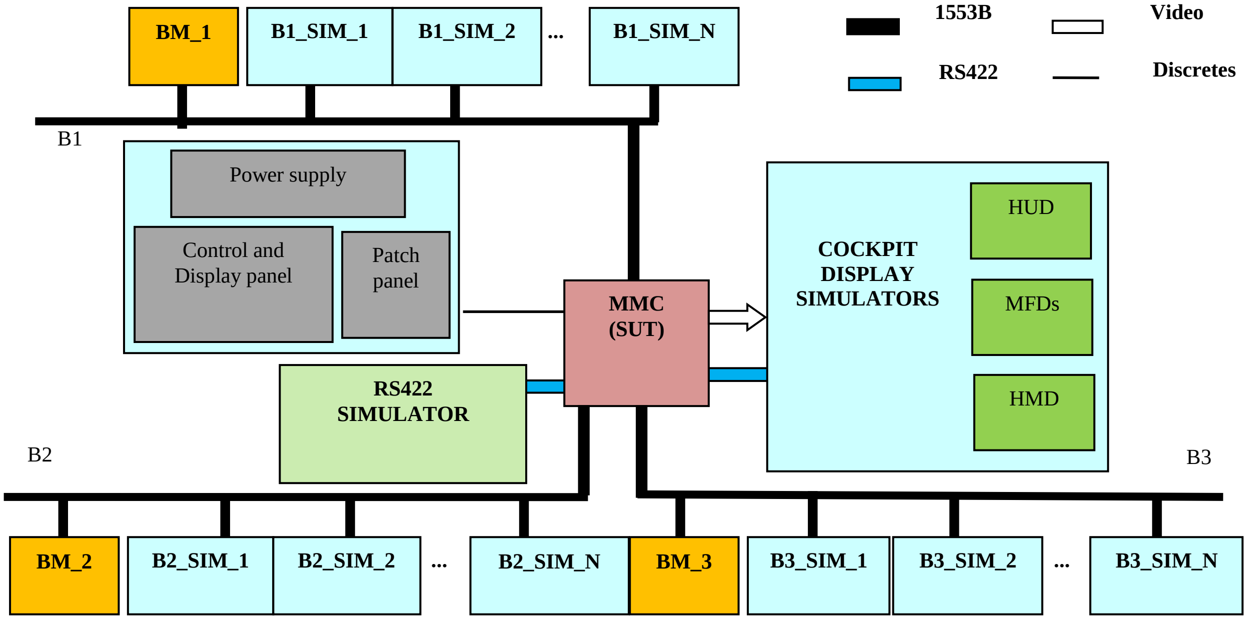


Fig. 1. Test Rig for Manual Testing of MMC.

- Control and display panel for ON/OFF control of SUT, isolation of SUT from bus, monitoring the health status of various components of the test rig etc.
- Bus Monitors (BM_1/2/3) for capturing data on the 1553B buses for analysis.
- RS422 Simulator for simulating RS422 interfaces to SUT same as in the target vehicle.
- Cockpit display simulators – Head Up Display (HUD), Multi-Function Displays (MFDs) and Helmet Mounted Display (HMD).

2) Simulated Interface Models

The test rig which is used for black box testing of MMC comprises of a network of Simulated Interface Models (SIMs) of various subsystems as depicted in Fig-1. The SIMs mimic the actual subsystems connected to the SUT in the target vehicle in terms of inputs and outputs to the SUT.

Each SIM will have the provision for feeding the inputs through GUI in engineering formats in a Transmit (TX) Window. These values are converted by the SIM into the required digital format and are updated on the corresponding data bus. Similarly, the outputs from the SUT which will be in digital format (Hexadecimal/Binary/Octal/ASCII/BCD are received by the SIM to which the data is addressed. They will be translated into engineering format and displayed in its GUI in Receive (RX) Window.

The UI of the SIMs shall also have the provision to feed transmit data in digital format which will get automatically updated in engineering format in the TX window. Similarly, there shall be provision to view the received data in digital format also. The test cases are static in the sense, for a particular test case, the values fed from the SIMs are constant. Hence the expected output of the SUT for the test case will be constant. However, the simulated subsystems and the SUT run dynamically in real time as per the bus scheduler functionality.

C. Manual Procedure for Generation and Execution of Test Cases for MMC

For each of the functionalities implemented in the MMC, test cases are generated and executed manually as follows:

1) The input parameters for the FUT are identified.

2) The address of the interfacing sub-system for each of the input parameters, the corresponding message details (the bits /words) and the range of values of the input parameters, are identified from the IRS.

3) Test cases are generated to ensure that the SUT is tested for different values of each of the input parameters and different combinations of input parameters. The values chosen for the input parameters include boundary values and mid value of the range. Additionally, as per the SRS, if there are any decisions/conditions with respect to specific values of any parameter, then values >, < and = to the specific value are added.

4) For each of the test case, the expected output value is calculated based on the SRS and converted into the format as per the IRS and specified in the test case document.

5) For running a test case on the test rig, the tester needs to feed the input values in various SIMs as per the test case, observe the output on the cockpit display surfaces and/or on the RX window of the SIMs which consume the output of the SUT corresponding to the FUT.

6) The observed output is compared with the expected output manually and the test result as PASS or FAIL is updated in the test report.

D. Drawbacks of Manual Methods of Generation and Execution of Test Cases

The process of generation of test cases manually is not very efficient and has many drawbacks. The test cases are not easily retrievable and reusable for regression testing of incremental software upgrades. The extent of combination coverage and path coverage depends on the randomly selected values of the input parameter.

The manual execution of test cases is time consuming, cumbersome and non repeatable as explained below.

- Values of different input parameters for the test case need to be provided across different subsystem terminals manually.
- If a sequence of inputs is needed to be provided within a specific timeframe consecutively, it is very difficult to achieve in the current approach. Requires multiple retries.
- Output result needs to be observed across multiple subsystem terminals manually.
- If the requirement is to update the output data only for a specific duration (for e.g., setting of a FLAG for one cycle), it is very difficult to observe the same. Tester needs to capture the data using the bus monitor terminal during run time and analyse offline whether the output data is updated correctly during the expected time duration in correlation with the input data.
- If any test case fails, then for demonstrating the failure to the designers, the whole process needs to be done again manually. If any observation is non-repeatable, getting the right scenario to get the observation becomes impossible some times.
- The systems are developed as unified systems suitable for different variants of target vehicle. For similar functionality, the expected behaviour of the system for the same input conditions will be different across different modes of operation (for e.g., navigation, approach, landing, weapon aiming, weapon releasing, exiting from attack mode, jettisoning etc.) and for different variants of the target vehicle (Airforce/Navy /Fighter/Trainer). The test cases are not easily reusable. A particular testing scenario if needs to be repeated for a different mode, then all the set of inputs need to be provided manually again across multiple terminals.
- Even if there are minor changes in the upgraded software releases, regression testing for clearance of

the upgraded version of the software takes same time as taken for the initial clearance of the software.

- Though modular design methodologies are used, every time when the software is upgraded for new functionalities, the impact of the changes on the existing software is very huge. The reason is the complex nature and huge size of the order of three-to-four million lines of source code. Performing detailed impact analysis manually is impractical. Even the usage of Computer Aided Software Engineering (CASE) tools for impact analysis of this type of software upgrades on the previously working software has been proven to be impractical. The tool shows hundreds of relationships across various objects of the source code. So based on gross level impact analysis carried out manually and with the knowledge of previous defect history most of the testing is repeated for previously implemented requirements. But as explained above, the effort for re-executing the test cases is very huge.

Because of above reasons, though the effort involved in automation is significant, it is one time effort which will help in reducing the regression testing time drastically for various upgrades of MMC software. During the development and maintenance of avionics systems which extend to about 15-20 years, there will be hundreds of software upgrades released incrementally. Once the setup for generation and execution of CIT for such systems is established, the same can be reused with no or minimum changes for testing the upgraded versions during each iteration. The test case generation and execution activities become more of process dependent than person dependent. In this domain where attrition of test engineers is very high, having this type of process dependent testing mechanism helps in a very big way for the organization.

V. Automation of CIT for RBT of MMC

In order to increase the efficiency of testing by using CIT and to ensure 100% requirements coverage for the RBT of MMC, the following activities are automated.

a) Generation of Expected Output for each test case by developing reference models for the FUT

b) Generation of Test data for the CIT cases

c) Execution of Combinatorial Interaction Test Cases for RBT of MMC

d) Generation of Test Report

Out of the above four activities that are automated, (b) & (c) explained in section B and C are unique to MMC testing. These methods are first time evolved and applied and are highly beneficial in many ways as explained in section VI. Though (a) and (d) are similar for systems belonging to various domains, (a) is explained in section A in brief as the same is used as the basis for (b). (d) is covered in section D for completeness.

A. *Automation of Generation of Expected Output*

Development of the reference models as shown in Fig-2 for different FUTs is carried out for generating the expected output values for every test case. The reference models are independently developed by the testing team based on the corresponding SRS. Generally the programming language used for the reference model is different from the one used by the design team in the actual SUT. The values for the input parameters for each test case are based on the IRS between the SUT and the interfacing subsystems for the FUT.

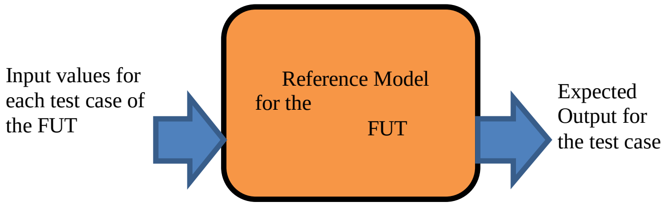

Input values for each test case of the FUT
Reference Model for the FUT
Expected Output for the test case

Fig. 2. Use of Reference Models for Generation of Expected Output.

B. *Automation of Test Case Generation*

Automation of generation of test data for the test cases of CIT involves enhancement of the reference model for the FUTs to include Constraint Checker (CC) and integration of the same with covering array generation tool. The test cases with the input test data and the expected output data are stored in the CIT Suite. For CIT to be effective for avionics systems, additional considerations are necessary for generation of optimal CIT suite with respect to CIT of highly configurable systems.

1) Additional steps to be taken for generation of test data for effective CIT of avionics systems

The steps for evolving the required combinations that need to be covered for evolving the test data for CIT of MMC software are depicted in Fig-3. The reasons for these additional considerations are:

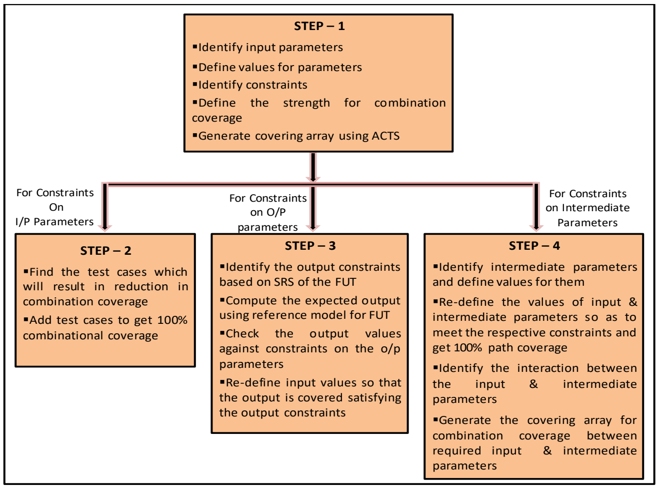

STEP – 1
▪Identify input parameters
▪Define values for parameters
▪Identify constraints
▪Define the strength for combination coverage
▪Generate covering array using ACTS
For Constraints On I/P Parameters
For Constraints on O/P parameters
For Constraints on Intermediate Parameters
STEP – 2
▪Find the test cases which will result in reduction in combination coverage
▪Add test cases to get 100% combinational coverage
STEP – 3
▪Identify the output constraints based on SRS of the FUT
▪Compute the expected output using reference model for FUT
▪Check the output values against constraints on the o/p parameters
▪Re-define input values so that the output is covered satisfying the output constraints
STEP – 4
▪Identify intermediate parameters and define values for them
▪Re-define the values of input & intermediate parameters so as to meet the respective constraints and get 100% path coverage
▪Identify the interaction between the input & intermediate parameters
▪Generate the covering array for combination coverage between required input & intermediate parameters

Fig. 3. Steps to be followed for Generating Combinatorial Test Cases for MMC.

a) Need for additional test cases to meet typical constraints on input parameters.

Most of the input parameters to MMC are associated with validity bits. As per the requirements whenever validity bit of any parameter is received as INVALID, MMC uses the previous valid value for a specified duration. Because of this type of constraints on the input parameters, the SUT will not get tested for some combinations of values of inputs. Hence CIT is an effective test generation technique for avionics systems only when additional test cases are added to meet the complex nature of input constraints and to get 100% combination coverage of required strength.

b) Need for redefinition of values of input parameters to meet typical constraints on intermediate parameters and output parameters.

Many of the functionalities that are implemented in the MMC software are computation intensive. The computed values are used either to display a symbol and/or data on the cockpit display surfaces and/or send the data to other systems for consumption. For a set of inputs, if the output which is the current_location of a moving symbol on a cockpit display surface is out of the dispay_area or Field Of View (FOV) then, based on the requirements, EITHER the symbol shall be made absent OR the symbol shall be displayed at the boundary in flashing.

Because of the reasons mentioned above, the choice of values of the input parameters should be such that the corresponding output values do not result in the location of the symbol being out of FOV. Only two test cases are required to test the symbol for absence/flashing. All other test cases should be such that resultant locations of symbol are distributed across the entire FOV instead of cluttered at some portions.

Similarly if the computed data is sent to other equipment, the values of the input parameters in the test cases should be such that the resulting output values are within valid output range, the values are distributed across the entire range of the output parameter and the Output_Data INVALIDITY bit is not set for more than one test case.

Every combination in the test suite needs to be checked by running on the corresponding simulated reference model (refer Fig-2) to ensure whether the resultant values of the output parameters are meeting the above output constraints. If not, the input values need to be redefined. Thus there is an impact of output constraints on the selection of values for various input parameters.

Similarly, in the algorithms for various functionalities, there will be constraints on the intermediate variables which are dependent on input variables. Based on the values of these intermediate variables, the program takes multiple paths. In order to ensure that the test cases are adequate enough to cover all the paths, it is essential that these types of constraints on intermediate values are met. Accordingly the values of the input parameters need to be redefined.

For effective CIT, wherever there are constraints on the values of intermediate/output parameters, the values of corresponding input parameters need to be redefined to meet those constraints. However, the size of the test suite and the combinations should not increase significantly.

c) Need for generation of covering arrays with combination coverage for the input parameters and the intermediate parameters.

The existing combination strategies [16] [17] are inadequate for handling intermediate parameters for combination coverage required for avionics software testing. In the algorithms for different functionalities, there will be decisions/paths based on conditions with combinations of input and intermediate parameters. Hence generation of covering array with combinations of input parameters alone will not be adequate. The covering array needs to be generated with combination coverage for the input parameters and the intermediate parameters to get 100% condition/decision coverage.

2) Development of facility for evolving CIT cases for RBT of MMC

We have developed the Combinatorial Interaction Test case Evolving Facility (CITEF). This facility is useful for evolving optimal test cases with input values for the parameters of the FUT such that the typical constraints on the input/intermediate and output parameters of the FUT in MMC are met. Fig-4 shows the Block Diagram of CITEF. It comprises of Input Value Generator (IVG), Reference Model (RM) of the FUT and CTCG tool for covering array generation. The RM in turn has two components: The Simulated Functionality Under Test (SFUT) and Constraint Checker (CC). SFUT is developed independently by test team members based on the corresponding requirements specified in the SRS of MMC. CC shall have the intermediate and the output constraints applicable for the FUT stored in it. The IVG is GUI based application which has the provision for entering the initial set of input values and range of data for each parameter.

The initial set of input values are derived through category partitioning [18] which involves selection of typical representative values based on input domain partitioning and boundary values as per the interface requirements. The IVG has the provision to manually update the values of the parameters or to automatically select random values (without duplication) from the given range of the parameters. This provision is given so that the size of the test suite does not increase abnormally. Else IVG was selecting random values from the valid range and with an increase in number of values of any parameter, the size of the test suite increases exponentially and the testing time will proportionally increase.

We have used ACTS Version 3.1 released in April 2018 for generating covering arrays for CIT of MMC. ACTS [19] [20] [14] is a GUI-based CIT tool developed by National Institute of Standards and Technology (NIST). ACTS has the ability to generate tests with interaction strength from 2-way to 6-way, with a user-friendly GUI and a command line version suitable for use in scripts or system calls from another tool.

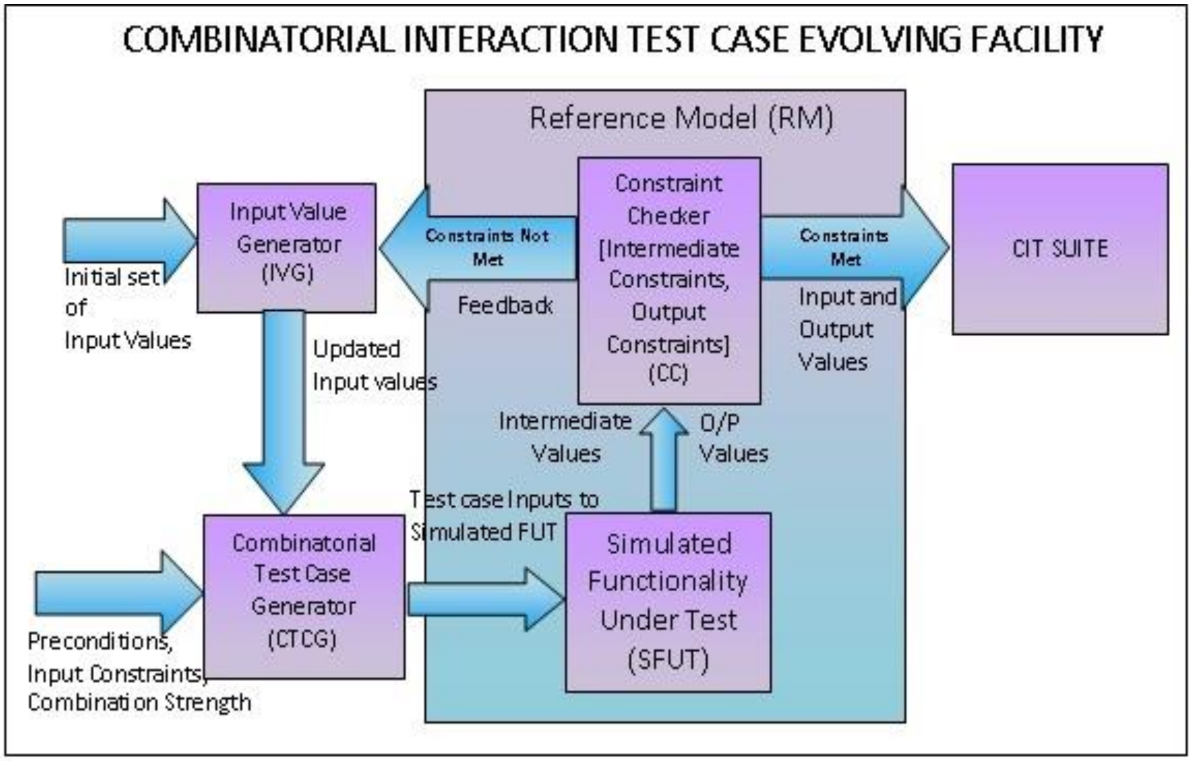

COMBINATORIAL INTERACTION TEST CASE EVOLVING FACILITY
Reference Model (RM)
Input Value Generator (IVG)
Initial set of Input Values
Constraints Not Met
Feedback
Constraint Checker [Intermediate Constraints, Output Constraints] (CC)
Constraints Met
Input and Output Values
CIT SUITE
Updated Input values
Intermediate Values
O/P Values
Test case Inputs to Simulated FUT
Combinatorial Test Case Generator (CTCG)
Preconditions, Input Constraints, Combination Strength
Simulated Functionality Under Test (SFUT)

Fig. 4. Combinatorial Interaction Test Case Evolving Facility (CITEF).

Combination Test Case Generator (CTCG) developed by us is an application to which ACTS tool is integrated. The values supplied by the IVG are accepted by an input file within the CTCG. These values along with the constraints on the input parameters and the combinations coverage strength (mixed strength wherever required) are fed to the ACTs tool. The generated test cases are stored in an output file. For the FUT, each of the generated combinatorial test cases shall be run on the SFUT in order. During execution of each test case, the run time intermediate values and the output values are sent from the SFUT to the CC in the RM. These values are checked by the CC against the respective constraints which are stored in it. If they are not satisfied, then feedback from CC is sent IVG and execution of further test cases on SFUT is stopped. The feedback information will contain the test case number, values of the intermediate and output parameters and the information about constraints that are not met. The same is displayed in the GUI of IVG. On selecting the EDIT option on the IVG for a particular input parameter, the values of that parameter will change randomly or a fixed value can be fed by the user. When the new values are applied by pressing the APPLY button, the same are sent to the CCTG Tool for generation of updated covering array. Each of the newly generated combinatorial test cases shall be again run on the SFUT. This process is repeated till the optimal values are assigned to the input parameters for every test case with values of all the input parameters satisfying the constraints on intermediate parameters and output parameters. The final test suite shall be such that on running all the test cases on the MMC, 100% path coverage and combination coverage of the parameters shall be achieved.

The test cases with the generated test data are stored in the Test Case Library in the CIT SUITE along with the preconditions and the expected output value for each test case. For each of the input and output parameters the corresponding address details (BUS ID, SIM ID, MESSAGE ID, Word number and Bit details) are also stored. The test data can be automatically fed to the SUT at black box level through the test rig as explained in the following section.

C. Automation of Execution of Test Cases

For Automation of execution of RBT of MMC, the test rig used for manual testing (Fig-1) is augmented and further integrated with CIT SUITE of CITEF as depicted in Fig-5.

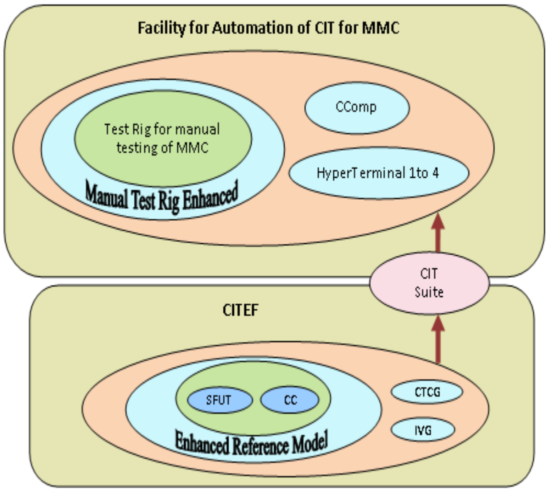


Fig. 5. Depiction of Components of Facility for Automation of CIT for RBT.

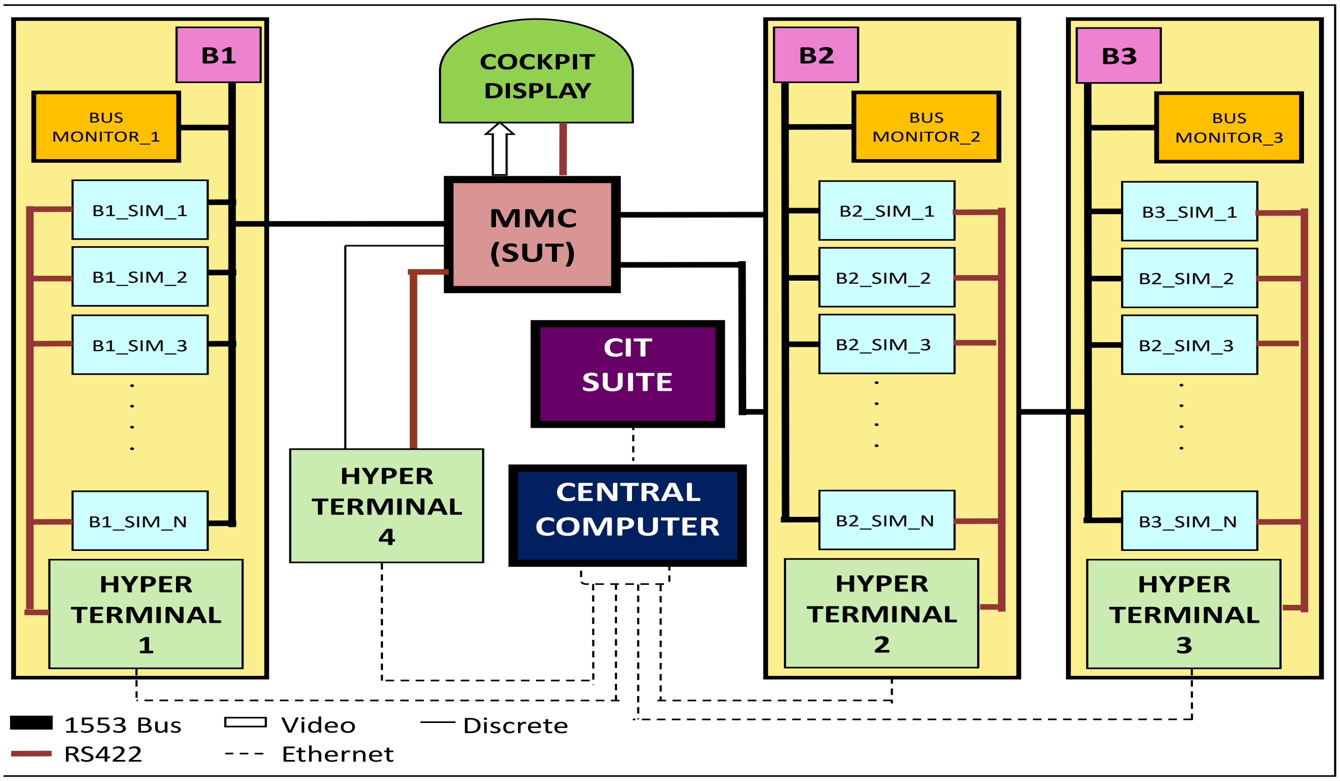


Fig. 6. Test Rig for Automation of CIT for RBT of MMC.

The block diagram of the Test rig for automation of execution of CIT for RBT of MMC is depicted in Fig-6.

Each SIM has a unique identification number BX_SIMY where X and Y are variables. X is the Bus 1/2/3 on which the SIM is connected. Y is the SIM number on that bus. For e.g.

Sub-System_1 on Bus1: B1_SIM1

Sub-System_2 on Bus 1: B1_SIM2

The test cases for the FUT are selected from the Test Case Library in the CIT SUITE. The Central Computer (CCOM) is the interface between the CIT SUITE and the Test Rig. For a chosen test case, the sources (respective SIMs) for input parameters are identified. To each of the four hyper-terminals connected on the different buses, the information about the values that the SIMs need to update in specific messages on the respective buses is sent by the CCOM. Each of the hyper-terminals 1 to 3 in turn will send the address and data blocks to corresponding SIMs on the respective 1553B buses. The SIMs will put the data accordingly in their transmit buffers for updating on the bus. Hyper- terminal 4 will update the values on the RS422 bus and set the discrete values as required to be fed to the SUT as per the test case. The output of the SUT is observed on the cockpit display surfaces and /or the RX windows of the SIMs to which the data is addressed. In the Automated approach, all the SIMs and the SUT would be working coherently in the same way as during manual testing except that the input to the SUT from the SIMs are fed without human intervention. For e.g., in order to test the SUT for computation of MACH NUMBER DATA, the 'total pressure' and 'static pressure' values have to be provided through B1_SIM4 (Simulated Air Data Computer) and Aircraft_On_Gnd_In_Air information has to be provided through B2_SIM2 (Simulated Engine System Interface Unit). This happens automatically on selection of the relevant test case from the CIT Suite.

Further, the computed MACH NUMBER DATA for the inputs fed, can be seen in the RX Window of the GUI of B2_SIM4 (Simulated Flight Control Computer) and B3_SIM5 (Simulated Fuel System Interface Unit) and on the cockpit display surfaces.

Floating point numbers are not supported by ACTS. For floating point type of values, the tool was not considering the decimal portion of the given input values for usage in the constraints defined. This limitation of the tool also had to be handled in the test harness.

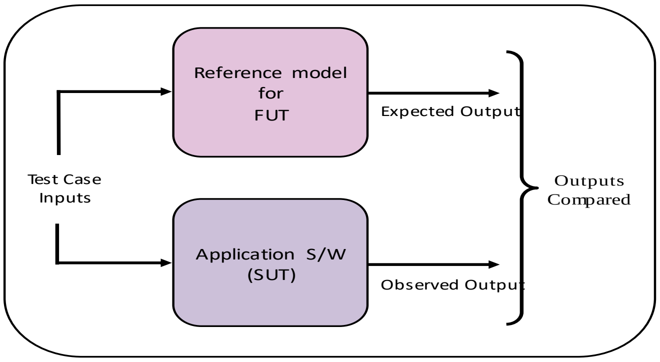


Fig. 7. Comparison of Expected and Observed Outputs for Test Report Generation.

D. Automation of Test Report Generation

The outputs from SUT captured by the bus monitors are compared with the expected output for automated test report generation. As shown in Fig-7, for a set of input values for a given test case if the corresponding outputs of the application software are same as that of the reference model, then the test is considered as PASS.

VI. Benefits of Automated Approach

- Tester need not switch to multiple terminals for providing inputs manually. Multiple Inputs from different SIMs will be provided automatically when test case is run from the CIT Suite.
- Sequence of inputs if required to be provided within a timeframe can be conveniently given as there is no human delay involved. The script has to be designed such that those input variables are placed and executed sequentially.
- The output that appears only for a short interval can also be easily verified as the checking of the output is automated.
- A particular testing scenario if need to be repeated, then it is sufficient to only re-run the test case instead of giving all the required inputs manually again.
- Regression testing for different variants of the target vehicle can be done at a faster pace and more efficiently.
- This process helps in effective CIT for RBT of MMC with typical constraints on values of inputs as well as intermediate and output parameters.
- The process of feeding the test data to the SUT automatically from the CIT SUITE and generation of Test report by comparing the output of the reference model and the actual output from the SUT reduces
- The overall testing time drastically as shown in the Table-1.

TABLE I. Comparison of Effort Involved in Manual Testing and Automated Testing

Functionality Under Test	*No. of Test cases*	*Time for Manual Execution (Mins)*	*Time for automated Execution (Mins)*
Jettisoning of selected stores	14	70	7
MARK & UPDATE Functions with different types of sensors	45	240	20
Attack Functions in different guided modes	56	560	56
Send Specific Data	60	180	10

VII. Conclusions

As the procedure involves generation of test cases by development of the simulated reference model for the functionality under test based on the detailed SRS, any ambiguity in the SRS can be reported to the authors for correction/elaboration. This helps in detailing the software requirement specifications without any ambiguity which is the main goal of RBT and hence the combinatorial test cases designed using this method will generate the most optimal test cases. Execution of these test cases shall not only provide the benefits of CIT but also provide benefits of RBT.

Automation is beneficial only if the components of the automated test rig and CITEF: SIMs, CCOM, Hyper-terminals, SFUT, CC, IVG etc. are validated for correctness. Errors in any of these components may result in the following which are not acceptable.

- FALSE NEGATIVE errors: Not detecting the errors present in the SUT which is very dangerous as errors in mission critical and safety critical avionics systems when encountered during flight can even lead to catastrophic consequences.
- FALSE POSITIVE errors: Though the implementation in the SUT is correct, this is highly undesirable as it results in waste of time in analysing and tracing the reason for the failure to a bug in the test facility.

Floating point numbers are not supported by ACTS. For floating point type of values, the tool was not considering the decimal portion of the given input values for usage in the constraints defined. This limitation of the tool also had to be handled in the test harness.

VIII. Scope for Future Work

Floating point numbers in constraints are not supported by covering array generation tools. There is scope for further work in development of tools for handling floating point data type. The automation test facility can be enhanced for optimisation of test cases in which multiple Functionalities can be tested together instead of sequential execution.

Acknowledgment

We wish to acknowledge our gratitude to the management of Aeronautical Development Agency (ADA), Bangalore for permitting publication of this paper.

References

[1] Sommerville Ian. "Software engineering", Pearson Education, Inc., publishing as Addison-Wesley, 2009.

[2] Predrag Skokovic, Marija Rakic-Skokovic, "Requirements - based testing process in practice", IJIEM International journal of industrial engineering and management, 2010, Vol. 1, p. 155-161.

[3] Yu-Wen Tung, , Wafa S Aldiwan, "Automating test case generation for the new generation mission software system" DOI 10.1109/Aero.2000.879426, 2001.

[4] Xiao Qu, Myra B Cohen, Katherine M Woolf., "Combinatorial interaction regression testing: A study of test case generation and prioritization", IEEE, 2007.

[5] Manuj Aggarwal, and Sangeeta Sabharwal, "Prioritization techniques in combinatorial testing : A survey", 1st India International Conference on Infromation Processing IICIP, 2016.

[6] Myra B Cohen, Matthew B Dwyer and Jiangfan Shi, "Interaction testing of highly configurable systems in the presence of constraints", ISSTA , 2007.

[7] Myra B Cohen and Charles J Colbourn, "Constructing test suites for interaction testing", IEEE 25th International conference on software engineering ICSE 03, 2003.

[8] Redge Bartholomew, "An industry proof-of-concept demonstration of automated combinatorial test", IEEE, 2013 p. 118 to124.

[9] Angelo Gargantini, Justyna Petke, and Marco Radavelli, "Combinatorial interaction testing for automated constraint repair", 10th IEEE International conference on software testing, verification and validation workshops, 2017.

[10] Aymeric Hervieu, Benoit Baudry and Arnaud Gotlieb, "Pacogen: Automatic Generation of pairwise test configurations from feature models". Proceedings of international symposium on software reliability engineering (ISSRE'11) Nov 2011, Hiroshima, Japan. 2011. <hal-00699558>.

[11] Dusica Marijan, Arnaud Gotlieb, Sagar Sen and Aymeric Hervieu, "Practical pairwise tesing for software product lines" SPLC 2013, Tokyo, Japan 2013 <hal-00859438>.

[12] Prasad Bokil, Priyanka Darke and Ulka Shrotri, "Automatic test data generation for C programs", Third IEEE international conference on secure softwawe integration and reliability improvement. 2009

[13] Dong Li, Linghuan Hu, Ruizhi Gao, W Eric Wong, D Richard Kuhn, and Raghu N Kacker, "Improving MC/DC and fault detection strength using combinatorial testing", IEEE International conference on software quality, reliability and security, 2017, p 297 to 303.

[14] Mehra N Borazjany, Linbin Yu, Yu Lei, Raghu Kacker and, Rick Kuhn, "Combinatorial testing of ACTS : A case study", IEEE Fifth international conference on software testing, verification and validation , 2012, p. 591 to 600

[15] https://www.softwaretestinghelp.com. Practical software testing, June 2018.

[16] Mats Grindal, Jeff Offutt, and Sten F Andler, "Combination testing strategies : A survey", GMU Technical report ISE-TR-04-05, July 2004.

[17] Mats Grindal, Bitgitta Lindstrom, Jeff Offut and Sten F, Andler, "An Evaluation of combination strategies for test case selection" GMU Technical report, 2006-10-06.

[18] Sunint Kaur Khasla and Yvan Labiche, "An extension of category partition testing for highly constrained systems", IEEE 17th International symposium on high assurance systems engineering, 2016 p. 47 to 54.

[19] Bestoun S Ahmed, Kamal Z Zamli, Wasif Afzal, and Miroslav Bures , "Constrained interaction testing: A systematic literature study", 2017, IEEE Access, DOI 10.1109/Access 2017.2771562 Vol. 5, p. 25706 to 25730.

[20] Sunint Kaur Khalsa and Yvan Labiche,"An Orchestrated survey of available algorithms and tools for combinatorial testing", Research Gate, 2014.

Smart Surveillance System using Background Subtraction Technique in IoT Application

Norharyati binti Harum[1], Mohanad Faeq Ali[2], Nurul Azma Zakaria[3], Syahrulnaziah Anawar[4]
Centre for Advanced Computing (C-ACT), Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100
Durian Tunggal, Melaka, Malaysia

***Abstract*—This paper presents a development of a security system based on Internet-of-Things (IoT) technology, where an IoT device, Raspberry Pi has been used. In the developed surveillance system, a camera works as a sensor to detect motion, and automatically capture the video of the view of area where the motion is detected. The motion is detected by image processing techniques; background subtraction technique. The technique is applied by comparing two different captured images using Pi NoIR camera. The system can be controlled from anywhere using Telegram application, and users will receive alert message with video using the application. The user can also play a siren from anywhere once detecting suspicious object can access images and videos using Telegram application. This can frighten the thief if the crime is suspected in home or office. Users can also deactivate and activate the system from anywhere at any time using the Telegram. The functionality tests have been done to ensure the developed product can work properly. Besides, tests to identify a suitable video length to be transmitted to the user and to identify the adequate location of the security in order to minimize false detection as well as false alert have been performed. The project is an IoT-based which significantly in line with the Industrial Revolution 4.0, supporting the infrastructure of Cyber-Physical System.**

Keywords*—Internet of things; raspberry pi; motion detection; home security system; surveillance system: ir4.0

I. Introduction

Nowadays, Internet-of-Things (IoT) application has become a hot topic for its capability to connect devices, enable people to reach any data using IoT platform. IoT buiding block consist of devices that acts as a data processing agent, sensors to detect any, connectivity to enable data transmission between devices, APPs that can help people to reach the data forwarded from the devices and cloud that can help big data management.

On the other hand, closed-circuit television (CCTV) camera is commonly used in houses, companies and organizations as a security system to prevent criminal. The current CCTV system is costly and not suitable for normal resident [1]. Although surveillance camera records video and helps the authorities to identify the cause of an incident such as crime or accident, it is just a passive monitoring device [2]. Most of the CCTV systems in the market provide a system that can only view the current image, and if the system is equipped with alarm system, the sirens will be activated once the intruders get in into the house. This system is not suitable for unattended house where, the intruders are able to deactivate the alarm. The system also can cause false alarm, where the alarm can be activated because of some unrelated motion. The frequent false alarm might cause inconvenience to the neighborhood and real alarm might be neglected. Furthermore, most of the CCTV fields focus on high-definition to record video which cost very high. Besides that, CCTV that comes with notification through SMS or email, the cost is extremely high which unreliable that for home uses. Due to these factors, this project will implement CCTV come with notification services, low cost and easy to manage video and pictures.

Raspberry Pi that has been introduced by Eben Upton, where it is a cheap but with a high mobility microprocessor [3], is one example of high potential IOT device, where it enables a Machine-to-Machine Communication using IEEE 802 standard. The developments of CCTV system using Raspberry Pi have also been done in [4] and [5]. In [4], the developed CCTV provides security with low cost smart security camera with night vision capability using Raspberry and Open Source Computer Vision (OpenCV). The system was designed to be used inside a warehouse facility. It has human detection and smoke detection capability that can provide precaution to potential crimes and potential fire. The problem is it has a higher delay to process the image due using a lot of function, and Raspberry Pi has limited capabilities in processing many function, including sound player. Paper [5] introduces low-cost non-intrusive sensor that can collect traffic data using a Raspberry Pi single board computer. In [6] a motion sensor, or known as PIR, "Passive Infrared", "Pyroelectric", or "IR motion" sensors are used in CCTV to detect any suspicious motion. Even these CCTVs are suitable for home and office security, but they require sensors that are sensitives and need to be maintained or exchanged frequently.

In this paper, motion detection using image processing technique [7][8] has been used where the camera in parallel performs video capturing and also motion detection. Image information acquired by Raspberry Pi HD camera module is analyzed for moving objects presence. After evaluation of detected object count, size, class and motion vector object of the properties are sent to server node by RF transceiver. This project however for traffic monitoring, which is not suitable for offices and home CCTV system. In this paper, a prototype of an IoT based security system using energy efficient microprocessor, known as raspberry Pi, using camera as motion sensor and Telegram as an IoT APP, used by a user to control the CCTV from anywhere has been developed. Our developed CCTV system can detect intruders using motion sensor, capture video of the intruders and send the video to the owner. The developed CCTV can act as a silent alarm to notify owner about the intruders. The received video can be used as a proof of the criminal activity for investigation. This paper is organized as follows. Section 2 describes the prototype design used to develop the IoT based security system. Section 3 introduces implementation stage used throughout this paper. The description of testing stage and discussion is shown in Section 3, followed by conclusion in Section 4.

II. Methodology

Rapid Application Development model is applied to develop the system as shown in Figure 1. The development process goes through the requirements planning phase, user design phase, construction phase and cutover phase.

- Requirements Planning Phase–In this phase, problems that occur among users such as working persons, building owners or premise security company is analysed to determine adequate solution/modules that might help them in ensuring the safety of the corresponding building. The hardware and software required for the development are also identified in this phase.
- User Design Phase–In this phase, the monitoring system is designed based on the information required and solution determined in the previous stage.
- Construction Phase–In this phase, the system based on design in the user design phase is developed. Early tests to ensure functionality of the system has been done.
- Cutover Phase–In this phase, the functionality of the system is improved based on testing in the previous stage. The overall tests for the developed system is then finalized.

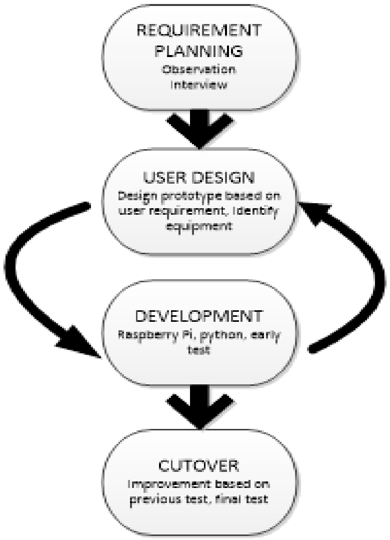

REQUIREMENT PLANNING
Observation
Interview
USER DESIGN
Design prototype based on user requirement, Identify equipment
DEVELOPMENT
Raspberry Pi, python, early test
CUTOVER
Improvement based on previous test, final test

Fig. 1. Rapid Application Development Phase.

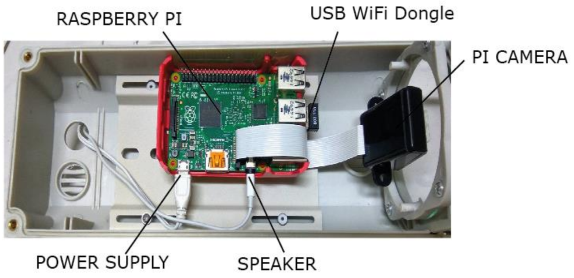


Fig. 2. Physical Design of Developed Ubiquitous Security Camera.

A. Prototype Design

Figure 2 shows the overall concept of this project. The system consists of Raspberry Pi 2, Pi NoIR camera, speaker, router (network connection) and computer (database to store pictures and videos). This security camera will capture pictures and record videos when detecting motion in coverage area and sending the notification to user through Telegram apps. Users receive pictures and videos when the phone is online or connected to internet. If a suspicious activity detected, user can send a command to play sound to frighten thief.

The Raspberry Pi is utilized as a microprocessor to enable the developed security camera sending captured picture and video via Telegram Application to user. The motion detection is configured at Raspberry Pi [9][10][11] which just using Pi NoIR camera. The picture and video is captured with pi camera when triggered by any motion. Python Programming code is used to synchronize the camera and sending notification (captured picture and video) to user through Telegram Application.

Figure 3 shows the physical design for Ubiquitous Security Camera using Raspberry Pi project. The camera is connected to the camera port in the Raspberry Pi board. USB WiFi dongle placed in the USB port and the power supply for the Raspberry Pi were connected in the power port. The 3.5-mm jack port is connected to a speaker which is used for playing police siren sound.

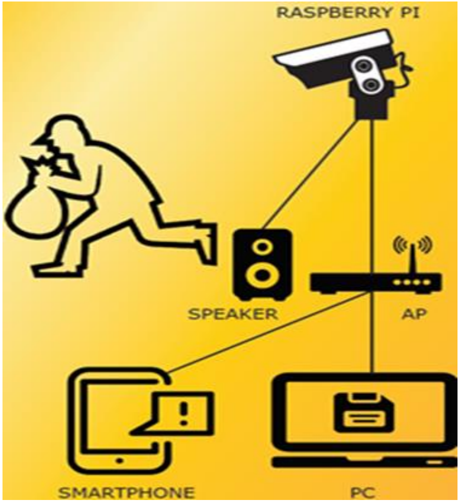


Fig. 3. Concept Diagram of Developed Ubiquitous Security Camera.

When the security camera is activated, it will wait for motion detection. Once the motion detection is detected, it will capture the picture and record video. The captured pictures and videos will be sent to server. Raspberry Pi will check whether the activation notification is activated or deactivated before send the captured images and video. If notification function is activated, the system will send the captured image and video to Telegram.

B. Motion Detection Technique

As discussed in previous section, motion detection is performed by using Background Subtraction Method or Foreground Detection as shown in [8]. The motion can be detected by the difference of image intensity between two consecutive frames. The image subtraction can be represented as:

$$\Delta I\ (i, j) = ICurrent(i, j) – IPrevious(i, j)$$

Where ΔI(i,j) is the difference in image intensity between two consecutive frames. ICurrent(i,j) and IPrevious(i,j) represent image intensities for current and previous frames or background frame respectively. This technique combines image processing and computer vision fields, where an image foreground is extracted for further processing (object recognition etc.). The background subtraction technique is used to detect the moving objects in videos captured by a static camera. The main task in this approach is that of detecting the moving objects from the difference between the current frame and a reference frame or background image. Figure 4 shows the concept diagram of Background Subtraction Method. Gaussian Mixture-based Background/Foreground Segmentation Algorithm has been used to detect motion as shown in Figure 5.

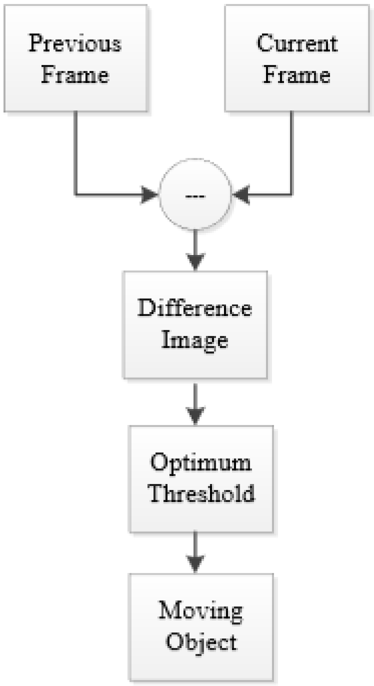
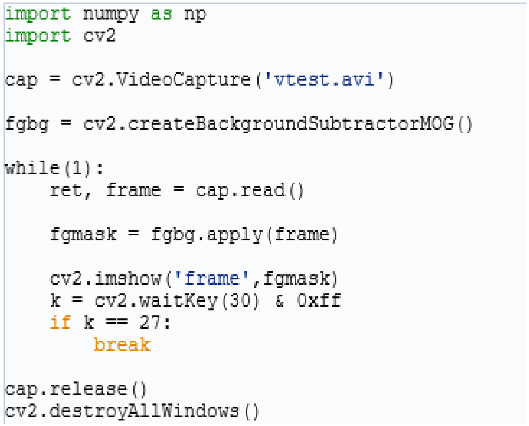


Fig. 4. Concept Diagram of Background Subtraction Technique for Moving Object Detection.



```
import numpy as np
import cv2

cap = cv2.VideoCapture('vtest.avi')

fgbg = cv2.createBackgroundSubtractorMOG()

while(1):
    ret, frame = cap.read()

    fgmask = fgbg.apply(frame)

    cv2.imshow('frame',fgmask)
    k = cv2.waitKey(30) & 0xff
    if k == 27:
        break

cap.release()
cv2.destroyAllWindows()
```

Fig. 5. Gaussian Mixture-based Background/Foreground Segmentation Algorithm.

C. Embedded Functions

In this section, functions need to be embedded in the developed security camera are discussed. As mentioned in previous section, this prototype consists of Raspberry Pi as CCTV and Telegram to control the CCTV from anywhere. Telegram that is installed in a smartphone is used, since most of the people todays will have at least one smartphone, so that ubiquitous concept can be applied, where users can access the security camera from anywhere at any time.

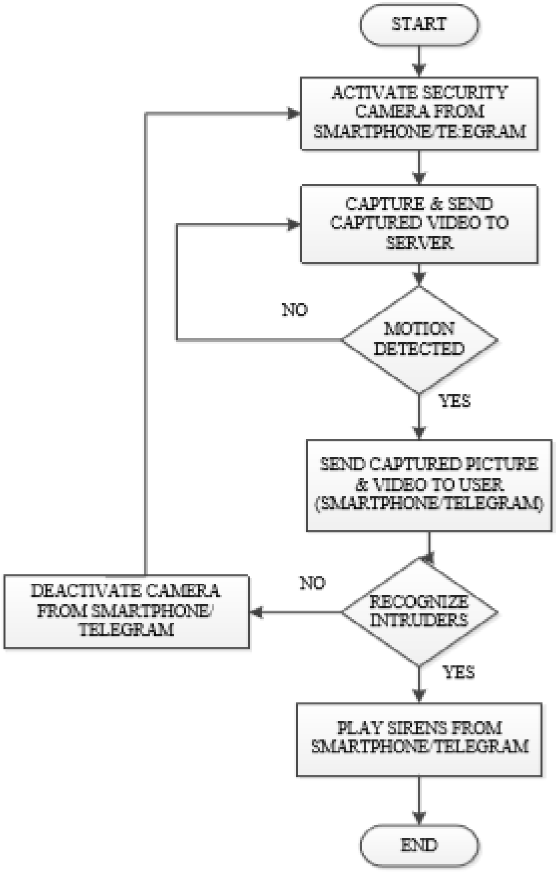
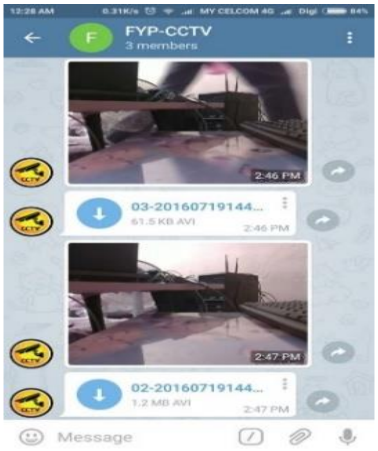

START
ACTIVATE SECURITY CAMERA FROM SMARTPHONE/TE:EGRAM
CAPTURE & SEND CAPTURED VIDEO TO SERVER
NO
MOTION DETECTED
YES
SEND CAPTURED PICTURE & VIDEO TO USER (SMARTPHONE/TELEGRAM)
NO
DEACTIVATE CAMERA FROM SMARTPHONE/ TELEGRAM
RECOGNIZE INTRUDERS
YES
PLAY SIRENS FROM SMARTPHONE/TELEGRAM
END

Fig. 6. Flowchart of the developed security camera.

The system consists of IoT devices, which are raspberry pi and smartphone. The raspberry pi is used to detect motion using captured video by embedded camera and to send the captured video to the server. The raspberry pi also enables the system to send a notification to users via Telegram that is installed in the smartphone. The system can be remotely controlled by a user using Telegram application, thus allows user to access the system from anywhere. The remotely control function includes activate and deactivate the system, remotely play sirens from the security camera if necessary and stop the sirens from if false alarm occurs. The Telegram also allows multiple monitoring of the system, so that the system can be easily used in home monitored by a family, and office monitored by responsible workers. Figure 6 shows flowchart describing operation of the embedded functions.

III. Result And Discussion

This chapter discusses about testing methods of the project. This phase has three types of tests which include functionality test, trade off test and false detection test. Functionality test is done to ensure all embedded functions in the developed product works properly. For trade off test, a number of experiments are done to identify suitable video length to ensure that the user can receive sufficient information from the received video without delay in reception. Next, time limit video testing shows the result of the most suitable time limit to capture the video for CCTV. False alert testing shows the most suitable direction to put the CCTV which to minimize false alert occur



FYP-CCTV
3 members

2:46 PM

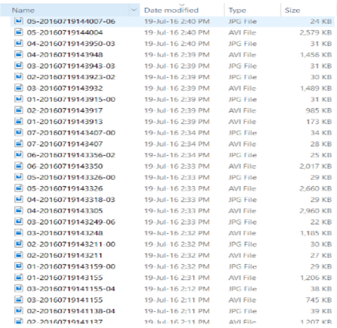
03-20160719144...
61.5 KB AVI 2:46 PM

2:47 PM

02-20160719144...
1.2 MB AVI 2:47 PM

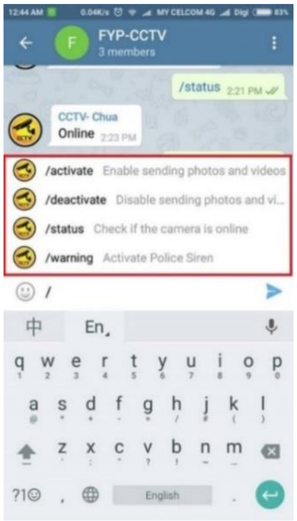
Message

Fig. 7. Alert Message when a Motion Detected.



Name	Date modified	Type	Size
05-20160719144007-06	19-Jul-16 2:40 PM	JPG File	24 KB
05-20160719144004	19-Jul-16 2:40 PM	AVI File	2,579 KB
04-20160719143950-03	19-Jul-16 2:40 PM	JPG File	31 KB
04-20160719143948	19-Jul-16 2:39 PM	AVI File	1,456 KB
03-20160719143943-03	19-Jul-16 2:39 PM	JPG File	31 KB
02-20160719143923-02	19-Jul-16 2:39 PM	JPG File	30 KB
03-20160719143932	19-Jul-16 2:39 PM	AVI File	1,489 KB
01-20160719143915-00	19-Jul-16 2:39 PM	JPG File	31 KB
02-20160719143917	19-Jul-16 2:39 PM	AVI File	985 KB
01-20160719143913	19-Jul-16 2:39 PM	AVI File	173 KB
07-20160719143407-00	19-Jul-16 2:34 PM	JPG File	34 KB
07-20160719143407	19-Jul-16 2:34 PM	AVI File	28 KB
06-20160719143356-02	19-Jul-16 2:34 PM	JPG File	25 KB
06-20160719143350	19-Jul-16 2:33 PM	AVI File	2,017 KB
05-20160719143326-00	19-Jul-16 2:33 PM	JPG File	29 KB
05-20160719143326	19-Jul-16 2:33 PM	AVI File	2,660 KB
04-20160719143318-03	19-Jul-16 2:33 PM	JPG File	29 KB
04-20160719143305	19-Jul-16 2:33 PM	AVI File	2,960 KB
03-20160719143249-06	19-Jul-16 2:33 PM	JPG File	22 KB
03-20160719143248	19-Jul-16 2:32 PM	AVI File	1,185 KB
02-20160719143211-00	19-Jul-16 2:32 PM	JPG File	30 KB
02-20160719143211	19-Jul-16 2:32 PM	AVI File	27 KB
01-20160719143159-00	19-Jul-16 2:32 PM	JPG File	29 KB
01-20160719143155	19-Jul-16 2:31 PM	AVI File	1,206 KB
03-20160719141155-04	19-Jul-16 2:12 PM	JPG File	38 KB
03-20160719141155	19-Jul-16 2:11 PM	AVI File	745 KB
02-20160719141138-04	19-Jul-16 2:11 PM	JPG File	39 KB
02-20160719141137	19-Jul-16 2:11 PM	AVI File	1,207 KB

Fig. 8. The Saved Captured Pictures and Videos.



FYP-CCTV
3 members

/status 2:21 PM

CCTV- Chua
Online 2:23 PM

/activate Enable sending photos and videos
/deactivate Disable sending photos and vi...
/status Check if the camera is online
/warning Activate Police Siren

/

Fig. 9. Commands in Telegram Application.

D. Functionality Test

The functionality test discussed in this subsection includes motion detection test, FTP service test, and Telegram service test. When the CCTV is boot, the motion detection camera will ready to detect motion, Windows file sharing ready to save the captured pictures and videos, the alert message through Telegram application will ready to send captured videos and pictures. To test motion detection function, a movement around the camera has been done and then wait for the captured pictures and videos. Figure 7 shows the result of alert message when a motion detected. Figure 8 shows the Windows file sharing that save captured pictures and videos in recording folder. From Figure 7 and 8, it is shown that the motion detection function using camera and FTP services work properly in this developed prototype. Then, all functions that enable remote access in Telegram applications are tested. The functions include /activate (Enable sending photos and videos), /deactivate (Disable sending photos and videos), /status (Check if the camera is online) and /warning (Activate Police Siren) as shown in Figure 9.

E. Trade Off: Video Loading Time vs Accuracy

In this subsection, testing process to determine suitable captured video length is discussed. The suitable captured video length is determined by considering tradeoff between video loading time and accuracy of the captured video image. The accuracy of the captured image is essential to ensure the received video can provide adequate information in order to determine any suspicious motion

The time limit video testing is done to determine the most suitable time to limit the length of videos. The captured videos must have enough information with the time limit of videos. At the same time, the videos must be transferred in adequate time length. In this testing, the maximum time limit of the video to 5 seconds (Scenario 1), 10 seconds (Scenario 2) and 20 seconds (Scenario 3) have been set. Each test is done in 10 times and average time for each test is calculated. The best decision for the maximum time out based on the information of the captured videos and the time to send the alert message.

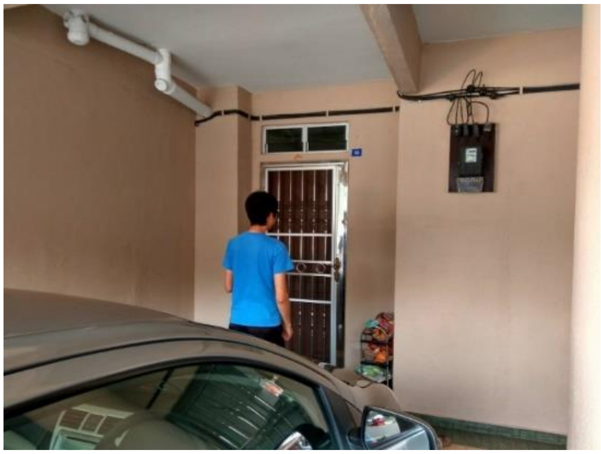


Fig. 10. Maximum Time Limit: 5 sec.

The captured video within 5 seconds is shown in Figure 10. Note that the video is captured once motion is detected. From Figure 10, it can be seen that within 5 seconds, the camera only captures a person that stand close to the door only. Any suspicious motion from the video cannot be determined. However, by setting the length to 5 seconds, the user enables to receive alert is just within 8 seconds.

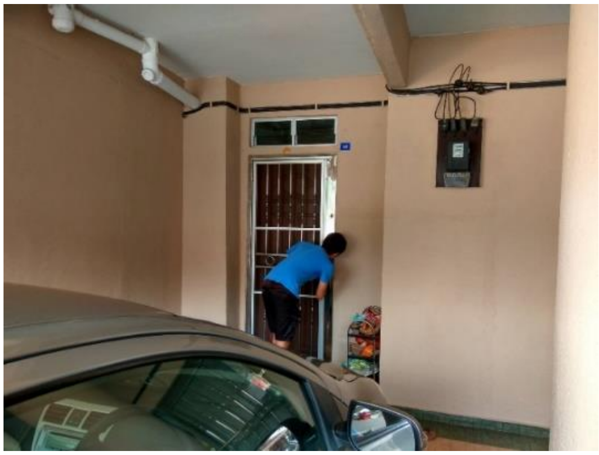


Fig. 11. Maximum Time Limit: 10 sec.

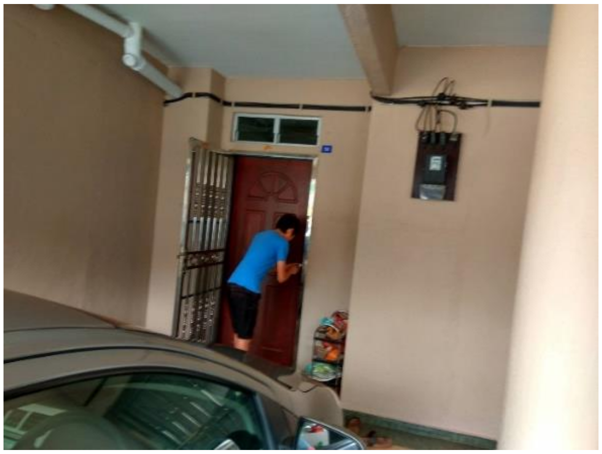


Fig. 12. Figure 12. Maximum Time Limit: 20 sec.

The captured video within 10 seconds is shown in Figure 11. Within 10 seconds which capture a person that try to open the door. With this captured videos or pictures, we can have sufficient information that this person is robber and surveillance activity is happening. The time to receive alert is within 13 seconds. The captured video within 20 seconds is shown in Figure 12. Within 20 seconds, which capture a person that already open first door and going to break second door. The problem occurs that the information for this surveillance activity to sending alert message is too slow. The time to receive alert is within 25 seconds which taking longer time because of the size of captured videos is bigger.

Table 1 and Table 2 shows testing results to identify an adequate video length, to ensure that video length can provide sufficient information in video received by a user, without causing delay in alert reception. From the results, 10 seconds video length can be concluded as sufficient because it can provide sufficient information with speedy alert transmission using Telegram.

TABLE I. RECEIVED ALERT TIME VS VIDEO LENGTH

TEST	**Alert received time (s): 5s video length**	**Alert received time (s): 10s video length**	**Alert received time (s): 20s video length**
2	8	13	24
3	8	14	26
4	7	13	25
5	7	13	25
6	8	14	26
7	8	13	25
8	9	14	25
9	8	14	26
10	7	13	26
Average	7.7	13.4	25.3

TABLE II. VIDEO LOADING TIME VS INFORMATION QUALITY

Video Length (sec)	**5**	**10**	**20**
Information	Insufficient	Sufficient	Sufficient
Average time (sec) for alert messages	7.7	13.4	25.3

F. *Minimizing False Detection*

This test is done to identify an optimal location for CCTV and to minimize the false alert message to user. Since the motion detection is done using camera, any object with motion can be detected as intruders. This will cause reception of false alert/warning message by a user. For the testing, CCTV with two directions; one is facing the CCTV to gate door shown in Figure 13 and the second test is done by facing the CCTV to main door shown in Figure 14 are setup.

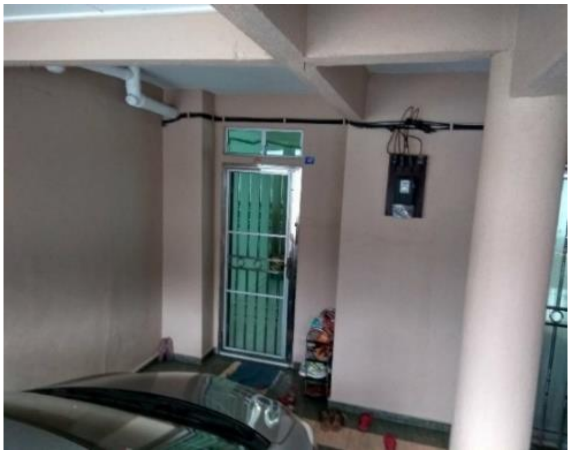


Fig. 13. Security Camera Facing Main Door.

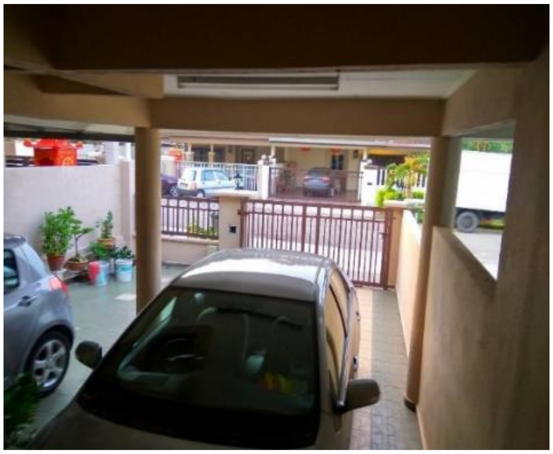


Fig. 14. Security Camera Facing Gate.

The detail of CCTV location in Figure 13 and 14 is setup as in Figure 14 and Figure 15, respectively. From Figure 14, we can see that the area of detect motion is very large if the CCTV facing gate door direction. This will get a lot motion detection because in front of gate door will have a lot of moving car, animal (dog, bird and cat), people in jogging and others. This direction to put this CCTV is not suitable because of the area detect motion is huge and false alert message will occur.

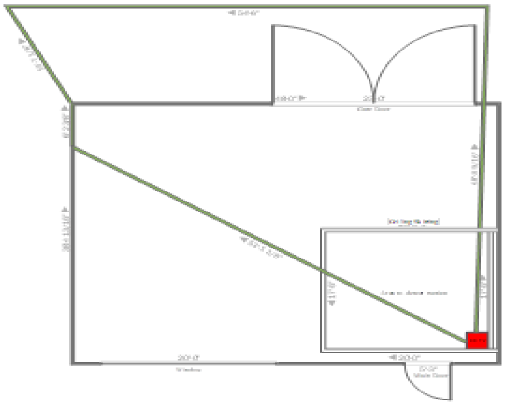


Fig. 15. Security Camera Facing Gate.

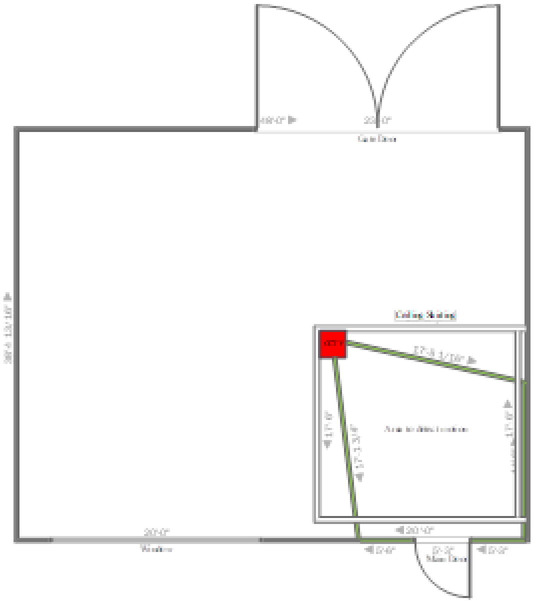


Fig. 16. Security Camera Facing Main Door.

From Figure 15, we can see that when the CCTV is located facing the main door, the camera view becomes narrower, thus can prevent the camera to detect motion of unrelated activities such as moving car in road, animal moving around, people in jogging and others. This direction only monitors the area that is main door which can be focused to moving object in front of the main door.

IV. Conclusion

In this paper, a development of IoT based security system, which can be accessed from anywhere at any time has been presented. The system uses image processing technique known as background subtraction to detect moving object. The security system includes a CCTV with a camera that can role as a sensor to detect motion, and automatically capture the video of the view of area where the motion is detected. The developed CCTV can be controlled from anywhere using Telegram application, and users will receive video using the application. The user can also play a siren from anywhere once detected suspicious object in front of the CCTV. The captured videos can be stored in Windows file sharing. It has been proved that our developed security system work properly by doing functionality test in previous subsection. Furthermore, we have found that the best time limit for capturing videos is 10 seconds because it captures sufficient information for a user to identify suspicious activities and it is an adequate value to avoid delay in receiving video through Telegram. Several tests to identify the suitable location and direction for the CCTV have been done. From the test, it can be conclude that the CCTV should be installed facing the main door to minimize the false alert message to user.

Acknowledgment

The authors would appreciate UTeM Zamalah Scheme. This research study is supported by Universiti Teknikal Malaysia Melaka (UTeM), to continue second author's study under UTeM Zamalah Scheme.

References

[1] M. Rouse (2016), Definition CCTV, [Online]. Available: FTP:http://whatis.techtarget.com/definition/CCTV-closed circuit-television.

[2] W.F. Abaya, J. Bassa, and M. Sy, “Low Cost Smart Security Camera with Night Vision Capability Using Raspberry Pi and OpenCV,” IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management 2014.

[3] C. Severence, “E. Upton:Raspberry Pi,” IEEE Computer Magazine, Vol. 46, Issue. 10, pp.14-16, 2013..

[4] M. Kochlan, “Wireless Sensor Network for Traffic Monitoring using Raspberry Pi Board” IEEE Computer Society India Symposium 2014.

[5] L. Ada, PIR Motion Sensor [Online]. Available: FTP: https://learn.adafruit.com/pir-passive-infrared-proximity motion-sensor/overview

[6] N. Yang, “Motion Sensor and Camera Placement Design for In-home Wireless Video Monitoring Systems”, IEEE Globecom 2011,

[7] D. S. Suresh and M. P. Lavanya, “Motion Detection and Tracking using Background Subtraction and Consecutive Frames Difference Method,” International Journal of Research Studies in Science, Engineering and Technology Vol. 1, Issue 5, August 2014, pp. 16-22.

[8] S. Joardar, A. Chatterjee, and A. Rakshit, “A Real-Time Palm Dorsa Subcutaneous Vein Pattern Recognition System Using Collaborative Representation-Based Classification” IEEE Transactions on Instrumentation and Measurement, Vol. 64, No. 4, April 2015.

[9] S. V. Gawande and P. R. Deshmukh, “Raspberry Pi Technology,” International Journal of Advanced Research in Compute Science and Software Engineering, Vol.5, No.4, April 2015.

[10] S. Singh, P. Anap, Y. Bhaigade, and J.P.Chavan, “IP Camera Video Surveillance using Raspberry Pi,” International Journal of Advanced Research in Computer and Communication Engineering , Vol.4, No. 2, February 2015

[11] D. Aishwarya and J. A. Renjith, “Enhanced Home Security Using IOT and Raspberry Pi,” International Research Journal of Engineering and Technology, Vol. 4, No.4 April 2017.

The Degree to which Private Education Students at Princess Nourah Bint Abdulrahman University have Access to Soft Skills from their Point of View and Educational Body

Saeb Kamel Allala[1], Ola Mohy Aldeen Abusukkar[2]
Department of Special Education,
Faculty of Education Princess Nourah bint Abdulrahman University
Saudi Arabia

***Abstract*—The study aimed at identifying the degree of ownership of special education students in the Department of Special Education, Faculty of Education, Princess Nourah University for soft skills from their point of view and the consideration of the educational body and its relation to some variables (level of study, specialization, teaching experience). The study consisted of (26) faculty members in the Department of Special Education, and the second consisted of (287) female students of the Department of Special Education at different levels and specializations, and the data using the Statistical Package for Social Sciences analysis (SPSS). The results of the statistical analysis of the study data indicated that the total score of students' possession of soft skills according to their point of view was low, except for some of the paragraphs in which the estimate was high. The degree of possession of soft skills from the point of view of faculty members was also low, and the results of the study also indicated that there were differences between the students due to specialization and the level of the school tend to favor the higher level of the study, while there were differences according to the view of the faculty for the higher experience, while the results did not indicate a difference attributed to R faculty members. The study recommended increasing the interest in soft skills for female students in particular, and for female students in general, by including these skills in the study courses and through the various student activities.**

Keywords*—Soft skills; special education students; faculty members; Princess Nourah Bint Abdulrahman University

I. Introduction

The phenomenon of unemployment among young people, especially among graduates of higher education institutions, which is increasing in number every year, is growing. This phenomenon is accompanied by a decline in the ability of public and private institutions to absorb more workers. In response to this phenomenon, many countries reviewed their educational and training systems A new dimension that aims to raise the interest of young students and develop their attitudes. Education is one of the most decisive factors in creating opportunities for young people to lead healthy, productive and responsible lives. Investments in youth education help make them good people.

The vision of the Kingdom of Saudi Arabia 2030 indicates that "one of our commitments education that contributes to the advancement of the economy We will seek to bridge the gap between the outputs of higher education and the requirements of the labor market, develop public education, guide students to appropriate career and career options, This is done only by qualifying university graduates with so-called solid skills and soft skills. Employers prefer the employee who has solid skills as well as soft skills. Solid skills are the specific cognitive and technical skills of each discipline, oft Skills) aspects of speech, good and distinct appearance, and special queens highlighting the applicant for the job among peers [5].

The term "soft skills" has become very frequent in recent times. It refers to "those basic skills that are associated with personality, competencies and positive traits that enhance a person's relationships and functioning. These skills include communicating effectively with others, presenting ideas in a convincing and tactful manner, With conflicts, teamwork, the use of leadership behaviors that characterize his relationships with others, initiative, accept criticism and manage time effectively, work under pressure, affection for others and demonstrate good manners [1,12].

In this research, we will discuss the soft skills, their importance, their definitions and types, and the different studies they have dealt with. We will examine the degree to which the special education department has a selection of soft skills. According to the researchers, this research is the first of its kind - the least.

II. Study Questions

This study attempts to answer the following questions, which are focused on the degree of ownership of special education students at Princess Nourah University Soft Skills from their point of view and consideration of the educational body and the questions are:

- What is the degree to which the students of the special education department have a soft skill from their point of view?
- What is the degree of ownership of female students of the Department of Special Education for soft skills

from the perspective of members of the Education Authority?

- Are there statistically significant differences at the level of (0,05) degree to the degree of possession of students of the Department of Special Education for soft skills from their point of view due to the variable specialization?
- Are there statistically significant differences at the level (0,05) for the degree of possession of female students of the Department of Special Education for soft skills from the point of view of the members of the Education Council due to the variable specialization?

A. *Objectives of the Study*

This research aims to identify the degree of ownership of special education students at Princess Nourah University for soft skills from their point of view and the educational body's view. The objectives of the research are as follows:

1) The availability of soft skills in the students of the Department of Special Education / Faculty of Education / University of Nourah bint Abdulrahman.

2) Identify the degree of ownership of soft skills according to the student's specialization in the section (learning difficulties, autism, talent and mental excellence).

3) Recognition of the estimates of faculty members to the degree of soft skills for female students according to the variable of specialization faculty member.

B. *Importance of Studying*

The importance of the study is as follows:

1) Theoretical importance: Although the study plans are interested in specialized information (hard skills), they often neglect to give the soft skills they deserve. This study attempts to draw attention to the importance of these skills, and is it available to university students enough to qualify for entry into the labor market by force and form.

2) Practical importance: Proposing a standardized scale for measuring soft skills for female university students. Enrich the Arabic and Saudi Library with research on soft skills.

The results of the study will show the nature of the availability of soft skills and therefore make recommendations to take into account in the study plans and characterization of the decisions of the different courses; and to draw the attention of faculty members to the importance of these skills and the need to include them in the activities of different courses.

C. *Terminology of Study Soft Skills*

Is the personal skills that enable the individual to deal positively with those around him, and know the procedure in this research degrees possessed by students from their point of view and the view of the faculty on the scale that will prepare for this goal, and will cover the following soft skills: Connect and communicate with others, Creative thinking, critical thinking, dialogue skills, teamwork, self-management and interpersonal skills, emotional / emotional intelligence skills, teamwork skills, decision making and problem solving skills.

D. *Students of Special Education Department*

They are students enrolled in the Department of Special Education, Faculty of Education, Princess Nourah University for the academic year 1438-1439.

E. *At Different Levels of Study*

Faculty members: They are members of the teaching staff in the special education department at the Faculty of Education, Princess Nourah University from the academic year 1438-1439. The study limits are as follows:

The sample of the study was limited to students of the Special Education Department at the Faculty of Education, Princess Nourah Bint Abdul Rahman University for the academic year 1348-1349 AH

Study tool: Soft skill scale (female students, faculty members) and their psychometric characteristics, which are prepared by researchers.

F. *Theoretical Framework and Previous Studies*

When reviewing the theoretical literature, a series of studies was examined, which examined soft skills in general, those that examined soft skills and their impact on recruitment and job acquisition after graduation. The researchers did not find, to their knowledge, any of the studies that examined soft skills Especially in the field of special education, but most of the studies examined the impact of the availability of these skills on success in employment and employment, and perhaps this scarcity in addressing the special education specialization as a humanitarian specialization is increasing demand as it relates to the category of persons with disabilities, Perhaps what s S gives importance to our research and opens up horizons and encourages researchers to methods of this subject, which is relatively recent.

In this paper, we will refer to the most important studies that examined soft skills in various variables. Academic institutions have simply gone beyond thinking about the technical skills required in the labor market to new skills called Soft Skills. These skills are related to social behavior in the workplace. Perhaps the last thing the West is interested in today is what they call "Soft Skills" Which translates to soft or gentle skills, and researcher in this type of skills work in the West easily find it is a values of work urged by Islam and values that govern the relationship of the individual to the group to which he belongs in the work environment, those values that promote the concepts of integration and cooperation and interdependence , For example Ok self-Muasher (God bless the slave Simha if sold Simha if bought) And the love of the other side and transparency (against fraud) and acceptance of the other, these values that the Western world is interested in today in the academic work is one of the fundamentals of work ethics in Islam and is of great importance in academic work is often lacking in our institutions and we are the first (Said, 2011). Soft skills are modern concepts associated with sociology, particularly emotional intelligence, as they are closely related to personal and social traits, as well as to their intrinsic connection to communication and relationships. The author in [12] identified them as traits and abilities that appear in attitudes and behavior rather than knowledge or technical competence. The author in [3] identified them as a set of personality traits related to the area of communication with others in a friendly atmosphere.

Collaboration and reflect the employee's sense of well-being with the work environment. It is also related to the ability to express oneself and communicate with the digital technology that is required by everyone in our time, such as computer skills, e-mail, social networking, Attractive.

Forbes Middle East and the British Broadcasting Corporation (BBC) reported that Soft Skills is a factor that is no less important than the specialized skills that will qualify the job seeker. For Labor [7] As reported by the BBC .on 14 January 2015 that soft skills have improved productivity and reduced operating costs in Britain by at least 88 billion pounds ($ 130 billion) a year, as these soft skills complement the hard skills Are related to the knowledge, experience and ability to carry out the tasks specified in the Professional Job Description. Experts speak of the fact that the lack of softness in the labor force leads to negative consequences [1][11].

The skills that were included under the name of soft skills include: communication skills, self-management skills, negotiation / rejection skills, emotional / emotional intelligence skills, teamwork skills, decision making and problem solving skills, creative and critical thinking skills With stress, dialogue skills in the community, a combination of these skills will be selected to be the current research vocabulary. Attention to these skills stems from university graduates in general, as we live in the age of knowledge-based economy. Where the competition between countries depends on the ownership of the workforce of skills consistent with the characteristics of this era, which led to different requirements of the workforce that aspire to the advancement of this economy, and the need for individuals to have skills to enable them to live and work in the society of the age of knowledge, , Effective communication with others depends on technology, and need is increased . To have the skills to solve non-stereotypical problems and to find creative solutions to these problems, all of the above requires the rethinking of the skills that learners need to prepare them for life and work in this age.

The importance of soft skills and the role that educational institutions can play in educating learners is evident in many reports and studies from institutions and study centers. The study by Young Enterprise, which included large institutions, indicated a lack of They are interviewed for the functional skills that come to soft skills within the office and the work environment at the forefront, which led them to search for foreign competencies or establish branches of their institutions Abroad to get out of this problem and find a solution, and the Foundation is due to the lack of efficiency of intensive graduates of educational institutions on academic skills and tests, which reduces the possibility of graduation competencies with the skills required for employment. [7][6]

The importance of care and attention to these skills, which is no less important than the scientific and cognitive skills that compete in universities and colleges in the competition and attention to this aspect of knowledge, while a few universities and colleges to turn to the so-called soft skills, and this research only an attempt to attract Attention to the importance of this type of competencies to be provided in the university graduates, and as one of the objectives of the Department of Special Education in the Faculty of Education at Princess Nourah University, the preparation of qualifications for the care of people with special needs and develop their abilities and energies and investment so as to enable them to merge The positive interaction in the community, and this cannot be achieved as seen researchers only through the possession alumna soft skills that qualify them for integration And this cannot be achieved, as the researchers see only through the possession of the graduate of the soft skills that qualify for integration and positive interaction in society was this research to investigate the availability of these skills in the female students in the section or not.

III. Previous Studies

The study of [2] aims to identify the role of soft skills in the process of capturing administrative functions in the Gaza Strip by identifying the most important soft skills that affect the process of capturing administrative functions and by identifying the most important soft skills required in the administrative labor market. The researcher used the descriptive analytical method to conduct the study. The researcher used the questionnaire as a main tool in the collection of information. The study population may be from the administrative staff. The questionnaire was distributed to a random sample of 150 employees. The study reached the following results: Statistic The study also found that the dependent variable is affected by the independent variable (networking, anger management, negotiation, etc.). , Crisis management, critical thinking, self-presentation, and professionalism), in a statistically significant and statistically significant manner. The study recommended that job seekers, especially in the field of administrative functions, should develop their soft expertise, the focus is on university specialization or university level only.[10] aims to identify students' perceptions of the importance of soft skills in their education and their employment, and that soft skills play a key role in achieving a successful career through social interactions in society, The researcher used the descriptive method in his study. The study tool was a questionnaire for collecting data, which was distributed to the study sample which reached (188) students specializing in business administration from four universities in Singapore. The results of the study showed that business students in general in Singapore were aware of the importance of soft skills For work And skills that contribute to their academic performance. The most important skills they possessed were teamwork, collaboration, decision making, problem solving, time management, and critical thinking skills. The study showed that other skills needed to be improved further, The study recommended some measures to improve the skills of students, which help improve the horizons of their work from their point of view, in addition to the need of educational institutions to raise awareness among students about the importance of soft skills in research and work [9], which was titled "The Need of Students and Professionals for Soft Skills Training," aimed to emphasize the need to integrate soft skills training programs with the curriculum to highlight the desired objectives of these skills and the different teaching methods to be applied. The soft skills training program consists of developing effective communication skills, presentation skills, team management, leadership skills, and focus on personal skills such as teamwork and behavioral skills improvement. The study received feedback from students The study concluded that the importance of effective communication and personal skills in increasing the chances of employment and competition strongly in the work environment, and soft skills programs develop and improve, and that these skills are characterized by the smooth transition of students From his ambition and direction to success as a business tycoon. The study recommended that more effective initiatives be taken to develop students to take into account the soft skills relevant to each discipline.

White's study [13] aimed to demonstrate the importance of soft skills for employers in the recruitment process, showing that 60% of employers do not employ university graduates and most job applicants due to their lack of soft skills, including communications , Interpersonal skills, and critical thinking.

The study of [8], entitled The Importance of Soft Skills: Teaching beyond Academic Knowledge, aimed at understanding the importance of soft skills in the life of students during undergraduate and postgraduate years. The researcher discusses in this study that soft skills complement solid skills), Which is a prerequisite for the job, and encourages responsibility especially for soft skills and students' access to these skills as well as solid skills through the undergraduate stage, which has a significant impact on students in developing their skills in this period, The study aims to raise awareness about the importance of soft skills and encourage students to improve their skills by including these skills in combination with solid skills in the curriculum. These skills play an important role in shaping the personality of the individual. It is no less important than academic knowledge.

After this presentation of previous studies shows the importance of attention to the soft skills of university students, as these skills are no less important than the scientific knowledge (hard skills), which often reduce the interest of universities in various disciplines, where most studies linked between students possess soft skills and increase The researchers have noticed that there are no studies conducted in Saudi Arabia (within the science of researchers) looking at the importance of soft skills, and this study may be the beginning to draw attention to the importance of this subject and attention to the wider, and Especially that the Kingdom is now moving strongly towards a new vision focused on human resources in light of the decline in the prices of natural resources, especially oil, and this promotes the graduation of a generation of students with solid skills and soft together.

IV. Method and Procedures

A. *Methodology of the Study*

The researchers used the quantitative approach that is appropriate to the nature of the study and its questions. The data were entered into the SPSS program and the appropriate statistical methods were used for each of the study questions.

B. *Study Society*

The study population is composed of all the students of the Special Education Department at the College of Education, Princess Nourah Bint Abdul Rahman University for the academic year 1438-1439 AH (2017-2018). The total number of students is 3000 students divided into 3 specializations (learning difficulties, autism, And mental excellence). The study society is composed of all faculty members of the Special Education Department, Faculty of Education, Princess Nourah University. The faculty consists of (26) teaching staff, specializing in learning difficulties, autism, talent and mental excellence.

C. *The Study Sample*

The study tools were applied to a random sample of students of the Special Education Department at the Faculty of Education, Princess Nourah Bint Abdul Rahman University, from different disciplines and levels of study. The sample of the members of the teaching staff was applied to all faculty members in the special education department, Princess Nora, and table (1) build the sample members.

D. *Study Tools*

The study tool was developed by reference to the literature related to soft skills, as well as through the adoption of the views of a number of specialists working in the field of special education, and through these procedures prepared the first two tools (6) paragraphs, each of the following skills: creative thinking, critical thinking, communication and communication with others, dialogue, dealing and self-management, intelligence Emotional, Teamwork, decision making, problem solving, and time management skills. The validity of the scale was ascertained by finding signs of honesty and persistence in the following ways:

TABLE I. Shows the Sample of the Study According to the Variables of Specialization, Experience, Level of Study

Total	**Total experience**				**Faculty members**
	older than 10 years	**from 8 years to 10 years**	**older than 4 years and less than 8**	**less than 4 years**	
14	2	2	2	8	**Autism** **Learning difficulties** **Talent** **Total**
7	0	0	4	3	
5	0	0	5	0	
26	2	2	12	11	

Total		**Total academic level**					
	Fourth year	**Third year**	**Second year**	**First year**	**Students**		
175	100	51	20	4	**Autism**		
112	41	48	21	2	**Learning difficulti**		
287	141	99	41	6	**Total**		

TABLE II. Coefficient of Consistency in the Internal Consistency Method of the Equation of the Cronbach Alpha Equation to the Extent of Possessing the Soft Skills of the Sample of Female Students and Faculty Members

Kronbach Alpha	Kronbach Alpha	Paragraphs	Dimensions
Paragraphs for faculty members	Paragraphs for female students		
,723	,761	7	Creative thinking
,732	,801	7	Critical thinking
,763	,733	7	Communicate and communicate with others
,772	,732	7	Dialogue
,802	,760	7	Dealing and self-management
,783	,731	7	emotional smartness
,742	,771	7	Teamwork (Team work)
,751	,750	7	Decision making and problem solving
,740	,741	7	time management skill
.756	.753		Total

1) Honesty: (7) arbitrators with specializations in the field of special education, psychology, measurement and evaluation, and asked each arbitrator to express his opinion in the paragraphs of the scale in terms of the degree of belonging to paragraph The second dimension concerns the linguistic language proposed by the arbitrators. The second dimension is the linguistic amendments proposed by the arbitrators. The second dimension concerns the linguistic changes proposed by the arbitrators. Adoption of a standard agreement (80%) of the arbitrators to accept paragraph and accordingly have been some linguistic amendments to some of the paragraphs in the standard procedure.

2) Stability: The stability of the scale was verified in an internal consistency method by the formula of the Cronbach alpha equation, after applying it to 30 students and 9 faculty members. Table (2) establishes the stability coefficient.

The table indicates that the total value of the Cronbach Alpha for the female students was (753), which is acceptable values for the purposes of the study, and the value of Cronbach Alpha for the scale of faculty members amounted to (756), which are acceptable values for the purposes of study.

E. Correction of Study Tools

The answer scale for soft skill measures was five alternatives; the answer ladder was (strongly agree, agree, neutral, disagree, not strongly agree). The researchers gave the paragraphs five scores for the answer, which represented the strongly approved alternative, , Three responses to the soft-skill scale (63- 315), and the researchers used the statistical criterion for judging the average number of vertebrates And dimensions; so if The average values of the arithmetical averages ranged from (1 - less than 2.33), medium between (3.66 - 2.33) and high if they ranged between (3.67 - 5.00).

F. Statistical Analysis

In order to answer the survey questions, the data in the scale were abstracted and statistically analyzed using the SPSS. Statistical methods were used to answer each of the study questions. The mean averages, the standard deviations, and the T test were used to answer the question. Third, the ANOVA test and the LSD test to answer the third and fourth question.

V. Results

This study aimed at the degree of ownership of special education students at Princess Nourah University for soft skills from their point of view and the consideration of the educational body. This study came out with a set of results, as follows:

To answer the first question of this study, which states: "What is the degree to which students of the Department of Special Education for soft skills from their point of view?" Mean averages and deviations were calculated for the degree of female students' possession of soft skills and standard deviations. Table (3) shows the results of the analysis.

It is noted in Table (3) that the highest average of the students' proficiency score for soft skills from their point of view was for the following paragraphs in the third dimension: After contacting and communicating with others, the high arithmetic mean was the second: "I have the ability to use the library (4.35). On the fourth dimension of the skill of dialogue, the first, second and third paragraphs, which read respectively: "Be careful to call the most recent of his name, or his nickname that he loves" and "I commend the other side "I think the counter-attack often leads to a futile confrontation with the other side" with a high arithmetic mean. The average of these paragraphs respectively was as follows: the first paragraph was average (4.39) and the second paragraph averaged 4.37), while the arithmetic average of the third paragraph (4,10), and in respect of the seventh dimension, after the skill of teamwork (work within the team), the third paragraph, which was "Share the members of the group in setting goals and making decisions" This dimension has an average of (4.26) and in terms of the eighth dimension related to the skill of decision-making and solving Problems, the second, fifth and sixth paragraphs had a high arithmetic mean, and the text of these paragraphs, respectively, was: "I advise everyone who will be affected by the decision before it is taken", "determine the benefits and disadvantages of the decision before it is taken" and "proceed with the satisfactory decision" The mean arithmetic mean for these paragraphs respectively is (4.12), (4.03) and (4.04). The third paragraph, which read "spend time for leisure", was the highest with an average of 4.11, 9, on the skill of time management, the third paragraph, which read "make sure to spend time to entertain" is the paragraph with the highest arithmetic mean in this The average score of students was 4.11, with a mean average of 4,11. Table 3 indicates that the students' degree of soft skills from their point of view was low on the first, third, sixth and seventh dimensions and the total score, which is intermediate on the second, fourth, eighth and ninth dimensions.

TABLE III. MATHEMATICAL AVERAGES AND STANDARD DEVIATIONS OF STUDENTS' DEGREE OF SOFT SKILLS FROM THEIR POINT OF VIEW

Class	Standard Deviation	Average	Extreme	Micro	N	Paragraph
Average	.91121	3.7046	5.00	1.00	281	1
Average	.97451	3.8198	5.00	1.00	283	2
Average	.76589	4.1853	5.00	1.00	286	3
Average	.78992	4.1469	5.00	1.00	286	4
Average	.76576	4.2203	5.00	1.00	286	5
Low	3.15966	19.9024	27.00	1.00	287	d1
Average	1.16480	3.2125	5.00	1.00	287	1
Average	.79655	4.0245	5.00	2.00	286	2
Average	.86293	4.0105	5.00	1.00	287	3
Average	.74783	3.9861	5.00	2.00	287	4
Average	.79933	3.8780	5.00	2.00	287	5
Average	.75955	4.1119	5.00	1.00	286	6
Average	.71945	4.2125	5.00	1.00	287	7
Average	3.46476	27.4077	35.00	14.00	287	d2
Average	.92342	4.0209	5.00	1.00	287	1
Average	1.00992	4.1359	5.00	1.00	287	2
Average	1.04420	3.8118	5.00	1.00	287	3
Average	.81616	4.3937	5.00	1.00	287	4
Average	.87232	4.2448	5.00	2.00	286	5
Average	.88905	3.2648	4.00	1.00	287	6
High	.84876	4.3589	5.00	1.00	287	7
Low	4.80334	29.2160	68.00	14.00	287	d3
High	.85966	4.3763	5.00	1.00	287	1
High	.76905	4.3986	5.00	2.00	286	2
High	.89528	4.1056	5.00	1.00	284	3
Average	.98427	3.8049	5.00	.00	287	4
Average	1.21453	3.2648	5.00	-4.00	287	5
Average	.89871	3.9965	5.00	1.00	287	6
Average	.85689	4.0000	5.00	1.00	287	7
Average	3.98706	27.8885	35.00	11.00	287	d4
Average	1.03746	3.7038	5.00	1.00	287	1
Average	1.08906	3.8153	5.00	1.00	287	2
Average	.76563	4.0348	5.00	2.00	287	3
Average	.94927	3.9126	5.00	1.00	286	4
Average	.85110	3.7805	5.00	1.00	287	5
Average	1.03382	3.8287	5.00	1.00	286	6
Average	1.08286	3.7666	5.00	1.00	287	7
Low	4.78240	26.8153	47.00	11.00	287	d5
Average	1.06239	3.5575	5.00	1.00	287	1
Average	.90284	3.9231	5.00	1.00	286	2
Average	.89952	4.0699	5.00	1.00	286	3
Average	.87907	4.2028	5.00	1.00	286	4
Average	.98163	3.8632	5.00	.00	285	5
Average	1.24380	3.4545	5.00	1.00	286	6
Average	1.10251	3.8077	11.00	1.00	286	7
Low	4.50664	26.7840	39.00	4.00	287	d6
Average	.75068	4.1608	5.00	1.00	286	1
Average	.90855	4.0979	5.00	1.00	286	2
High	.78881	4.2622	5.00	1.00	286	3
High	.81150	4.1364	5.00	1.00	286	4
Average	.86516	4.1719	5.00	1.00	285	5
Average	.89600	3.8239	5.00	1.00	284	6
Average	.77699	4.1014	5.00	1.00	286	7
Low	4.49167	28.6132	41.00	19.00	287	d7
Average	.92382	4.1538	5.00	1.00	286	1
High	.77112	4.1259	5.00	1.00	286	2
Average	.81170	3.9720	5.00	.00	286	3
Average	.82347	3.9021	5.00	2.00	286	4
High	.79835	4.0350	5.00	2.00	286	5
High	.73735	4.0456	5.00	1.00	285	6
Average	.72500	4.2657	5.00	1.00	286	7
Average	4.30078	28.3868	37.00	19.00	287	d8
Average	.88158	3.9580	5.00	1.00	286	1
Average	1.06088	3.6748	5.00	1.00	286	2
High	.84998	4.1158	5.00	1.00	285	3
Average	.76486	4.2993	5.00	1.00	284	4
Average	.98346	3.4196	5.00	1.00	286	5
Average	.88239	4.1783	5.00	1.00	286	6
Average	1.04103	3.7552	5.00	1.00	286	7
Average	4.11096	27.2613	35.00	19.00	287	d9
Low	24.91379	242.2753	329.00	127.00	287	Total

To answer the second question of this study, which states: "What is the degree to which the students of the special education department have soft skills from the point of view of the members of the faculty?" The average and standard deviations of students' degree of soft skills and standard deviations were calculated. Table (4) shows the results of the analysis.

It is noted in Table (4) that the highest assessment of skills according to the point of view of the faculty was the second dimension, which relates to the skill of critical thinking, and the high estimate of the fourth paragraph only, which was written "students have the ability to compare different views" Paragraph 3.38, as well as the assessment of the fifth dimension related to the skill of communication and communication with others. The high assessment of paragraph 5, which read "listen carefully to the speaker", reached the mathematical average of this paragraph (3.73) VII Skill Teamwork (Action Combine "The students are seeking to find effective means of work and appropriate tools to achieve the objectives of the working group." The average of this paragraph is 3,46. Table 4 indicates that the degree of the total evaluation of faculty members for female students Overall soft skills were low, also low on the first, second, fifth, sixth, eighth and ninth dimensions, which are medium on the third, fourth and seventh dimensions. There is no high rating except for some of the paragraphs to which we referred.

TABLE IV. THE AVERAGE ASSESSMENT OF FACULTY MEMBERS OF FEMALE STUDENTS ON THE PARAGRAPHS AND DIMENSIONS AND THE TOTAL SCORE

Class	**Standard Deviation**	**Average**	**High**	**Micro**	
Average	1.0318	3.2308	5	2	1
Average	0.97744	2.6538	5	1	2
Average	1.05903	3.1923	5	1	3
Low	0.8709	2.9615	5	1	4
Average	0.97033	3.3077	5	2	5
Average	0.85934	3.5385	5	2	6
Low	0.97665	2.9231	5	2	7
Low	4.648	21.81	35	15	d1
Average	0.98684	3.4231	5	1	1
Average	0.84853	3	4	2	2
Average	0.8709	3.0385	4	2	3
High	0.75243	3.3846	4	2	4
Average	0.78446	2.8462	4	2	5
Average	0.52769	3.9615	5	3	6
Low	1.10732	2.8846	5	1	7
Low	2.901	22.54	28	19	d2
Average	1.0198	3	5	2	1
Average	1.03255	2.8846	5	1	2
Average	1.03255	3.1154	5	1	3
Average	0.80861	3.5769	5	2	4
High	0.45234	3.7308	4	3	5
Average	0.98917	3.4615	5	1	6
Average	1.2083	3.5	5	1	7
Average	4.172	23.27	30	15	d3
Average	0.71036	3.7692	5	2	1
Average	0.84943	3.8077	5	1	2
Average	0.70711	3.5	4	2	3
Average	0.91903	3.2692	5	2	4
Average	0.8339	2.8462	4	1	5
High	0.53349	3.7308	4	2	6
Average	0.96157	2.7308	4	1	7
Average	2.697	23.65	28	20	d4
Average	0.8339	3.1538	5	1	1
Average	0.88405	3.3077	5	1	2
Average	0.81146	3.4615	5	2	3
Low	0.86291	3.2308	5	2	4
Low	0.8709	2.9615	5	2	5
Average	0.9389	3.1923	5	2	6
Low	0.89529	3.1923	5	1	7
Low	4.283	22.5	35	15	d5
Average	0.97665	3.0769	5	1	1
Average	0.91568	3.0385	4	1	2
Low	0.95836	2.9615	5	2	3
Low	0.83758	3.3077	5	2	4
Low	0.78838	3.3077	4	1	5
Low	0.67937	3.6923	5	2	6
Average	0.71036	3.7692	5	3	7
Low	3.184	23.15	31	18	d6
Average	1.15825	3.3077	5	1	1
Average	0.98917	3.4615	5	1	2
Average	0.78838	3.6923	5	1	3
Average	1.04954	3.3077	5	1	4
Average	0.90469	3.4615	5	1	5
Average	0.84853	3	4	1	6
High	0.80861	3.4231	4	1	7
Average	5.161	23.65	31	7	d7
Low	0.86291	3.2308	5	2	1
Average	0.89184	3.3462	5	1	2
Average	0.86291	3.2308	5	2	3
Low	0.74936	3.1923	5	2	4
Low	0.86291	3.2308	5	2	5
Low	0.80096	3.1923	5	2	6
Average	0.94787	3.4615	5	1	7
Low	4.246	22.88	33	13	d8
Average	1.00231	3.2692	5	1	1
Average	1.03849	3.0385	5	1	2
Average	0.89786	3.6154	5	1	3
Average	0.68836	4.0769	5	2	4
Average	0.82741	3.7308	5	2	5
Average	0.96157	2.7308	5	1	6
Average	0.99305	3.1154	5	2	7
Low	4.081	23.58	32	17	d9
Low	26.386	207.04	274	157	

In order to answer the third question: "Is there a significant difference at the level of (0.05) to the extent that the students of the Department of Special Education for soft skills from their point of view are attributed to the variable of specialization and level of school? The specialization and Table (5) shows the results of the analysis.

Table (5) indicates that there is a virtual difference on the dimensions of the total score due to the specialization variable and to see if the difference D was statistically tested (T) for the difference between the averages as in the following table.

TABLE V. SHOWS THE CALCULATION OF PERFORMANCE AVERAGES BY SPECIALIZATION VARIABLE

Mean Standard Error	Standard Deviation	Average	N	Specialty	
.22510	2.97783	19.8914	175	Autism	d1
.32489	3.43836	19.9196	112	Learning difficulties	
.26494	3.50479	27.5486	175	Autism	d2
.32176	3.40517	27.1875	112	Learning difficulties	
.38231	5.05745	29.0514	175	Autism	d3
.41450	4.38663	29.4732	112	Learning difficulties	
.30547	4.04105	28.0571	175	Autism	d4
.36895	3.90455	27.6250	112	Learning difficulties	
.36186	4.78694	26.7886	175	Autism	d5
.45323	4.79650	26.8571	112	Learning difficulties	
.32658	4.32030	26.9143	175	Autism	d6
.45319	4.79609	26.5804	112	Learning difficulties	
.32594	4.31184	28.5657	175	Autism	d7
.45148	4.77801	28.6875	112	Learning difficulties	
.29669	3.92489	28.2457	175	Autism	d8
.45738	4.84044	28.6071	112	Learning difficulties	
.28433	3.76135	27.2857	175	Autism	d9
.43675	4.62214	27.2232	112	Learning difficulties	
1.87175	24.76090	242.3486	175	Autism	Total
2.38704	25.26207	242.1607	112	Learning difficulties	

TABLE VI. THE RESULTS OF THE TEST (T) FOR THE DIFFERENCE BETWEEN THE AVERAGES ACCORDING TO THE SPECIALIZATION VARIABLE

Variance of Standard Deviation	Average Variation	Significance	Degree of freedom	T	
.38301	-.02821	.941	285	-.074	d1
.41945	.36107	.390	285	.861	d2
.58172	-.42179	.469	285	-.725	d3
.48263	.43214	.371	285	.895	d4
.57971	-.06857	.906	285	-.118	d5
.54594	.33393	.541	285	.612	d6
.54443	-.12179	.823	285	-.224	d7
.52090	-.36143	.488	285	-.694	d8
.49832	.06250	.900	285	.125	d9
3.02002	.18786	.950	285	.062	total

Table (6) indicates that there is no statistically significant difference between the averages due to the specialization variable.

In order to answer the question about the effect of the variable of the academic level, to the extent that students have soft skills according to their point of view, which were: Are there statistically significant differences at the level of (0, 05) Study Analysis of the common variation of the differences between the averages by the variable of the study level was conducted as shown in Table (7).

TABLE VII. THE RESULTS OF THE ANOVA TEST ARE BASED ON THE DIFFERENCE BETWEEN THE AVERAGES BY THE VARIABLE OF THE STUDY LEVEL

Level of Significance	F VALUE	Average Squares	Degree of freedom	Total Squares		
.387	1.013	10.108	3	30.324	Between groups	d1
		9.982	283	2824.945	In the groups	
			286	2855.268	Total	
.450	.884	10.622	3	31.866	Between groups	d2
		12.019	283	3401.437	In the groups	
			286	3433.303	Total	
.462	.861	19.891	3	59.673	Between groups	d3
		23.106	283	6538.933	In the groups	
			286	6598.606	Total	
.007	4.126	63.510	3	190.529	Between groups	d4
		15.392	283	4355.903	In the groups	
			286	4546.432	Total	
.069	2.387	53.814	3	161.443	Between groups	d5
		22.543	283	6379.770	In the groups	
			286	6541.213	Total	
.295	1.242	25.158	3	75.473	Between groups	d6
		20.258	283	5733.134	In the groups	
			286	5808.606	Total	
.094	2.154	42.939	3	128.816	Between groups	d7
		19.934	283	5641.254	In the groups	
			286	5770.070	Total	
.434	.915	16.947	3	50.842	Between groups	d8
		18.513	283	5239.227	In the groups	
			286	5290.070	Total	
.010	3.821	62.721	3	188.163	Between groups	d9
		16.414	283	4645.238	In the groups	
			286	4833.401	Total	
.041	2.782	1695.052	3	5085.155	Between groups	total
		609.308	283	172434.099	In the groups	
			286	177519.254	Total	

Table (7) shows that there is a statistically significant difference in the fourth, ninth, and total dimensions. There is no statistically significant difference between the first, second, third, fifth, sixth, seventh and eighth dimensions.

LSD test (least significant differences) was performed for the post-comparisons as shown in Table (8).

Table (8) shows that the difference in the fourth, ninth, and total score tends to favor the larger school years.

The fourth answer: Are there significant differences at the level of (0,05) to the degree of possession of female students of the Department of Special Education for soft skills from the point of view of the members of the Education Council due to the variables of experience and specialization?

ANOVA was analyzed to answer the question. Table (9) analyzes the results of this question.

TABLE VIII. Shows the Distance Comparisons (LSD) of the Difference between the Averages

Level of Significance.	Standard Error	Average Contrast (I-J)	Level of Study	Level of Study	Dimension
.331	1.71485	-1.67073	Second year	First year	d4
.303	1.64948	-1.70202	Third year		
.053	1.63538	-3.18085	Fourth year		
.331	1.71485	1.67073	First year	Second year	
.966	.72862	-.03129	Second year		
.031	.69611	-1.51012*	Fourth year		
.303	1.64948	1.70202	First year	Third year	
.966	.72862	.03129	Second year		
.004	.51443	-1.47883*	Fourth year		
.053	1.63538	3.18085	First year	Fourth year	
.031	.69611	1.51012*	Second year		
.004	.51443	1.47883*	Second year		
.077	1.77089	-3.14634	Second year	First year	d9
.011	1.70338	-4.35354*	Second year		
.006	1.68882	-4.70213*	Fourth year		
.077	1.77089	3.14634	First year	Second year	
.110	.75243	-1.20719	Third year		
.031	.71886	-1.55579*	Fourth year		
.011	1.70338	4.35354*	Second year	Third year	
.110	.75243	1.20719	Second year		
.512	.53124	-.34859	First year		
.006	1.68882	4.70213*	First year	Fourth year	
.031	.71886	1.55579*	Second year		
.512	.53124	.34859	Third year		
.192	10.78946	-14.10976	Second year	First year	Total
.069	10.37815	-18.91414	Second year		
.027	10.28944	-22.86879*	Fourth year		
.192	10.78946	14.10976	First year	Second year	
.296	4.58430	-4.80439	Third year		
.046	4.37978	-8.75904*	Fourth year		
.069	10.37815	18.91414	First year	Third year	
.296	4.58430	4.80439	Second year		
.223	3.23666	-3.95465	Fourth year		
.027	10.28944	22.86879*	First year	Fourth year	
.046	4.37978	8.75904*	Second year		
.223	3.23666	3.95465	Third year		

Table (9) indicates that there is a statistically significant difference in the degree of female students possessing soft experience from the point of view of the faculty members according to the variable of experience in the first, fifth and the total degree.

TABLE IX. SHOWS THE RESULTS OF THE VARIANCE ANALYSIS FOR THE QUESTIONNAIRE ACCORDING TO THE VARIABLE OF EXPERIENCE

Level	**F**	**Average Squares**	**Freedom degree Degree**	**Total Squares**	**Source of Contrast**	**Dimension**
.012	4.625	69.624	3	208.872	Between groups	d1
		15.053	22	331.167	In the groups	
			25	540.038	Total	
.157	1.913	14.515	3	43.545	Between groups	d2
		7.587	22	166.917	In the groups	
			25	210.462	Total	
.515	.786	14.038	3	42.115	Between groups	d3
		17.864	22	393.000	In the groups	
			25	435.115	Total	
.198	1.689	11.350	3	34.051	Between groups	d4
		6.720	22	147.833	In the groups	
			25	181.885	Total	
.008	5.066	62.444	3	187.333	Between groups	d5
		12.326	22	271.167	In the groups	
			25	458.500	Total	
.548	.725	7.600	3	22.801	Between groups	d6
		10.481	22	230.583	In the groups	
			25	253.385	Total	
.063	2.819	61.628	3	184.885	Between groups	d7
		21.864	22	481.000	In the groups	
			25	665.885	Total	
.063	2.817	41.690	3	125.071	Between groups	d8
		14.799	22	325.583	In the groups	
			25	450.654	Total	
.081	2.561	35.921	3	107.763	Between groups	d9
		14.027	22	308.583	In the groups	
			25	416.346	Total	
.023	3.872	2004.598	3	6013.795	Between groups	
		517.780	22	11391.167	In the groups	
			25	17404.962	Total	

TABLE X. SHOWS THE DIFFERENCE TRENDS. LSD (LEAST SIGNIFICANT DIFFERENCES) WAS PERFORMED FOR THE REMOTE COMPARISONS

Variance	**Standard error**	**Intermediate Teams (I-J)**	**Experience j**	**Experience**	**dimension**
.150	1.82632	2.72727	From 4-8 years	Less than4 Years	d1
.173	3.29245	-4.63636	8-10 years		
.351	3.29245	-3.13636	More than10 years		
.150	1.82632	-2.72727	Less than4 Years	Less than4 Years	
.036	3.29245	-7.36364*	8-10		
.089	3.29245	-5.86364	More than 10 years		
.173	3.29245	4.63636	Less than 4	8-10	
.036	3.29245	7.36364*	4-8		
.730	4.28311	1.50000	More than 10 year		
.351	3.29245	3.13636	Less than 4 years	More than 10	
.089	3.29245	5.86364	4-8		
.730	4.28311	-1.50000	8-10		
.295	1.52453	1.63636	4-8	Less than 4 years	d5
.015	2.74839	-7.22727*	8-10		
.099	2.74839	-4.72727	ˡmore than 10 years		
.295	1.52453	-1.63636	Less than 4 year	4- 8	
.004	2.74839	-8.86364*	8-10		
.030	2.74839	-6.36364*	More than 10 years		
.015	2.74839	7.22727*	Less than 4 years	8-10	
.004	2.74839	8.86364*	4-8		
.492	3.57534	2.50000	More than 10 years		
.099	2.74839	4.72727	Less than 4 years	More than 10 years	
.030	2.74839	6.36364*	4- 8		
.492	3.57534	-2.50000	8-10		
.055	9.36817	19.00000	4- 8	Less than 4 years	
.062	16.88871	-33.18182	8-10		
.203	16.88871	-22.18182	More than 10 years		
.055	9.36817	-19.00000	Less than 4 years	4- 8	
.005	16.88871	-52.18182*	8-10		
.023	16.88871	-41.18182*	More than 10 years		
.062	16.88871	33.18182	Less than 4 years	8-10	
.005	16.88871	52.18182*	4- 8		
.622	21.97030	11.00000	More than 10 years		
.203	16.88871	22.18182	Less than 4 years	More than 10 years	
.023	16.88871	41.18182*	4- 8		
.622	21.97030	-11.00000	8-10		

It is noted in Table (10) that the difference in the comparisons for the benefit of the higher experience. To find out if there is a statistically significant difference in the degree of faculty members' assessment of soft skills due to the specialization variable.

TABLE XI. ANALYSIS OF VARIANCE OF THE QUESTIONNAIRE ACCORDING TO THE SPECIALIZATION VARIABLE

Level of significance	**F**	**Average squares**	**Degree of freedom**	**Total squares**	**Source of Contrast**	d
.458	.808	17.726	2	35.453	Between groups	d1
		21.939	23	504.586	In the groups	
			25	540.038	Total	
.268	1.394	11.374	2	22.747	Between groups	d2
		8.161	23	187.714	In the groups	
			25	210.462	Total	
.580	.558	10.065	2	20.130	Between groups	d3
		18.043	23	414.986	In the groups	
			25	435.115	Total	
.480	.758	5.621	2	11.242	Between groups	d4
		7.419	23	170.643	In the groups	
			25	181.885	Total	
.351	1.095	19.929	2	39.857	Between groups	d5
		18.202	23	418.643	In the groups	
			25	458.500	Total	
.624	.481	5.085	2	10.170	Between groups	d6
		10.575	23	243.214	In the groups	
			25	253.385	Total	
.307	1.245	32.514	2	65.027	Between groups	d7
		26.124	23	600.857	In the groups	
			25	665.885	Total	
.756	.283	5.405	2	10.811	Between groups	d8
		19.124	23	439.843	In the groups	
			25	450.654	Total	
.103	2.511	37.309	2	74.618	Between groups	d9
		14.858	23	341.729	In the groups	
			25	416.346	Total	
.399	.956	667.616	2	1335.233	Between groups	Total
		698.684	23	16069.729	In the groups	
			25	17404.962	Total	

Table (11) indicates that there are no statistically significant differences in the degree of female faculty members' assessment of soft skills due to specialization variable.

VI. DISCUSSION OF RESULTS

It should be noted here that most of the previous studies that were reviewed examined soft skills in terms of their impact on post-graduate employment, and how the lack of these skills among university graduates prevents them from getting the jobs they want in the face of great competition among graduates from The findings of our current study indicate that the female students of the special education department at Princess Nora University from their point of view and the view of the faculty members were low on the dimensions of the nine scale. This result is consistent with the results of the [8] The study also noted the need to raise awareness about the importance of soft skills and encourage students to improve their soft skills. Our current study is similar in its results with the study of [9], which noted that the importance of training students and professionals in soft skills, And personal skills, where the sample of the study indicated that the possession of skills increases the imposition of employment and competition for jobs, and the current study alerted to the importance of possessing soft skills, where the study of pilgrims (2014) to the existence of statistical significance between the capture of administrative functions and variable Will reduce the soft skills, and recommended the study of job seekers need to develop their soft skills and refined and gain the missing.

As the loss of these skills reduces their access to administrative work as the study showed, and the results of the study show that students in the Department of Special Education lack soft skills in sufficient form, and with regard to the study of Shaheen and his colleagues [10], which noted that business students in general in Singapore were aware of the importance of soft skills for work and career advancement, and that soft skills contribute greatly to their academic performance. The results of this study are one of the objectives of the current study The study recommended the need to raise awareness of the importance of these skills, and intersect the results of this study with the results of our current study, which showed that the soft skills of students are low and this [13] has shown that 60% of employers do not employ university graduates and most job applicants due to their lack of soft skills, so believe The researchers said that it is appropriate to pay attention to the results of the current study and the necessary work with regard to the inclusion in the various university courses of soft skills required, and include them in various university activities, and awareness of the importance of availability of university students [4].

RECOMMENDATIONS

In the light of the findings of the researchers, they suggest the following:

- Educating students about the importance of soft skills and the need to possess these skills in order to achieve success in practical life.

- More attention is paid to soft skills by faculty members, by including courses in various activities that focus on these skills.

- Conduct other studies on other samples of female students, so that the samples are larger, and the disciplines are different.

- Provide various activities through student clubs, through which workshops and courses and lectures on soft skills in different dimensions, to contribute to shaping the character of the student better.

- The researchers also recommend conducting follow-up studies on female graduates in particular, and the university in general, to find out the impact of the availability of soft skills or not on the chances of employment and success in their various jobs.

References

[1] Al-Jamri, Mansour (2015). Soft skills, article published in the newspaper Al-Wasat, No. 4598, Friday 10 April 2015 corresponding to 20 Jumada II 1436 AH.

[2] Hajjaj, Ola (2014). The role of soft skills in the process of capturing administrative positions, applied case study on administrative jobs in the Gaza Strip, Master Thesis, Faculty of Commerce, Islamic University, Gaza.

[3] Khamis, Abdullah (2013) Soft skills they are looking for, Vision Foundation for Press and Publishing, Amman, Jordan

[4] Saeed, Omar (2011). Principles of evaluation and classification of the universities of the Islamic world, a vision of originality, Sudan, African International University

[5] Sweilem, Faiza (2013). Soft Skills Personal qualities put their owners at the forefront of the recruitment marathon, "Vision Foundation for Press and Publishing, Amman, Jordan.

[6] Chalabi, Nawal (2014). Proposed Framework for Integrating 21st Century Skills into Science Curricula in Basic Education in Egypt, Specialized International Journal of Education, No. 10, 3.

[7] Arfaj, Maher (2-14). Soft Skills, an introduction to the College of Education, King Faisal University, Dammam, Saudi Arabia.

[8] Schulz, B. (2008). The Important of Soft Skills: Education Beyond Academic Knowledge. NAWA Journal of Language and Communication Polytechnic of Namibia.

[9] Seetha, S, (2013). Necessity of Soft Skills Training for Stu dents and Professionals. 4(2) March- may, 2013, pp. 171-174. India: International Journal of Engineering, Business and Enterprise Applications.

[10] Shaheen, M, Zhang, L, Shen, T,& Siti, (2012). Impotence of Soft for Education and Career Success, Volume 2, Issue 2. Nanyang Technological University, Singapore, International Journal for Cros - disciplinary Subject in Education (IJCDSE), special.

[11] Tobin, p. (2006). Managing ourselves leadership: Experiential Leadership: Experiential learning and leadership Development. Vol.12, pp36-42. Journal, Barthay, Ambleside.

[12] Vijayalakshmi, V. (2016). Soft Skills- The Need of the Hour for Professional competence : A Review on Interpersonal skills and Theories. Volume. 11, Number 4, pp. 2859-2864. International Journal of Applied Engineering Reseaech.

[13] White, Martha (2012). “The Real New College Grads Can’t Get Hired”.

FTL Algorithm using Warm Block Technique for QLC+SLC Hybrid NAND Flash Memory

Wanil Kim[1], Seok-Bin Seo[2], Jin-Young Kim[3], Se Jin Kwon[4]
Department of Computer Engineering
Kangwon National University
Samcheok, South Korea

***Abstract*—When applying the existing flash translation layer technique to a mixed NAND flash storage device composed of Quad Level Cell and Single Level Cell, because the characteristics of a semiconductor chip are not taken into consideration, the data are stored indiscriminately, and thus the performance and stability are not guaranteed. Therefore, this study proposes a flash translation layer algorithm using the warm block technique in a NAND flash storage device that combines a large capacity Quad Level Cell and a high performance Single Level Cell. The warm block technique avoids overloading of the read/write/erase operations in the Quad Level Cell flash memory by efficiently placing hot data that are frequently updated on a long-living Single Level Cell. It was confirmed experimentally that the lifetime extension and performance of hybrid NAND flash memory are improved using the warm block technique.**

Keywords—Quad level cell; single level cell; composed flash memory; flash translation layer

I. Introduction

Flash memory is classified into various types of semiconductor chips including Single Level Cell (SLC), Multi- Level Cell (MLC), Triple Level Cell (TLC), and Quad Level Cell (QLC) depending on the number of bits that can be stored in a single memory cell. SLC, MLC, and TLC/QLC are treated as advanced, intermediate, and TLC/QLC entry-level types, respectively. A hybrid SSD, in which various types of chips are mixed and used in a single storage device, has recently been proposed [1]. Because such a storage device has high-grade and intermediate/entry-level semiconductor chips in a single SSD, it is necessary to decide where to store the data, namely, in either the high-grade or intermediate/low-cost semiconductor chips when a write request is made from the file system. If the data classification and storage technique are inefficient, unnecessary data movement from a merge operation frequently occurs, which may degrade the overall performance.

To improve this, studies on the flash translation layer (FTL) [2] have been carried out. Most existing approaches [3] write updates to the log block. However, when such a log block-based FTL is applied to hybrid NAND flash memory [4], the characteristics of each semiconductor chip are not considered and data transfers between the high-grade and intermediate/low-cost semiconductor chips frequently occur, resulting in a degraded performance.

The algorithms in this paper are designed to minimize the erase operation of intermediate/supplied semiconductor chips and increase I/O performance of high-end semiconductor chips, which are for large storage devices based on QLC+SLC mixed NAND flash.

Section 2 describes related studies and their background. Section 3 provides the structure of the proposed method. Section 4 details the operation of the proposed method. Section 5 shows the performance test results of the proposed method as compared with the existing techniques. Finally, some concluding remarks are provided in Section 6.

II. Related Studies and their Background

Previous studies on hybrid NAND flash memories [5] have focused on the development of cold data storage on a relatively low-performance chip (TLC, MLC) depending on the characteristics of the flash memory chip, and on storing hot data on a high-performance chip (SLC).

In an existing study, where mixed NAND flash memory is applied with the technology of FAST [6], MLC is used as a data block and SLC is used as a log block. Figure 1 shows the structure of the existing technique [7]. This technique is written to the log block when the data are updated, and when the log area is full, the data blocks associated with the log block are fully merged [8], resulting in a degraded performance. Partial merges, which are relatively low in terms of computational cost, are applied in existing studies only when the sequential write condition of the data block is satisfied.

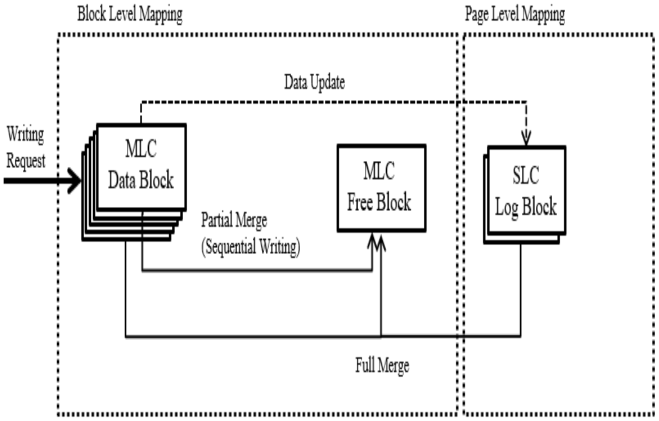


Fig. 1. Existing Log-Block based Mixed NAND Flash Storage Architecture.

Therefore, full merge operations are frequently conducted, meaning that the frequently used data in the log block are handled as cold data indiscriminately and written as MLC data block. When a conventional log block FTL technique is applied to hybrid NAND flash memory, the characteristics of a semiconductor chip are not considered, the data are indiscriminately written, and the performance and stability are thus not guaranteed. With this background, the present study proposes FTL algorithm based on the warm block technique in a NAND flash storage system that combines a high-capacity QLC and a high-performance SLC when considering the characteristics of a semiconductor chip.

III. Structure of the Proposed Warm Block Technique

This study uses large QLC NAND flash as both data and warm blocks, and uses SLC NAND flash as a log block. The proposed method uses the warm block technique with a log-block based algorithm, and has four types of physical blocks namely, a data block, warm block, hot data block, and free block.

In a conventional FTL [9], 3–5% of the data blocks are fixedly allocated as log blocks; however, this study proposes the allocation of 3% log blocks and 2% warm blocks.

The warm block exists between the data and log blocks. On the other hand, in the existing log-block based FTL algorithm, the updated data are written to the log block immediately when the data in the data block are updated. The proposed method regards the data updated from the data block as warm data before writing to the log block, and stores the updated data according to the offset after the warm block allocation of the QLC.

When the data in the data block are updated, as shown in Figure 2, the updated data are written to the warm block according to the offset of the corresponding logical address.

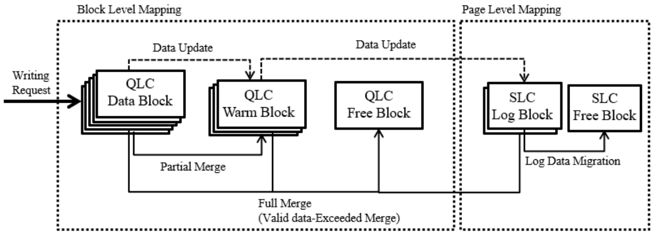


Fig. 2. Mixed NAND Flash Storage Architecture Applied using the Proposed.

This implies that partial merges will be possible in the future regardless of whether the write requests are sequential or arbitrary. Therefore, unlike the existing technique, which can apply partial merge only when the sequential writing condition is satisfied, the proposed technique can be partially merged between the data and warm blocks under any situation.

When the data in the warm block are updated, the data are treated as hot data, and the updated data are written in the SLC log block. The data treated as hot data are maintained in the SLC log block for a long period of time. This occurs because the hot data are located in the log block as long as possible through a self-applied garbage collection in the log area owing to the presence of invalid data in the log block.

If there are no invalid data in the log area, then a full merge operation will take place only then. This means that all log blocks in the log area are filled with valid data, indicating that the space utilization of the log area has already been maximized.

The merge operations of the proposed technique are explained in detail in Section 4.

IV. Operation of the Proposed Warm Block Technique

A. Data-Warm Partial Merge

If the data in a data block are updated, and an allocation to a warm block is required but there are no more free blocks to allocate, Data-Warm partial merge is conducted by selecting the block with the smallest merge operation cost as a victim block. If there are no valid data in the selected victim block, it is possible to execute a switch merge that has the least computation cost during the merge operation.

Switch merges and partial merges of the log-block based existing techniques can be conducted only on sequential data of log blocks handled by page mapping; however, because the warm block data are stored at a position depending on the offset regardless of the writing request type, sequential or random, the proposed method can switch and partially merge under any type of situation.

The Data-Warm partial merge in Figure 3 shows that the D1 and F1 data in the data block apply a “copy and write” operation using the D1' and F1' data of the warm block. Next, the warm block is replaced with the data block after the erase operation of the data block occurs. In this case, the erase operation occurs only once.

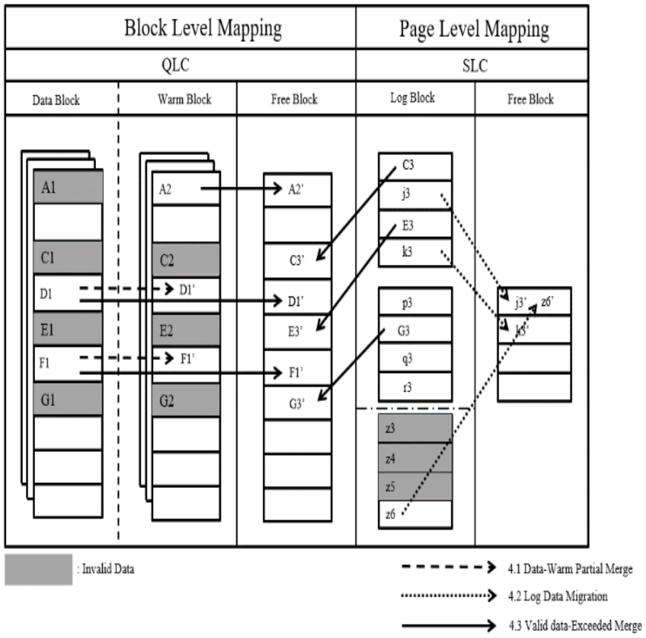


Fig. 3. Three Merge Operations of the Proposed Technique.

B. Log Data Migration

If the SLC log blocks are filled with data, a block with a large amount of invalid data is selected as a victim block, and a valid data migration, which is moving data from a victim block to a free block, is conducted. As shown in Figure 3, the block in which the z3, z4, z5, and z6 data are stored within the log block is selected as a victim block because it has the largest amount of invalid data. Because the block whose valid data in the victim block is z6 data has the greatest amount of invalid data, it is selected as the victim block, and the Log Data Migration is performed from the victim block to a free block as moving valid z6 data. After conducting the data migration, the victim block is erased to make the next available free space. This results in only one erase operation.

In contrast, after assuming that there are no blocks in which z3, z4, z5, and z6 data are stored, all of the log blocks are stored as valid data. In this case, if the data in the worm block is to be written to the log block at a request to be updated, a Valid Data-Exceeded Merge operation occurs. The Valid Data-Exceeded Merge operation is discussed in Section 4.3. The j3 and k3 data remaining after the Valid Data-Exceeded Merge are positioned as doing the Log Data Migration into a free block.

C. Valid Data-Exceeded Merge

This is a full merge of data blocks, warm blocks, and log blocks, which is performed when log blocks are filled with valid data. A block with the lowest degree of association among the log blocks is selected as a victim block, and the Valid Data-Exceeded Merge is performed on the data block and warm block having the largest amount of data included in one of the LBNs.

However, if data are included in the LBN other than the victim block, as shown in G3 of Figure 3, the data are only migrated to the free block of the QLC, but the log block does not execute the Log Data Migration. Only one log block selected as a victim block conducts the Log Data Migration, and the victim block is erased to make the next free space available. This is because what is needed in that situation is only one block, and is also to avoid the problem of further updates if there is data in a location other than the corresponding LBN-consistent victim block.

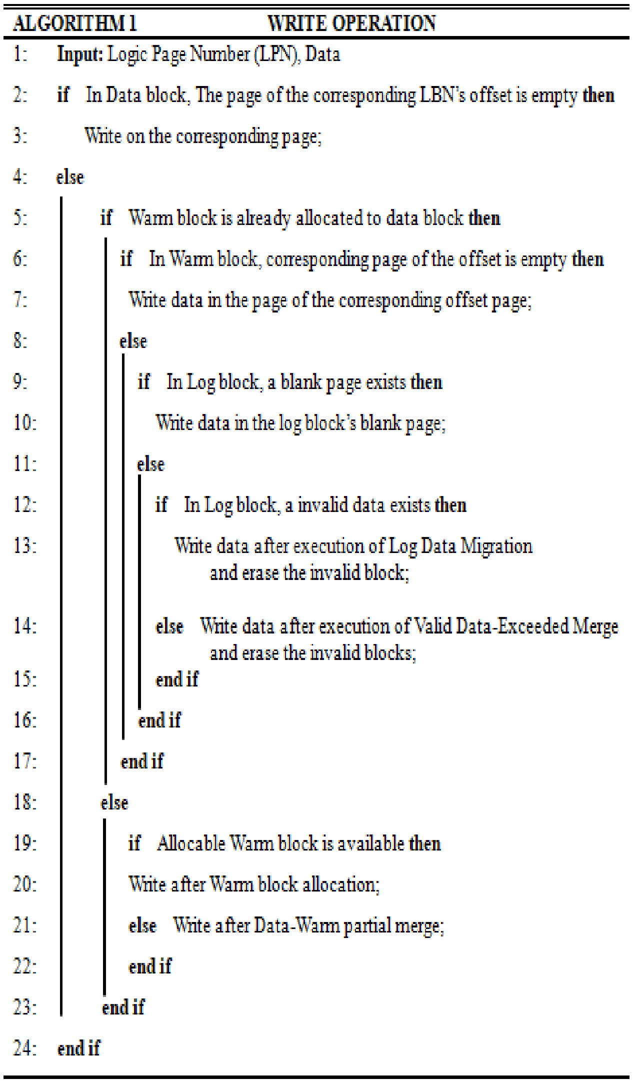
In this case, an erase operation is applied three times, namely, to a data block, a warm block, and a log block. Valid Data-Exceeded Merge occurs when the entire log block is filled with valid data, which means that the space utilization of the log area has already been maximized.

D. Write operation

Algorithm 1 describes the write operation algorithm of the proposed scheme. The empty space in the offset of the data block corresponding to the input logical page number is checked, and if empty, the data are written in the page of the corresponding offset (lines 1 and 3). If the page of the offset is not empty, the presence or absence of the assigned warm block should be identified.

If there is a warm block already allocated to the data block, the page space of the corresponding offset of the warm block is checked; if empty, the data are written into the corresponding warm block (line 7), but if not, the logical page number is considered hot and written on the log block.

Here, if there is a blank page space in the log block, the data can be written on the page (line 10). If there is no empty page space in the log block, a free space needs to be secured in the log area. To do so, the presence of invalid data in the entire log block is first identified; if invalid data exist, after Log Data Migration and one erase operation, log data are written into the empty space secured (line 13). However, if all log block data are full of valid data, because it is impossible to secure a free space in the log area, then garbage collection in the related data block, warm block, and log block is conducted by Valid Data-Exceeded Merge to occupy free space to write new data (line 14). If there is no warm block assigned to the data block, the warm block that can be allocated in the warm area is inspected. If an allocable warm block exists, data are written after allocation (line 20). If no warm blocks are available, Data-Warm partial merge is conducted to execute the garbage collection and secure the assignable warm block. The generated warm block is assigned to the corresponding data block and the data are written (line 21).



```
ALGORITHM 1                    WRITE OPERATION
1:  Input: Logic Page Number (LPN), Data
2:  if  In Data block, The page of the corresponding LBN's offset is empty then
3:      Write on the corresponding page;
4:  else
5:          if  Warm block is already allocated to data block then
6:              if  In Warm block, corresponding page of the offset is empty then
7:              Write data in the page of the corresponding offset page;
8:              else
9:                  if  In Log block, a blank page exists then
10:                     Write data in the log block's blank page;
11:                 else
12:                     if  In Log block, a invalid data exists then
13:                         Write data after execution of Log Data Migration
                                and erase the invalid block;
14:                     else  Write data after execution of Valid Data-Exceeded Merge
                                and erase the invalid blocks;
15:                     end if
16:                 end if
17:             end if
18:         else
19:             if  Allocable Warm block is available then
20:             Write after Warm block allocation;
21:             else  Write after Data-Warm partial merge;
22:             end if
23:         end if
24: end if
```

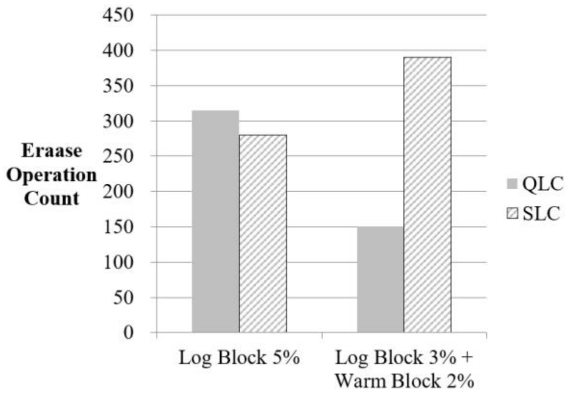


Fig. 4. Comparison of Erase Operations before and after Applying Warm Block Technique.

V. Conclusion and Future Plans

In this study, we proposed an algorithm for minimizing the erase operations of a QLC semiconductor chip and increasing the input/output performance of an SLC semiconductor chip by utilizing the warm block technique in the mass storage system, which is a mixture of QLC and SLC.

To measure the performance of QLC + SLC mixed NAND flash memory, QLC and SLC hardware characteristics were implemented based on existing data sheets in [10] and [11], and a performance evaluation was conducted after the warm block technique was implemented in the device drive of mixed NAND flash memory. In this study, the data sent from the file system were modified to generate about 10,000 write requests to improve an intuitive understanding.

Figure 4 shows the number of operations conducted in QLC + SLC mixed NAND flash memory when a conventional log block algorithm and the warm block technique are applied. Although the total numbers of erase operations in both the existing algorithm and the warm block algorithm are similar, the number of erase operations conducted in the QLC and SLC chips is quite different. A more intensive erase operation is conducted on a QLC chip for the existing log block algorithm (QLC 315, SLC 281), whereas an SLC chip is more intensely used for the warm block technique (QLC 153, SLC 392).

These experimental results support the idea that the warm block technique can improve the overall lifetime of QLC + SLC mixed NAND flash memory.

Considering the characteristics of the semiconductor chip mentioned in Section 1, the lifetime of QLC flash memory is 1,000 operations or less, whereas the lifetime of SLC flash memory is 100,000 operations or more [10], [11]. Therefore, in this study, an overload of read/write/erase operations in QLC flash memory can be prevented by efficiently allocating hot data with frequent updates to a long-lived SLC. In other words, the lifetime extension and performance of QLC + SLC hybrid NAND flash memory are improved using the warm block technique, as proposed in this study.

Future studies will apply the warm block technique to various FTL algorithms in addition to the log block algorithm, and we plan to propose a wear-leveling algorithm that optimizes the compatibility with the warm block technique. In addition, based on actual workloads used in smartphones and servers, we plan to compare the performance with existing techniques, including the number of erase operations and the operation speed of the write and read requests in various environments.

Acknowledgment

This work was supported by Basic Science Research through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B04031440). This study was also supported by 2018 Research Grant from Kangwon National University (No. 000000000).

References

[1] S. Hong and D. Shin, "NAND Flash-Based Disk Cache Using SLC/MLC Combined Flash Memory," 2010 International Workshop on Storage Network Architecture and Parallel I/Os, Incline Village, NV, pp. 21-30, 2010.

[2] Dongzhe Ma, Jianhua Feng, and Guoliang Li, “A survey of address translation technologies for flash memories,” ACM Computing Surveys (CSUR), 46(3), pp. 1-39, 2014.

[3] Jesung Kim, Jong Min Kim, S. H. Noh, Sang Lyul Min and Yookun Cho, "A space-efficient flash translation layer for CompactFlash systems," in IEEE Transactions on Consumer Electronics, vol. 48, no. 2, pp. 366-375, May 2002.

[4] Rino Micheloni, “Solid-state drive (SSD): A nonvolatile storage system,” Proceedings of IEEE 105(4), pp. 583-588, 2017.

[5] Se Jin Kwon and Tae-Sun Chung, “Data pattern aware FTL for SLC+MLC hybrid SSD,” Design Automation for Embedded Systems 19(1-2), pp. 101-127, 2015.

[6] Sang-Won Lee, Dong-Joo Park, Tae-Sun Chung, Dong-Ho Lee, Sangwon Park, and Ha-Joo Song, “A log buffer-based flash translation layer using fully-associative sector translation,” ACM Transactions on Embedded Computing Systems (TECS) 6(3), no. 18, pp. 1-27, 2007.

[7] Byung-Woo Nam, Gap-Joo Na, and Sang-Won Lee, “A hybrid flash memory SSD scheme for enterprise database applications,” Web Conference (APWEB), 2010 12th International Asia-Pacific, IEEE, 2010.

[8] Lee, Sang-Won, et al. "A log buffer-based flash translation layer using fully-associative sector translation." ACM Transactions on Embedded Computing Systems (TECS) 6.3 (2007): 18.

[9] Gupta, Aayush, Youngjae Kim, and Bhuvan Urgaonkar. DFTL: a flash translation layer employing demand-based selective caching of page-level address mappings. Vol. 44. No. 3. ACM, 2009.

[10] Toshiba Semiconductor, “BiCS3 768 Gb 3D QLC NAND chips,” Toshiba Semiconductor Technical Notes, 2018.

[11] Micron Electronics, “MT29E128G08CECAB, MT29E256G08CMCAB, MT29E512G08CUCAB,” Micron Electronics Datasheet, 2018.

Energy-Efficient Security Threshold Determination Method for the Enhancement of Interleaved Hop-By-Hop Authentication

Ye Lim Kang[1], Tae Ho Cho[*2]
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea

***Abstract*—Wireless sensor networks allow attackers to inject false reports by compromising sensor nodes due to the use of wireless communication, the limited energy resources of the sensor nodes, and deployment in an open environment. The forwarding of false reports causes false alarms at the Base Station and consumes the energy of the sensor nodes unnecessarily. As a defense against false report injection attacks, interleaved hop-by-hop authentication was proposed. In interleaved hop-by-hop authentication, the security threshold is a design parameter that influences the number of Message Authentication Codes; the sensor nodes must verify, based on the security requirements of the application and the node density of the network. However, interleaved hop-by-hop authentication fails to defend against false report injection attacks when the number of compromised sensor nodes exceeds the security threshold. To solve this problem, in this paper we propose a security scheme that adjusts the security threshold according to the network situation using an evaluation function. The proposed scheme minimizes the energy consumption of the sensor nodes and reinforces security.**

Keywords—Component; wireless sensor networks; false report injection attack; network security; interleaved hop-by-hop authentication

I. Introduction

Wireless Sensor Networks (WSNs) are densely deployed with many sensor nodes and use wireless communication [1]. WSNs are used in various applications that require real-time observation, such as fire detection and enemy movement detection [2]. For this reason, it is important to transmit accurate information to the Base Station (BS).

Security is an essential consideration in WSNs. WSNs are extremely vulnerable to false report injection attacks, due to their use of wireless communication, their deployment in open environments, and the limited energy resources of the sensor nodes [3-4]. In a false report injection attack, an attacker injects a false report into a WSN by compromising certain sensor nodes. The BS causes false alarms upon receiving false reports. In addition, the forwarding of false reports unnecessarily consumes the energy of the sensor nodes. Therefore, false reports must be detected early and dropped before they arrive at the BS.

As a defense against false report injection attacks, Z. Sencun, S. Sanjeev, J. Sushil., N. Peng proposed Interleaved Hop-by-hop Authentication (IHA) [5-6]. IHA is a security protocol in which sensor nodes detect and drop false reports during transmission if the number of compromised nodes does not exceed a certain Security Threshold (T). IHA is used when strong security is desired because it involves two-step report verification with one-hop neighbors and T+1-hop neighbors. However, IHA does not defend against false report injection attacks when the number of compromised nodes exceeds T.

In this paper, we propose a security scheme that determines a suitable T according to the network situation by means of an evaluation function. The BS resets T according to the network situation and alters the number of Message Authentication Codes (MACs) to be included in the report. By resetting T, it is possible to defend against false report injection attacks when the number of compromised nodes exceeds T. As a result, the proposed scheme improves the total energy efficiency of the network and reinforces security.

The composition of this paper is as follows. In Section 2, false report injection attacks and IHA are described. In Section 3, the proposed scheme is described. Section 4 illustrates the performance of the proposed scheme through experimental results. In Section 5, conclusions are drawn.

II. Related Work

A. False Report Injection Attacks

Fig. 1 depicts a false report injection attack in a WSN. An attacker compromises sensor nodes and obtains an authentication key. Then, the attacker generates a false report about an event that did not occur using the acquired authentication key. This causes a false alarm at the BS, and the sensor nodes consume energy unnecessarily during the transmission of false reports. There are various security protocols to defend against such attacks, including Dynamic En-route Filtering (DEF) and Probabilistic Voting-based Filtering (PVFS) [7-8].

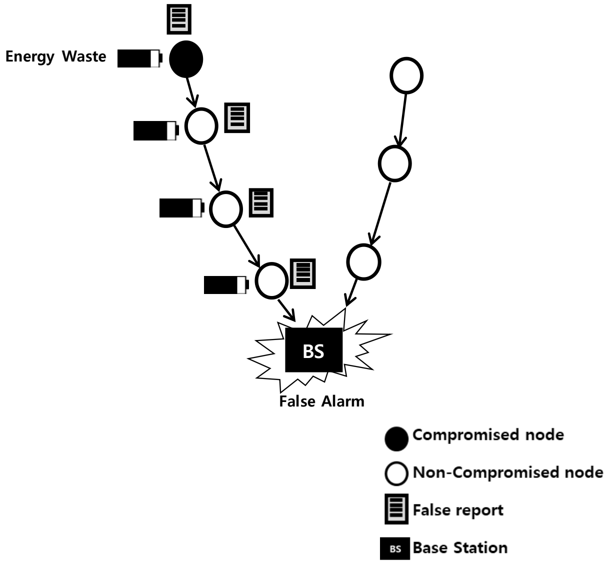


Fig. 1. False Report Injection Attack

B. IHA

1) Node Initialization and Deployment

The key server is located at the BS and distributes and manages keys. The key server preloads unique integer IDs and individual keys onto all nodes. After node deployment, all the nodes discover one-hop neighbor nodes and establish pairwise keys with them [9-10].

2) Association Discovery

Fig. 2 displays the Association Discovery step, in which all nodes discover the IDs of association nodes. For the initial path setup, there are association discovery steps, such as BS Hello and Cluster Acknowledgement [11]. The BS Hello step allows the node to discover an upper association node. The BS broadcasts a Hello message. The node receiving the Hello message discovers the ID of a T+1-hop upper association node within the path. The Cluster Head (CH) receiving the Hello message allocates T+1 IDs to the cluster nodes. The Cluster Acknowledgement (ACK) step allows the node to discover a lower association node. After the BS Hello step, the CH transmits ACK toward the BS. The node receiving ACK discovers the ID of a T+1-hop lower association node.

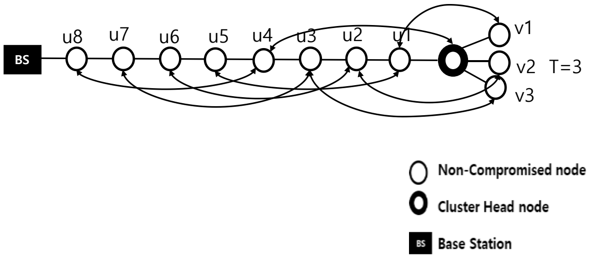


Fig. 2. Association Discovery

3) Report Endorsement

If an event occurs, the cluster nodes generate an endorsement message including the MAC, and transmit it toward the CH. The CH authenticates the endorsement message using pairwise keys shared with the cluster nodes. If the endorsement message is authenticated, the CH generates an event report including the MAC, and transmits it toward the BS.

4) En-route Filtering

Fig. 3 depicts en-route filtering in IHA. The node receiving a report from the CH verifies the report using a pairwise key shared with its downstream node. Then, the node checks the number of pairwise MACs within the report. The node verifies the last MAC within the pairwise MAC list of the report using the pairwise key shared with its lower association node. If the verification succeeds, the last MAC is eliminated from the pairwise MAC list. The node generates a MAC using the pairwise key shared with its upper association node and adds it to the beginning of the pairwise MAC list of the report. Then, the report is transmitted to the next node. All forwarding nodes repeat the same process.

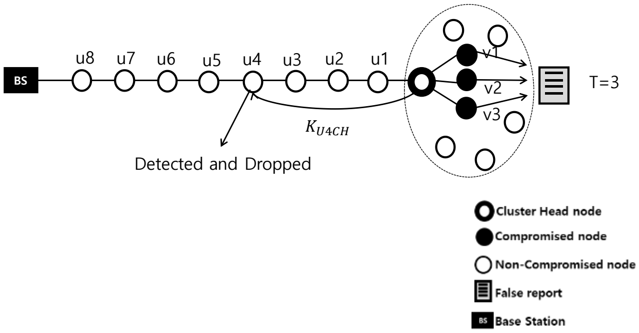


Fig. 3. En-route Filtering

5) BS Verification

If a report arrives at the BS, the BS verifies the compressed individual MACs [12]. The BS computes T+1 MACs about the event using the authentication keys of the nodes in the ID list of the report. Then, the BS determines whether or not these MACs match the individual MACs within the report. If the verification succeeds, the report is authenticated. However, if the verification fails, the report is discarded.

III. Proposed Scheme

A. Problem Statement

IHA is an effective security protocol that defends against false report injection attacks. However, it has several problems.

1) IHA loses en-route filtering during false report injection attacks when the number of compromised nodes exceeds T. The event data of a false report that the number of compromised nodes exceeds T are false, but the report is composed of normal MACs. There are not methods to detect and drop such reports before the BS receives them.

2) In IHA, if the node receives a report, the node goes through a verification process with its T+1-hop neighbor node in addition to its one-hop neighbor node. IHA has powerful security, but the energy of the sensor nodes is consumed much more than with other security protocols.

To solve these problems, we propose the following solutions

3) The network administrator presets standard range of event data values of a normal report. The BS judges that a received report is false if it deviates from the standard range of a normal report set by the network administrator.

4) The BS minimizes the energy consumption of the sensor nodes as much as possible by determining the suitable T according to the network situation.

B. Assumptions

- We assume the energy of the sensor nodes is not limited.
- We assume the BS is not compromised.

C. Proposed Scheme

1) Factors Considered to Determine T

Fig. 4 demonstrates that after the BS resets T based on the False Traffic Ratio (FTR), Residual Energy of the Node (REN), and Max Hop Count (MHC), it broadcasts T. The evaluation function is executed whenever the number of received reports is 100. The BS performs the evaluation function considering the cluster in which the FTR is the largest and the REN is the smallest among the clusters in which the false report injection attack has occurred. The evaluation function is not executed if T is changed from the initially set T to a new T. The node detecting the false report deletes the false MAC without discarding the false report, and forwards the false report to the BS by adding the pairwise MAC of its upper association node. The BS can determine the number of compromised nodes and their IDs using this received false report. The BS determines the value to which it should reset T according to the number of compromised nodes. T should be reset to a value equal to the number of compromised nodes or greater than the number of compromised nodes. If all the nodes receive a broadcasting message to reset T from the BS, all the nodes will stop transmitting false reports after detecting them. Once T is reset, all the nodes operate by detecting and dropping false reports, as in existing IHA.

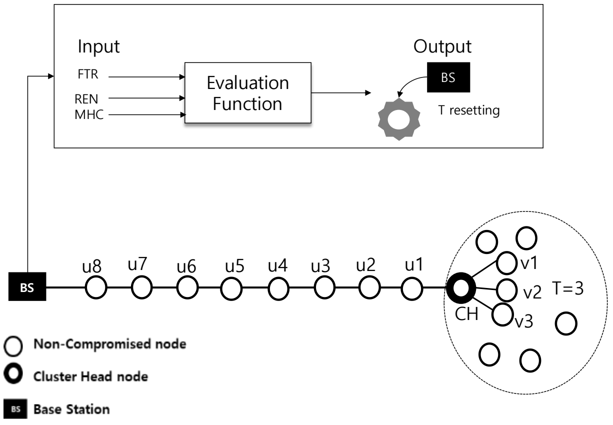


Fig. 4. Interleaved Hop-by-hop Authentication Using the Evaluation Function

Evaluation function (1) determines whether to reset T based on the FTR, REN, and MHC of Path (P).

$$\mathrm{T(p)} = \frac{FTR(p)}{FTR(p) + MHC(p)} + REN(p) \qquad (1)$$

Table 1 displays the output according to the inputs of the evaluation function. If the FTR is 0-10% and the REN is 0-100%, the BS resets T to a value smaller than the initially set T through the evaluation function. Then, the BS broadcasts T. At this time, the execution condition of the evaluation function is that the number of compromised nodes is smaller than the initially set T. This minimizes unnecessary energy consumption by the sensor nodes.

If the FTR is 11-100% and the REN is 0-100%, the BS resets T to a value greater than the initially set T through the evaluation function. Then, the BS broadcasts T. At this time, the execution condition of the evaluation function should be the situation of a false report injection attack when the number of compromised nodes exceeds the initially set T. In other words, the BS restores en-route filtering by resetting T to a value greater than the initially set T. Therefore, false reports are dropped before they arrive at the BS. Thus, it is possible to defend against attacks that cannot be defended against by the existing IHA.

Determining the suitable T involves adjusting T differently in various network situations, such as attack situations.

TABLE I. EVALUATION FUNCTION

Input			Output	
FTR(%)	**REN(%)**		**T**	
0~10	0~10	**MHC**	0~10	Down
	11~21		11~21	
	22~100		22~100	
11~100	0~10		0~10	Up
	11~21		11~21	
	22~100		22~100	

2) Reassociation Setup of Nodes

The BS broadcasts the new T to all the nodes when T is reset as a result of the evaluation function. Fig. 5 depicts the process in which all the nodes reset T+1 association node pairwise keys with the new T. For example, if T is changed from 3 to 4, node u3 resets its upper association node pairwise key from K_{u3u7} to K_{u3u8}.

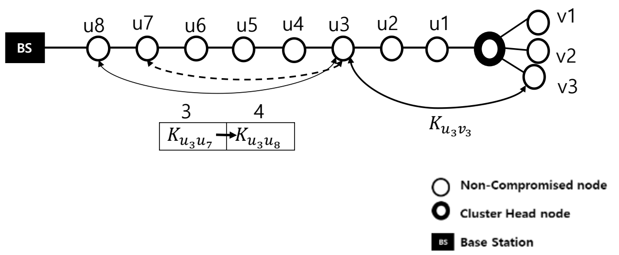


Fig. 5. Reassociation Setup

IV. PERFORMANCE EVALUATION

A. Experimental Environment

The experimental environment is as follows: 2000 nodes are randomly deployed in the sensor field, which is 1000 x $1000m^2$ in size, and 2000 events are generated. The BS is positioned at (x, y = 1000, 1000) of the sensor field. The energy required to transmit 1 byte is 16.25 µJ, and the energy required to receive 1 byte is 12.5 µJ [13]. The energy required to verify a MAC is 75 µJ, and the energy required to generate a

MAC is 15 µJ [14]. The size of a MAC is 1 byte and the size of the original report is 12 bytes. The size of the Hello Message is 30 bytes [15]. The initial energy resource of the nodes is 1 J.

B. Experimental Results

Fig. 6 displays the total energy consumption of the network according to the FTR in the situation of a false report injection attack when the number of compromised nodes is less than T. To demonstrate the energy efficiency of the proposed scheme, we generated 2000 events at random positions and analyzed the total energy consumption of the sensor nodes. As the FTR increases, the total energy consumption decreases. IHA maintains the initially set T. In the proposed scheme, the BS changes its broadcast value from the initially set T to the reduced T if the FTR is 0-10% and the REN is 0-100%.

Comparing IHA and the proposed scheme, when the FTR is 100%, the energy efficiency is improved by up to 27.17%, as shown in Fig. 6.

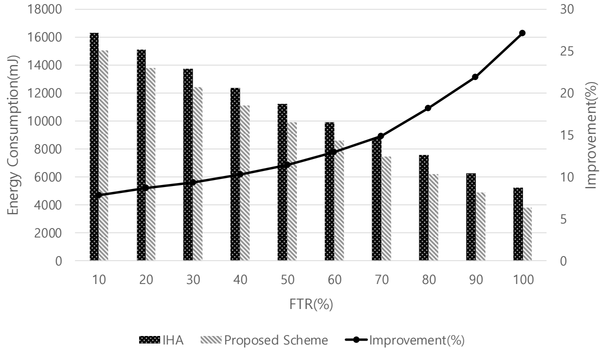


Fig. 6. Energy Consumption versus the FTR when the number of compromised nodes is less than T.

Fig. 7 displays the total energy consumption of the network according to the FTR in the situation of a false report injection attack when the number of compromised nodes exceeds T. IHA maintains the initially set T. In the proposed scheme, when the FTR is 11-100% and the REN is 0-100%, the BS changes its broadcast value from the initially set T to the increased T. Comparing IHA and the proposed scheme, when the FTR is 90%, the energy efficiency increases by up to 24.18%, as shown in Fig. 7.

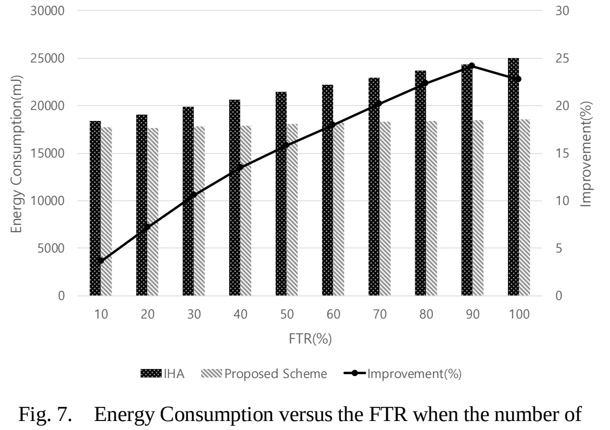


Fig. 7. Energy Consumption versus the FTR when the number of compromised nodes exceeds T.

Fig. 8 displays the number of dropped false reports according to the FTR in the situation of a false report injection attack when the number of compromised nodes exceeds T. To demonstrate the security of the proposed scheme, we generated 2000 events at random positions and analyzed the number of dropped false reports. IHA maintains 0% en-route filtering, while the proposed scheme progressively improves en-route filtering as the FTR increases. Thus, the security of the proposed scheme is better than that of IHA.

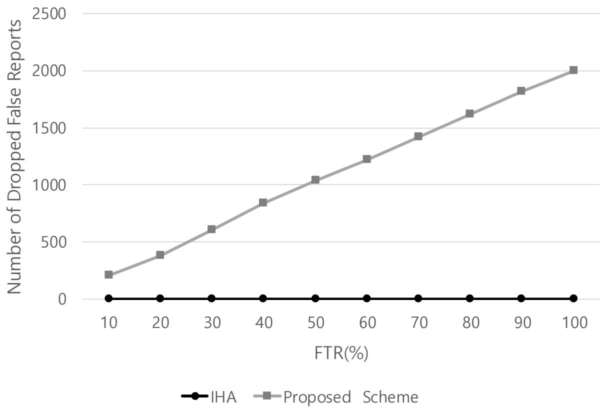


Fig. 8. Number of Dropped False Reports versus the FTR

V. Conclusions

WSNs are vulnerable to false report injection attacks. IHA is an effective security protocol to defend against false reports. However, with IHA, it is impossible to defend against false report injection attacks when the number of compromised nodes exceeds T. In this paper, we proposed a WSN security scheme that adjusts T according to the network environment. When sensor nodes detect a false report that the number of compromised nodes is less than T, the BS reduces T and prevents unnecessary energy use by the sensor nodes through the evaluation function. The BS broadcasts an increased value of T when the sensor nodes detect a false report that the number of compromised nodes exceeds T. Therefore, en-route filtering is restored, security is enhanced compared to the existing IHA, and unnecessary energy wasting by the sensor nodes is prevented because false reports are discarded in advance. However, in the node initialization deployment step, all the nodes have many keys, so much of the memory of the sensor nodes is consumed. Therefore, in the proposed scheme, costs increase because expensive sensor nodes must be used. Through this experiment, we have demonstrated the performance improvement of the proposed scheme compared with IHA.

Acknowledgment

This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) (2015-0-00914) supervised by the IITP (Institute for Information & communications Technology Promotion)" (2015-0-00914)

References

[1] A. Ian F., S. Weilian, S. Yogesh, C. Erdal, "A survey on sensor networks." IEEE communications magazine 40.8 (2002): 102-114.

[2] A. K .Jamal N., and K. AHMED E. "Routing techniques in wireless sensor networks: a survey." IEEE wireless communications 11.6 (2004): 6-28.

[3] Jeba, S. A., and B. Paramasivan. "False data injection attack and its countermeasures in wireless sensor networks." European Journal of Scientific Research 82.2 (2012): 248-257.

[4] K. Chris, and W. David. "Secure routing in wireless sensor networks: Attacks and countermeasures." Sensor Network Protocols and Applications, 2003. Proceedings of the First IEEE. 2003 IEEE International Workshop on. IEEE, 2003.

[5] Z. Sencun, S. Sanjeev, J. Sushil., N. Peng, "An interleaved hop-by-hop authentication scheme for filtering of injected false data in sensor networks." Security and privacy, 2004. Proceedings. 2004 IEEE symposium on. IEEE, 2004.

[6] Z. Sencun, and N. Peng. "Interleaved hop-by-hop authentication against false data injection attacks in sensor networks." ACM Transactions on Sensor Networks (TOSN) 3.3 (2007): 14.

[7] Y. Zhen, and G. Yong. "A dynamic en-route filtering scheme for data reporting in wireless sensor networks." IEEE/ACM Transactions on Networking (ToN) 18.1 (2010): 150-163.

[8] L. Feng, S. Avinash and W. Jie. "PVFS: a probabilistic voting-based filtering scheme in wireless sensor networks." International Journal of Security and Networks 3.3 (2008): 173-182.

[9] Z. Sencun, S. Sanjeev, and J. Sushil. "LEAP+: Efficient security mechanisms for large-scale distributed sensor networks." ACM Transactions on Sensor Networks (TOSN) 2.4 (2006): 500-528.

[10] B. Carlo, S. Alfredo De, H. Amir, K. Shay, V. Ugo, Y. Moti, "Perfectly-secure key distribution for dynamic conferences." Annual international cryptology conference. Springer, Berlin, Heidelberg, 1992.

[11] W. Yong, A. Garhan, and R. Byrav. "A survey of security issues in wireless sensor networks." (2006).

[12] B. Mihir, G. Roch, and R. Phillip. "XOR MACs: New methods for message authentication using finite pseudorandom functions." Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 1995.

[13] P. Dongjin, and C. Taeho. "A Fuzzy Rule-based Key Re-Distribution Decision Scheme of Dynamic Filtering for Energy Saving in Wireless Sensor Networks." International Journal of Information Technology and Computer Science (IJITCS) 9.4 (2017): 1-8.

[14] Y. Fan, L. Haiyun, L. Songwu, Z. Lixia, "Statistical en-route filtering of injected false data in sensor networks." IEEE Journal on Selected Areas in Communications 23.4 (2005): 839-850.

[15] R. Sajjad, Q. Hassaan Khaliq, K. Syed Ali, R. Veselin, R. Muttukrishnan, "A1: An energy efficient topology control algorithm for connected area coverage in wireless sensor networks." Journal of Network and Computer Applications 35.2 (2012): 597-605.l

WordNet based Implicit Aspect Sentiment Analysis for Crime Identification from Twitter

Hajar El Hannach[1], Mohammed Benkhalifa[2]
ANISSE research team, Computer science department
Faculty of science, Mohammed V University
Rabat, Morocco

***Abstract*—Crime analysis has become an interesting field that deals with serious public safety issues recognized around the world. Today, investigating Twitter Sentiment Analysis (SA) is a continuing concern within this field. Aspect based SA, the process by which information can be extracted, analyzed and classified, is applied to tweet datasets for sentiment polarity classification to predict crimes. This paper addresses the aspect identification task involving implicit aspect implied by adjectives and verbs for crime tweets. The proposed hybrid model is based on WordNet semantic relations and Term-Weighting scheme, to enhance training data for (1) Crime Implicit Aspect sentences detection (IASD) and (2) Crime Implicit Aspect Identification (IAI). The performance is evaluated using three classifiers Multinomial Naïve Bayes, Support Vector Machine and Random Forest on three Twitter crime datasets. The obtained results demonstrate the effectiveness of WN synonym and definition relations and prove the importance of verbs in training data enhancement for crime IASD and IAI.**

Keywords—Implicit aspect based sentiment analysis; information retrieval; machine learning; supervised approaches; frequency model; WordNet; crime detection; hate crime twitter sentiment (HCTS)

I. Introduction

Sentiment Analysis (SA) has become one of the most active topics in information retrieval and text mining due to the large expansion of the World Wide Web. SA is the field of study that deals with automatic analysis of people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text [1]. The entities can be products, services, organizations, individuals, events, or topics such as crimes. SA research has been mainly carried out at three levels of granularity: document, sentence or aspect level. Aspect level SA is the most fine-grained model, which extracts opinions expressed against different aspects/features of the entity.

Classifying opinion text at the document level or at the sentence level as positive or negative is insufficient for most applications. These classifications do not tell what each opinion is about, that is, the target of opinion. Indeed, when a document or a sentence evaluates a single entity, it does not mean that this evaluation is true for all aspects of the entity [2]. For a more complete analysis, aspects need to be discovered before to determine whether the sentiment is positive, negative, or neutral about each aspect. To obtain this level of fine-grained results, Aspect-based Sentiment Analysis (ABSA) is applied [3]. This latter considers relations between the aspects of the object of the opinion and the document polarity (positive or negative feeling expressed in the opinion). An aspect is a concept on which the author expresses his/her opinion in the document. The aspects can be of two types: explicit aspects and implicit aspects. Explicit aspects correspond to specific terms that explicitly appear in the document. In contrast, an implicit aspect is not specified explicitly in the document. The implicit aspects (which can be indicated by adjectives, adverbs, verbs or phrasal verbs) are very important that they can convey the opinions and help in improving the performance of SA systems.

Within the next few years, SA and more particularly IASA is set to become a promising approach for crime prediction [4]–[6]. Nowadays, IASA is applied for crime prevention systems such as neighborhood crime rating systems and safety of school platforms that are developed to support crime prevention and fear reducing. The Most challenging task in crime prediction area is identifying the set of committed crimes according to their types, locations and individuals, especially when this information is implicitly implied and not mentioned explicitly in data. In this scenario, Implicit Aspect based Sentiment Analysis (IASA) can be used to highlight the patterns of crimes.

When applied to crime prediction, IASA operates in three steps: (1) implicit aspect sentences detection (IASD), (2) implicit aspect identification (IAI) and (3) sentiment classification.

For crime datasets, Twitter is a defensible and logical source of data widely used in crime prevention and pattern detection approaches[7]–[9]. When gathering implicit aspect sentences from this popular social networking site, the main issue is the huge number of tweets returned with poor grammar and spelling, hashtags, URL, and irrelevant sentences. Thus, the construction of implicit aspect crime datasets requires preprocessing treatment and information retrieval techniques in order to classify relevant and irrelevant sentences. This process is known as "implicit aspect tweets or sentences detection".

After building crime datasets, Implicit Aspect Identification (IAI) is performed. IAI encompasses implicit aspect term (IAT) extraction and IAT aggregation. For each implicit aspect sentence, IAT extraction aims at extracting adjectives, verbs implying aspects. Afterward, extracted terms suggesting the same aspect are assembled into one implicit aspect in IAT aggregation.

After the implicit aspect identification, sentiment classification can be applied to classify opinions, toward each aspect, into positive or negative classes.

In this paper, the focus is made on Implicit aspect sentence detection and Implicit Aspect identification. A hybrid model, coupling WordNet Synonym and Definition semantic relations and Term Weighting scheme, is proposed for training data improvement to support both IASD and IAI steps. The proposed hybrid model is empirically evaluated using three classifiers Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Random Forest(RF) on three Twitter crime datasets. The study shows that our approach helps the three classifiers achieve good performance for IASD and IAI tasks.

The remainder of the paper is organized as follows. In section 2, related works are reviewed. In section 3, the proposed hybrid model based approach is presented in details. In section 4, the experimental setting adopted is exposed. In section 5, obtained results are presented and discussed. Finally, conclusions and future work are presented.

II. Related Works

A considerable amount of works have been published in aspect based sentiment analysis [10], [11], while few have attempted to address the implicit aspect identification. The methods applied for this task are based on two major methods: lexical based and supervised learning approaches. Among the lexical based approaches, the semantic orientation methods are used to supports binary classification [12]. Dictionary based techniques are one of the most popular lexical approach used in this field. In [13], authors try a new approach based lexical method, Part of speech tagging, SentiWordNet and WordNet combined with a weighted model provided by Natural language processing NLP(weight assignment policies) in sentiment classification. Their results outperform the basic use of WEKA Naïve Bayes Classifier and prove the effectiveness and contribution of the lexical approach in opinion mining.

Several studies investigating machine learning have been carried out on sentiment analysis. Machine learning algorithms have been used to solve the sentiment analysis as a regular text classification problem. In [14] performed a comparative study involving different machine learning algorithms. Naïve Bayes, Support Vector Machine and maximum-entropy-based classifiers are applied for sentiment polarity classification for movies reviews. Compared to the human generated baselines, the ML techniques achieve the better performances. Data representation is also among factors that impact ML performances. In [15], authors aim at investigating the effectiveness of vector representation for explicit aspect extraction. Their approach is hybrid based on Semantic Role Labelling, Conditional Random Fields and Structural Support Vector Machines (SVM-HMM). The evidence presented in their work suggests that the vector space approach support explicit aspect extraction and SA classification.

Much of the current studies on SA pays particular attention to Twitter trends and opinions. A lot of research has been done in this field by researchers and scholars all around the world [16]–[18]. Sentiment analysis in tweets is done according to major steps, identifying opinion target, explicit or implicit aspect, and classifying the sentiment polarity of tweets. To perform Sentiment classification in twitter, most of the research applied the followed process: data collection, information retrieval and sentiment classification [19], [20]. For information retrieval, Term Frequency-Inverse Document Frequency (TF-IDF) is among the most popular technique used for text categorization and tweets selection [21]. This term weighting scheme is easy to compute, implement, and understand. However, its shortcoming is very well recognized. For imbalanced datasets, the TF-IDF need to be enhanced to allows better performances[22].

Sentiment analysis is fast becoming a key instrument in Crime prevention and data detection. In [4], authors elaborate a sentiment analysis approach based on lexicon methods and combined with kernel density estimation based on historical crime incidents to predict the time and location in which a specific crime will occur. Their approach provides a significant achievement comparing to the benchmark model. Others in [8], addressed the aspect-based sentiment analysis for crime tweets through the use of hybrid model. Based on Natural Language Processing techniques and SentiWordNet, the hybrid model detects the subjectivity of crime and then predicts the hate crime tweets polarity.

III. Proposed Framework

The proposed study is motivated by considering WordNet extracted terms according to Synonym and Definition subsets for adjectives and verbs coupled with a new Term Weighting model to represent implicit aspects and improve training data. This motivation is driven by two curiosities: (1) How these WN extracted terms can be exploited and combined with their corpus adjectives and verbs to best represent implicit aspects and (2) How this combination can be made optimally informative to both tasks: implicit aspect crime sentence detection and implicit aspect identification.

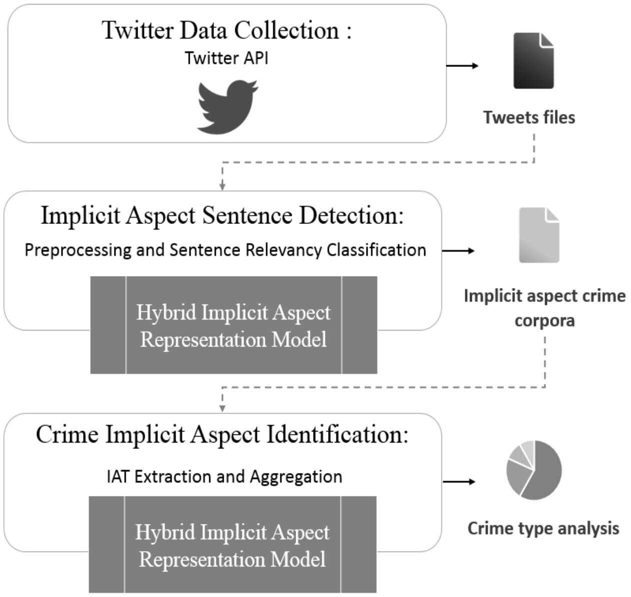

Twitter Data Collection :
Twitter API
Tweets files
Implicit Aspect Sentence Detection:
Preprocessing and Sentence Relevancy Classification
Hybrid Implicit Aspect Representation Model
Implicit aspect crime corpora
Crime Implicit Aspect Identification:
IAT Extraction and Aggregation
Hybrid Implicit Aspect Representation Model
Crime type analysis

Fig. 1. Abstract Process of the Proposed Framework.

The proposed approach is supported by a hybrid representation model. It operates in three phases according to the schema shown in Fig. 1. The first phase collects tweet datasets using the official Twitter Search API v1.1. The second phase proceeds into 2 steps: The preprocessing that prepares tweet datasets and the sentence relevancy classification that detects implicit aspect crime sentences. The third phase performs IAT extraction and IAT aggregation for crime implicit aspect identification.

Before presenting the exhaustive outline of the proposed approach, the Hybrid Implicit Aspect representation model (as shown in Fig. 2) is explained in details in the following section.

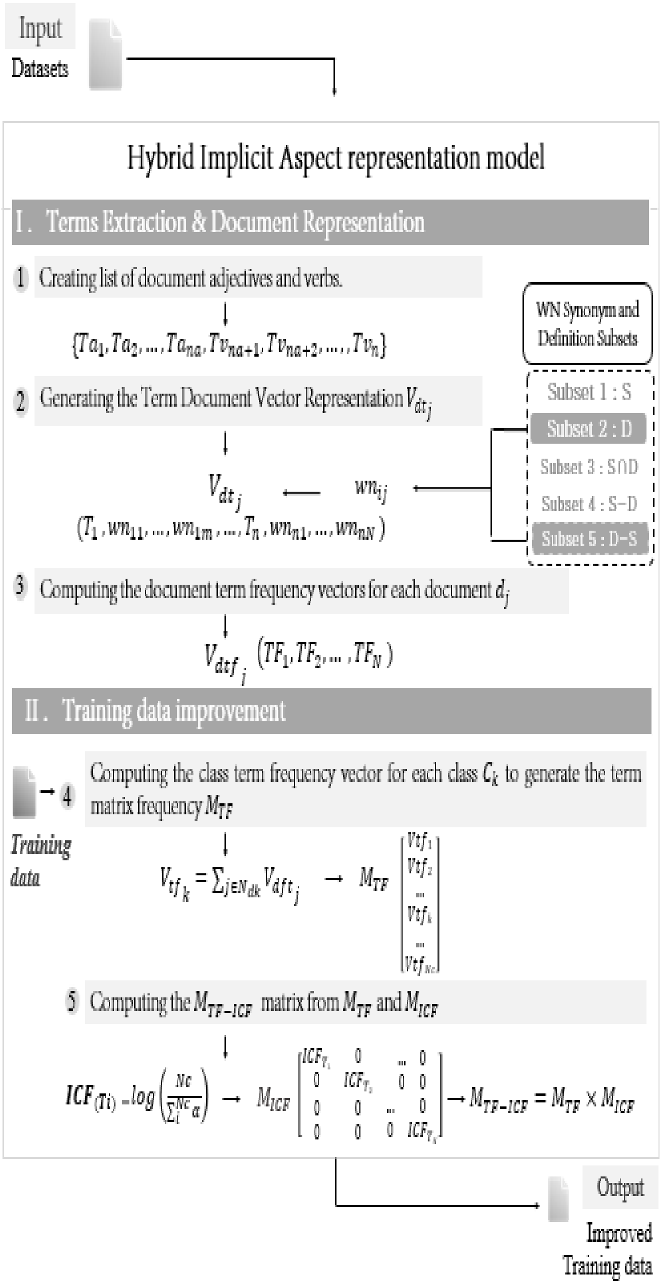

Input
Datasets
Hybrid Implicit Aspect representation model
I. Terms Extraction & Document Representation
1 Creating list of document adjectives and verbs.
$\{Ta_1, Ta_2, \ldots, Ta_{na}, Tv_{na+1}, Tv_{na+2}, \ldots, Tv_n\}$
WN Synonym and Definition Subsets
Subset 1 : S
Subset 2 : D
Subset 3 : S∩D
Subset 4 : S−D
Subset 5 : D−S
2 Generating the Term Document Vector Representation V_{dt_j}
$V_{dt_j} \leftarrow wn_{ij}$
$(T_1, wn_{11}, \ldots, wn_{1m}, \ldots, T_n, wn_{n1}, \ldots, wn_{nN})$
3 Computing the document term frequency vectors for each document d_j
$V_{dtf_j}\ (TF_1, TF_2, \ldots, TF_N)$
II. Training data improvement
4 Computing the class term frequency vector for each class C_k to generate the term matrix frequency M_{TF}
Training data
$V_{tf_k} = \sum_{j \in N_{dk}} V_{dft_j} \rightarrow M_{TF} \begin{bmatrix} Vtf_1 \\ Vtf_2 \\ \ldots \\ Vtf_k \\ \ldots \\ Vtf_{Nc} \end{bmatrix}$
5 Computing the M_{TF-ICF} matrix from M_{TF} and M_{ICF}
$ICF_{(Ti)} = log\left(\frac{Nc}{\sum_1^{Nc} \alpha}\right) \rightarrow M_{ICF} \begin{bmatrix} ICF_{T_1} & 0 & \ldots & 0 \\ 0 & ICF_{T_2} & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & ICF_{T_N} \end{bmatrix} \rightarrow M_{TF-ICF} = M_{TF} \times M_{ICF}$
Output
Improved Training data

Fig. 2. Summary of the Proposed Hybrid Implicit Aspect Representation Model.

A. Hybrid Implicit Aspect Representation Model

To represent crime implicit aspects, our hybrid model proceeds in five steps. Steps 1, 2, and 3 deal with extracting implicit aspect terms for document representation whereas step 4 and 5 bring improvements to training data.

Step 1 creates a list of extracted adjectives and verbs called terms T_i.

$\{Ta_1, Ta_2, \ldots, Ta_{na}, Tv_{na+1}, Tv_{na+2}, \ldots, Tv_n\}$

Where Ta_i and Tv_i denotes adjective and verb term respectively, and n represents the number of terms T_i.

To represent dataset documents, step 2 generates a document term vector V_{dt_j} from WN extracted terms vectors V_{T_i}. The V_{T_i} vectors are generated using the appropriate WN semantic relation subsets according to adjectives and verbs. Indeed, V_{T_i} (as shown in (1)) are constructed from the best supportive subsets of WordNet semantic relations, empirically identified in [23]. In this latter work, five WN subsets are considered for adjectives and verbs:

- Subset 1: S which contains all words extracted from synonym relation.
- Subset 2: D containing all synonyms words and nouns appearing in phrases describing a given word from definition relation.
- Subset 3: S ∩ D that contains words appearing in synonym and definition relations.
- Subset 4: S-D, composed of words appearing in synonym relation and not in definition relation.
- Subset 5: D-S, representing words appearing in definition relation and not in synonym relation.

For Ta_i , V_{T_i} vectors are constructed from Subset 2 (D) containing synonyms and nouns appearing in Ta_i description. For Tv_i, V_{T_i} vectors are generated from subset 5 (D-S) that is composed of nouns appearing in verb definition and not in synonym relation.

$$\mathrm{V_{T_i}} = (\mathrm{wn_{i1}}, \mathrm{wn_{i2}}, \ldots, \mathrm{wn_{im}}) \tag{1}$$

The document term vector V_{dt_j} representing a document j is generated as follows:

$$V_{dt_j} = (T_1, wn_{11}, \ldots, wn_{1m}, \ldots, T_n, wn_{n1}, \ldots, wn_{nN} \tag{2}$$

Where wn_{in} is the n-th WN related word extracted for Term T_i and N denotes the number of terms and theirs WN extracted terms.

After the term document vector generation, step 3 computes the document term vector frequency V_{dtf_j} for each document j. TF is calculated for T_i and their WN extracted terms wn_{im} . The wn_{im} term frequency is equal to the number of times term Ti occurs in document d_j.

$$V_{dtf_j} = (TF_1, TF_2, \ldots, TF_N) \tag{3}$$

Where TF_i is the document term frequency of term T_i.

Instead of using Term Frequency-Inverse Document Frequency (TF-IDF) the hybrid model uses TF-ICF which brings class information from training data.

$$TF-IDF(T_i,d_j)=tf(T_i,d_j)\times log\left(\frac{N}{N(T_i)}\right) \qquad (4)$$

Where $tf(T_i,d_j)$ represents the number of times term T_i occurs in documents d_j , N denotes the number of documents, and $N(T_i)$ stands for the number of documents in which term T_i occurs at least once.

In fact, TF-IDF, that computes term weighting scores regardless the class information of documents, can't effectively deal with crime datasets which are imbalanced.

The next steps aim at including class category information from training data to provide the new term weighting ICF (inverse class frequency). This basically implies that the new ICF is class category specific and is computed using the class terms frequency vector V_{tf_k} (5) based on document term frequency vectors V_{dtf_j}

In step 4, the class terms frequency vector V_{tf_k} is generated. For each class C_k , V_{tf_k} presents the number of times that term T_i occurs in training data of class C_k. The class term frequency is obtained from V_{dtf_j} as follows:

$$V_{tf_k}=\sum_{j\in N_{dk}} V_{dft_j} \qquad (5)$$

Where V_{tf_k} is the class term frequency of class C_k , and N_{dk} denotes the number of training document of C_k and V_{dtf_j} is the document frequency term for document j computed in (3). $M_{TF}(N_C,N)$ is defined as terms frequency matrix representing all V_{tf_k} vectors where N_C stands for the number of classes and N is the number of terms T_i.

Finally, in step 5, the ICF is computed for each term T_i as follows:

$$ICF(T_i)=log\left(\frac{Nc}{\sum_{i}^{Nc}\alpha}\right) \qquad (6)$$

Where α takes 0 if term T_i does not appear in class C_k, and 1 in otherwise. The new ICF boosts the importance of terms appearing only at one class and penalizes irrelevant terms.

The final $M_{TF-ICF}\,(N_C,N)$ matrix is obtained by

$$M_{TF-ICF}=M_{TF}\times M_{ICF} \qquad (7)$$

Where the $M_{ICF}(N,N)$ is the diagonal matrix of ICF.

As mentioned earlier, our approach proceeds in three phases (shown in Fig.1) as follows:

Phase 1: Twitter Data Collection

The data collection is done from twitter through the use of the official Twitter Search API v1.1. The Twitter API allows real time access and extraction of tweets according to a specific query. With more than 50 requests, we create three crime different datasets. The two first datasets consider the major crime types (Homicide, Rape, Robbery, Assault, Kidnapping). whereas the third one is a Hate Crime Twitter sentiment (HCTS) dataset with different aspects of Hate Crime as racism, terrorism, religious tolerance… The obtained datasets contain two types of tweets: (1) irrelevant tweets which refer to contexts not related to crimes (i.e., movies, games) or tweets without implicit aspects and (2) implicit aspect Crime tweets. Furthermore, certain tweets contain grammatical and spelling mistakes, abbreviations, URLs, sources of data, hashtags… These hurdles are addressed by the preprocessing step of the IASD phase to ensure better crime implicit aspect identification.

Phase 2: Implicit Aspect Sentence Detection

IASD phase, as shown in figure 3, consists of preprocessing and sentence relevancy classification process:

1) Preprocessing

The first step of the preprocessing is the removal of noisy data. The process begins with the removal of URL, @usernames and #hashtags. Then, the Part of speech tagger (POS) is used to parse tweets to extract adjectives and verbs as they represent potential implicit aspect terms implying crimes. For the elongate extracted terms, with more than three following occurrence of the same letter, we applied the compression words process commonly used for tweets. It's used to obtain the right form of word acceptable by the WordNet dictionary. At last, the stop words are removed from tweet datasets.

2) Sentence relevancy classification

Sentence Relevancy Classification, which encompasses two sub-steps, focuses on classifying relevant/irrelevant tweets in order to create an implicit aspect crime corpus from each dataset. The first sub-step preprocesses tweet datasets and uses the proposed hybrid model to enhance training data. The second sub-step employs the improved training data to build a classification model for crime implicit aspect sentences and then generate crime implicit aspect corpora.

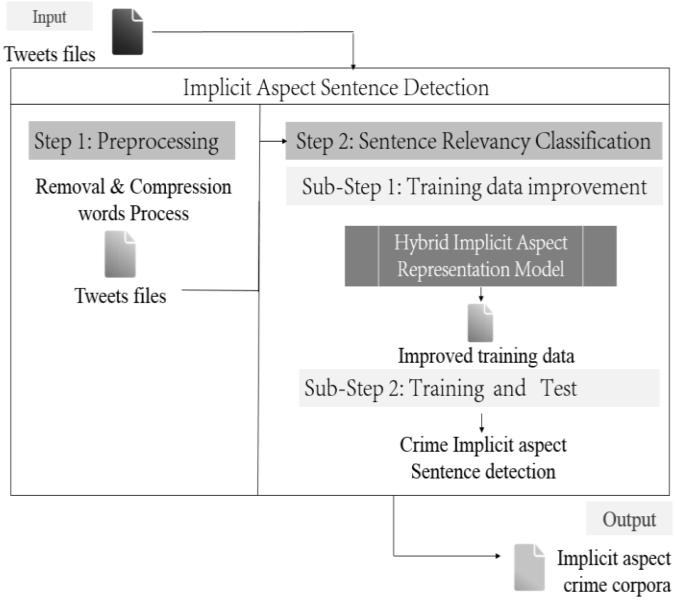

Input
Tweets files
Implicit Aspect Sentence Detection
Step 1: Preprocessing
Removal & Compression words Process
Tweets files
Step 2: Sentence Relevancy Classification
Sub-Step 1: Training data improvement
Hybrid Implicit Aspect Representation Model
Improved training data
Sub-Step 2: Training and Test
Crime Implicit aspect Sentence detection
Output
Implicit aspect crime corpora

Fig. 3. Crime Implicit Aspect Sentences Detection using Hybrid Model.

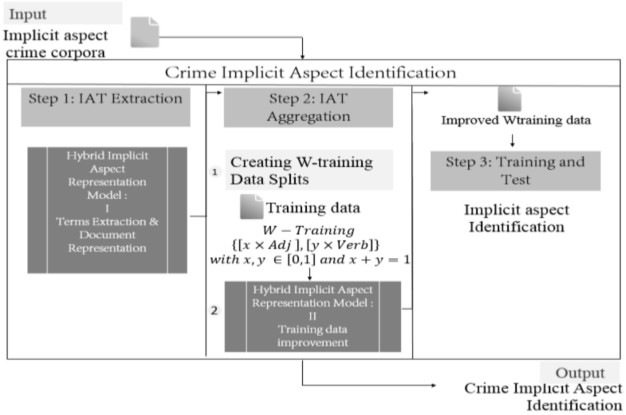

Input
Implicit aspect crime corpora
Crime Implicit Aspect Identification
Step 1: IAT Extraction
Hybrid Implicit Aspect Representation Model : I Terms Extraction & Document Representation
Step 2: IAT Aggregation
1
Creating W-training Data Splits
Training data
$W - Training$
$\{[x \times Adj], [y \times Verb]\}$
$with\ x, y \in [0,1]\ and\ x + y = 1$
2
Hybrid Implicit Aspect Representation Model : II Training data improvement
Improved Wtraining data
Step 3: Training and Test
Implicit aspect Identification
Output
Crime Implicit Aspect Identification

Fig. 4. Crime Implicit Aspect Identification using Hybrid Model.

Phase 3: Crime implicit aspect identification

As shown in Fig.4, the task aims at extracting crime implicit aspects from corpora prepared in phase 2. IAT extraction and aggregation are addressed using the two steps of the proposed hybrid model.

In IAT Extraction, the Terms Extraction & Document Representation steps of the hybrid model are applied to extract potential implicit aspect implied by adjectives and verbs. Then, for each dataset document, the hybrid model provides the document term frequency vectors V_{dtf_j}, which represents the contribution of adjectives and verbs and their WN extracted terms for a given document.

In IAT aggregation, the Training data improvement steps of the hybrid model are applied using several W-Training data splits. These splits are obtained using weighting schema assigning different weights for adjectives and verbs. This weighting schema is used to evaluate the impact of using different proportions of adjectives and verbs on the improvement of training data for crime datasets. Each W-Training data split is computed by equation 9 as follows:

$$W - Training\ split = \{[x \times Adj\ ], [y \times Verb]\} \quad (8)$$

$$where\ \ x, y\ \in [0,1]\ and\ x + y = 1$$

IAT aggregation task aims at identifying the implicit aspect for each document. To this end, IAT aggregation uses weighting model that measures the document terms reliability according to a given implicit aspect (class). Thus, IAT aggregation computes term matrix frequency M_{TF-ICF} , that reflects the term's strength of representing a specific class.

IV. Experiments and Evaluation

The experiments conducted to validate the proposed approach are presented in this section with the experimental design adopted, i.e. the pre-processing techniques utilized, classifies used, datasets chosen, the performance evaluation metrics used, and the results obtained based on those measures with the discussion.

A. Experimental Setup

1) Preprocessing: After gathering data from Twitter by means of the Twitter API within data collection phase, the preprocessing is done. it applied filtering text techniques to obtain a clear text without irrelevant content. At last, the POS tagger is used for parsing data and extracting a list of adjectives and verbs used at sub-step 1 of Sentence Relevancy Classification process.

2) Classifiers used: Three supervised classifiers are used to validate the proposed approach: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Random Forest (RF).

Multinomial Naïve Bayes is the most variation of NB that is mostly used in text categorization and sentiment analysis [24]. MNB is a probabilistic model based on the Bayes theorem. It uses the joint probabilities of features and categories to estimate the probabilities of classes given a document and makes the assumption that features are conditionally independent of each other to make the computation of joint probabilities simple.

In text categorization and sentiment analysis, *Support Vector Machine* is often considered as the best classifier providing the greatest performances for those tasks [25]. It's among the class of classifiers based on kernel substitution [26].In this work, the version Sequential Minimal Optimization (SMO) developed in [27] is used.

Random Forest is a popular tree classifier based on many classification trees, used for text categorization and sentiment analysis for Twitter [19], [28]. The forest construction is the base step in this classification. Each individual tree is constructed based on two procedures proposed by [29]*:* (1) to create decision tree nodes, subspace of features is randomly chosen, then (2) to generate training data subsets for building individual trees, the classifier relies on bagging method and finally (3) to obtain the random forest classifier all individual trees created are combined.

3) Datasets: The proposed approach is assessed using crime datasets collected and prepared in this work. The three crime datasets are extracted from twitter with different size and aspects.

The first crime dataset contains 2k tweets, of which 357 include implicit aspect sentences involving adjectives and verbs. The dataset covers the four major crime types namely, homicide, rape, robbery and aggravated assault.

The second dataset considers more specific type of crime as shooting, kidnapping, vehicle theft, violent crime, rape and homicide. It contains more than 600 implicit aspect sentences extracted from 3k tweets.

The hate crime dataset involves 6k tweets of which 648 include implicit aspect sentences and cover different predefined aspect racism, disability abuse, religious tolerance, terrorism and rape.

4) Evaluation measures: To evaluate the performance obtained after using the proposed approach, we use the standard metric F1-score which is commonly used to evaluate the classification task. F1, introduced by Van Rijsbergen [30] is the equally weighted average of recall and precision as stated in (9). The Recall is defined to be the ratio of correct assignments by the system divided by the total correct assignments. The Precision is the proportion of correct assignments by the system within the total number of the system's assignments. All experiments are carried out using Weka platform [31]. We use the 10 Fold cross validation to reduce the uncertainty of data split between training and test data.

$$F1 = \frac{2\times Precision \times Recall}{Precision+Recall} \quad (9)$$

B. Experimental Protocols

Experiments aimed at evaluating the effectiveness of the proposed approach for Crime IASD and crime IAI.

1) Crime implicit aspect sentence detection IASD: Experiments have been conducted according to three points relating to performance of the three classifiers for Crime IASD:

a) The use of TF-IDF versus TF-ICF for document representation

The first point pertains to evaluating and comparing the impacts of using TF-IDF and TF-ICF on the performances for the classifiers for IASD. Two categories of experiments are defined:

- TF-IDF (denoted baseline): it refers to the use of the three classifiers with hybrid model and without considering WordNet semantic relations. TF-IDF uses only the terms extracted from datasets and presents documents using the Term Frequency-Inverse Document Frequency vector model.
- TF-ICF: it represents the use of the three classifiers with hybrid model and without considering WordNet semantic relations. TF-ICF uses only the terms extracted from datasets and presents documents using the Term Frequency-Inverse Class Frequency vector model.

b) The integration of WordNet Synonym relations of adjectives and verbs in the document representation model

The second point relates to the comparison of the impacts on the performances of the classifiers for IASD using TF-IDF and TF-ICF with the integration of WordNet synonym relations for adjectives and verbs. Synonyms are considered here due to their wide use in SA. Two types of experiments are defined:

- TF-IDF+ Synonyms: it concerns the use of the three classifiers with the hybrid model using TF-IDF and integrating synonyms of adjectives and verbs.
- TF-ICF+ Synonyms: it refers to using the three classifiers with the hybrid model using TF-ICF and integrating Synonyms of adjectives and verbs.

c) The integration of the best WN subsets of adjectives and verbs in the document representation

The third point is similar to the second point except here the integration of WordNet relations concerns the best WN subsets for adjectives and verbs. Two types of experiments are defined:

- TF-IDF + Best-WN-subsets represents the use of the three classifiers with the hybrid model using TF-IDF and integrating the best WN subsets (subsets D and D-S for adjectives and verbs respectively).
- TF-ICF + Best-WN-subsets represents the use of the three classifiers using TF-ICF and integrating the best WN subsets (subsets D and D-S for adjectives and verbs respectively).

Experiments have been conducted according to two points that evaluate the effectiveness of the proposed approach for Crime implicit aspect sentence detection IASD and crime implicit aspect identification IAI.

2) Crime implicit aspect identification IAI: Experiments have been conducted according to two points relating to performance of the three classifiers for Crime IAI:

a) The use of adjectives and verbs for training data enhancement

The first point concerns the comparison of the impacts of adjectives and verbs on training data improvement. Experiments are done here using several W-Training data splits. These splits are obtained using weighting schema assigning different weights for adjectives and verbs. For each dataset, six testing datasets are prepared where each set combines adjectives and verbs with different weighting. These weightings are defined as follows: $(1\,,0)$, $(0.8\,,0.2)$, $(0.6\,,0.4)$, $(0.4\,,0.6)$, $(0.2\,,0.8)$ and $(0{,}1)$. Three experiments are defined: MNB, SVM and RF that respectively refers to MBN, SVM and RF classifier using hybrid model, W-training data and integrating the best WN subsets of adjectives and verbs.

b) The Absence of WN terms of adjectives and verbs in the document representation

The second point deals with the effects on classifiers performances of the absence of WN terms of adjectives and verbs in the hybrid model. Three experiments are defined: MNB_{NoWn}, SVM_{NoWn} and RF_{NoWn} that respectively represents MNB, SVM and RF using the hybrid model with W-training data and without the best WN subsets of adjectives and verbs.

C. Results and Discussion

In this section, experiments results are presented according to the points mentioned in the experimental protocols section.

1) Crime implicit aspect sentence detection IASD: The three classifiers are assessed for crime Implicit aspect sentence detection using three crime datasets with varying sizes presented in table 1. Table 2 shows the performances with the Average Improvement Rates (AVG.Imp.R) of the three classifiers obtained from experiments related to the three points above mentioned in the experimental protocols for this phase.

TABLE I. SIZE OF DATASETS

	Crime dataset 1	*Crime dataset 2*	*Hate crime dataset*
Number of sentences	2k	3k	6k
Number of implicit aspect sentences	357	641	648
Number of irrelevant sentences	1643	2359	5352
Number of Training data for implicit aspect	180	350	300
Number of Training data for implicit aspect	670	1500	3500

TABLE II. MNB, SVM AND RF FOR RELEVANT / IRRELEVANT CLASSIFICATION

	Crime dataset 1			*Crime dataset 2*			*Hate crime dataset*		
	MNB	*SVM*	*RF*	*MNB*	*SVM*	*RF*	*MNB*	*SVM*	*RF*
(1)	0.51	0.63	0.63	0.52	0.65	0.63	0.63	0.69	0.68
(2)	0.57	0.65	0.65	0.59	0.68	0.68	0.74	0.77	0.75
(1)/(2)	11.6 %	3.1%	3.1%	13.4 %	4.6 %	7.9 %	17.4 %	11.5 %	10.2 %
(3)	0.71	0.68	0.68	0.69	0.71	0.69	0.76	0.78	0.78
(4)	0.74	0.71	0.71	0.74	0.74	0.72	0.74	0.78	0.78
(5)	0.78	0.71	0.71	0.74	0.73	0.73	0.80	0.81	0.80
(6)	**0.83**	**0.88**	**0.87**	**0.79**	**0.80**	**0.79**	**0.82**	**0.89**	**0.87**
(3)/(5)	9.8%	4.4%	4.4%	7.2%	2.8 %	5.7 %	5.2%	3.8%	2.5%
(4)/(6)	12.1 %	23.9 %	22.5 %	6.7%	8.1 %	9.7 %	10.8 %	14.1 %	11.5 %

(1) Baseline, (3) TF-IDF+Synonyms, (5) TF-IDF + Best WN Subsets
(2) TF-ICF, (4) TF-ICF+Synonyms, (6) TF-ICF + Best WN Subsets (The proposed Hybrid Model)

Firstly, it can be seen, from table 2, that the use of TF-ICF helps better the three classifiers deal with IASD than using TF-IDF for the three datasets. In fact, TF-IDF does need cope effectively with document representation for the three datasets because these latter are class imbalanced. Normally, terms with low TF-IDF are considered irrelevant terms since they appear in large part of documents. This is not definitely true for imbalanced datasets, because although these terms occur more often in one class than others they are relevant and important to distinguish between classes. On the contrary of TF-IDF, TF-ICF takes advantage of those unevenly distributed words by considering term contribution in class representation rather than document representation.

Secondly, table 2 shows that the integration of WN synonyms helps the three classifiers improve their performances for IASD when using TF-IDF and TF-ICF. Moreover, the use of TF-ICF is proven to be consistently more helpful for the three classifiers than using TF-IDF even with the integration of WordNet synonyms.

Thirdly, table 2 proves that the integration of the best WN subsets allows the three classifiers achieve their best performances for both TF-IDF and TF-ICF cases. Also, using TF-ICF is shown to help the three classifiers achieve better performances than using TF-IDF.

In fact, the integration of WN semantic relations promotes training data vocabulary by creating a large set of relevant terms that support system to learn better from data. However, the selection of WN semantic relation is crucial. The integration of synonym relation allows classifiers to achieve better scores, yet, it induces more noisy terms than definition subsets. WN semantic relations for adjectives and verbs must be appropriately selected (subsets D for adjectives and subsets D-S for verbs) so that they can help the classifiers achieve their best performances for IASD.

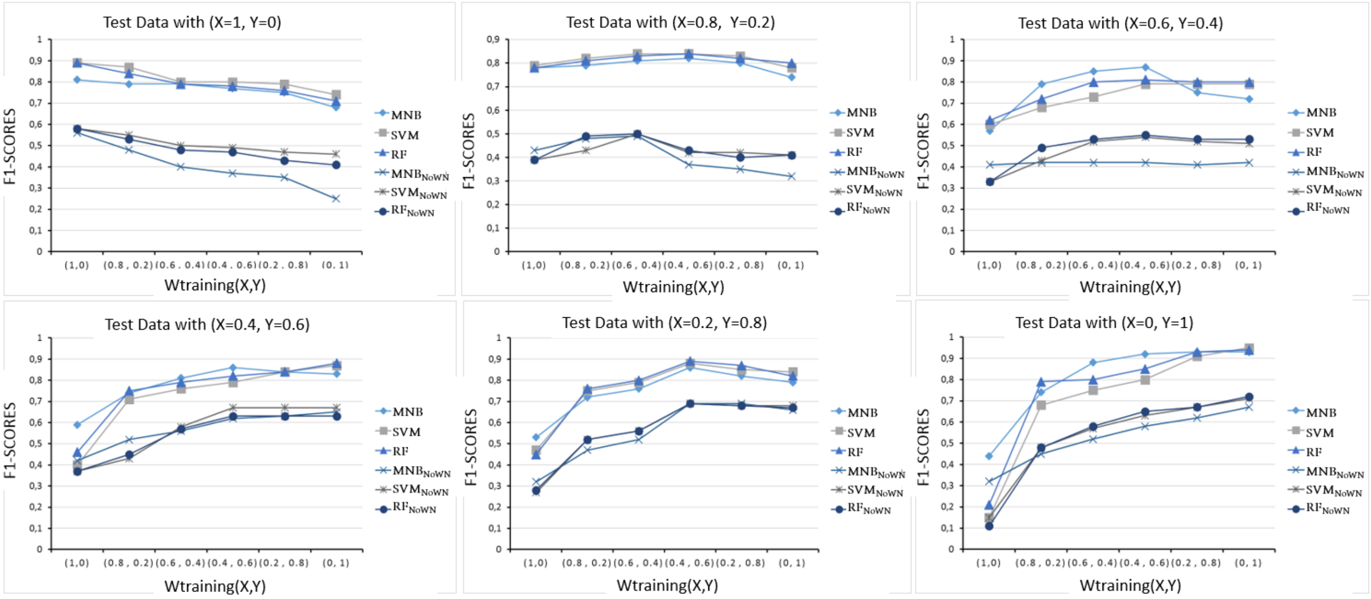


Fig. 5. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on **CRIME CORPUS 1** for IAI Phase.

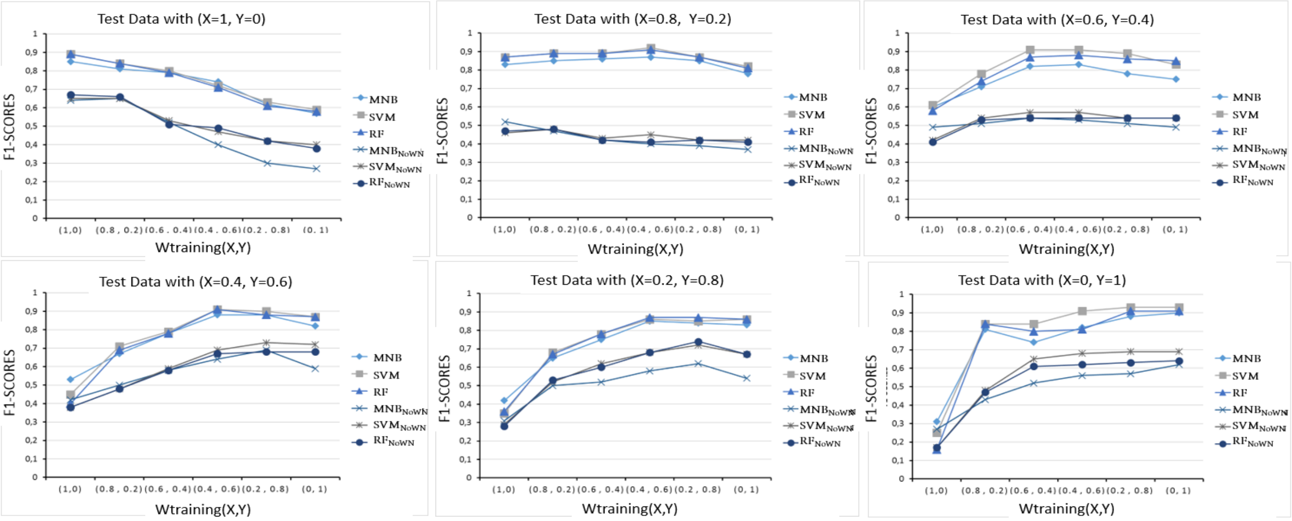


Fig. 6. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on CRIME CORPUS 2 for IAI Phase.

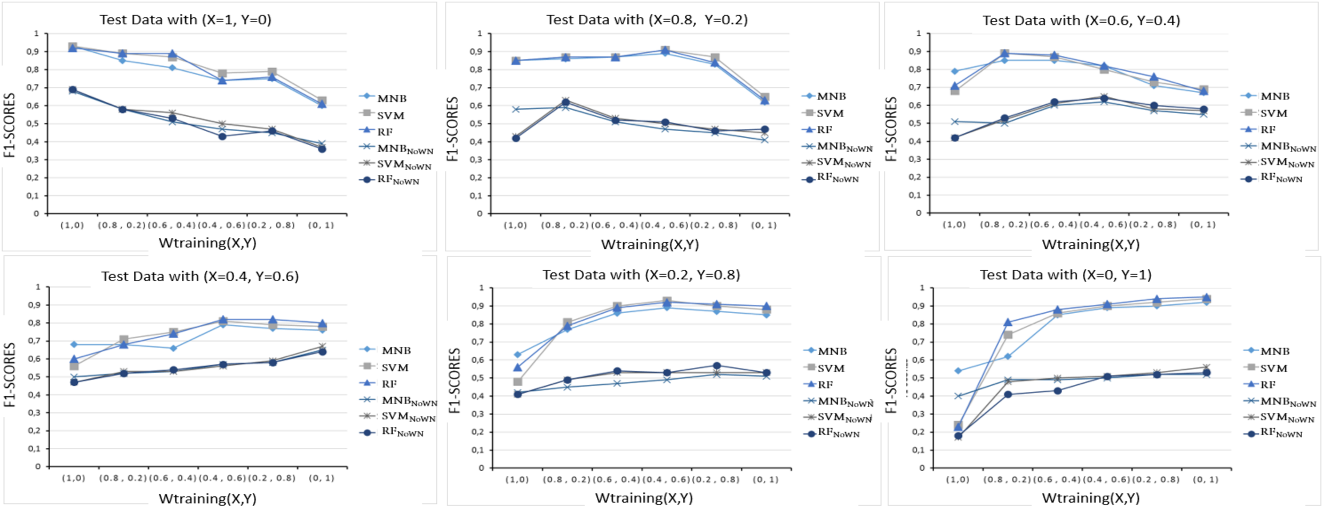


Fig. 7. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on HATE CRIME CORPUS for IAI Phase.

2) Crime implicit aspect identification IAI: Fig. 5, 6 and 7 show the performances of the three classifiers obtained from experiments pertaining to the two points already introduced in experimental protocols for this phase. For each testing data, X-axis denotes the different weights (x, y) assigned to adjectives and verbs and Y-axis indicates F-1 performances. The number of adjectives and verbs of the three crime corpora is shown in table 3.

TABLE III. NUMBER OF ADJECTIVES AND VERBS IMPLYING IMPLICIT ASPECT FOR EACH CRIME DATASET

	Crime dataset 1	*Crime dataset 2*	*Hate crime dataset*
Number of sentences	357	641	648
Number of adjectives	406	841	773
Number of verbs	446	872	729

As shown from Fig. 5, 6 and 7, the use of different weights (x, y) assigned to adjectives and verbs leads to variant F1-performances for MNB, SVM and RF.

The case of training and test data involves adjectives only (test data with (x=1, y=0) and W-training with (1,0)) leads to the best performances for all classifiers. One unanticipated finding is that when considering only verbs for training (Wtraining with (0,1)), all classifiers are able to achieve considerable F1- performances that exceed 60% for implicit aspect identification implied by adjectives.

In contrast, adjectives do not support verb identification (test data with (x=0, y=1)). In this test data case and for the three datasets, classifiers achieve their worst performances when verbs are completely absent in training data (W-training with (1 ,0)). For the same test data, the best F-1 scores are attained when considering only verbs for training (Wtraining with (0,1)).

Using adjectives in training data supports IAI involving more adjectives than verbs (test data with (x=1, y=0) and (x=0.8, y=0.2)). Implicit aspect identification including verbs is known to be more challenging than adjective. Using only adjectives in training to predict implicit verbs does not support classifiers identifying the implicit aspect for crime datasets. However, using verbs for implicit aspect identification is more beneficial for classifiers. This can be explained by the fact that, verbs used to imply a crime aspect are more descriptive and useful than adjectives. In other words, for each crime aspect there are a number of verbs specifically used to imply this aspect, for example '*to kill'*, '*to kidnap'* and *'to steal'*, each verb is used to imply a single type of crime which is *'Homicide'*, *'Kidnapping'* and *'Robbery'* respectively. However, as often happens, one adjective can be used to imply different crime aspects such as *'blooded', 'atrocious'*, '*hostile'*, *'agonizing'*, *'cruel'* that can be used not only for *'homicide'* but for *'violent crime'* and *'kidnapping'* as well. As a result, adjective extracted terms can represent more than one aspect. However, when considering more verbs for training than adjectives, the WN extracted terms are more descriptive and contain more reliable terms that better represent the implicit aspect which supports adjective and verb identification.

For training and test data using a combination of adjectives and verbs, and for the same reasons explained above, the highest performance is achieved in general when considering training with more verbs than adjectives. Overall, for the three datasets, the best performing W-training is $(0.4, 0.6)$.

On the other hand, for each testing data of the three crime corpora, $\mathrm{MNB}_{\mathrm{NoWn}}$, $\mathrm{SVM}_{\mathrm{NoWn}}$ and $\mathrm{RF}_{\mathrm{NoWn}}$ have the same behavior than MNB, SVM and RF, but the performances reached, without WN extracted terms for verbs and adjectives, are consistently lower.

Considering more verbs than adjectives in training data supports implicit aspect identification for adjectives and verbs. While using more adjectives for learning conducts to better classifiers performances for test data involving only adjectives. However, the observed decrease in F1-performances can be attributed to the lack of WN extracted terms. Without WN, classifiers are not able to enlarge training vocabulary. This makes it extremely hard to identify adjectives and verbs appearing only twice in datasets. Even worse, it's completely impossible to identify terms appearing only once either in training or test set. The Absence of WN terms of adjectives and verbs severely penalizes performances of all classifiers for crime IAI. Hence, considering a weighted training data based on verbs and their WN extracted terms not only is required and undeniable but also improves the performance of the considered classifiers for implicit aspect identification for crimes.

Finally, Fig. 8 presents the extracted implicit aspects of the three considered crime datasets using the proposed framework.

V. Conclusion

We presented a hybrid approach for training data improvement of MNB, SVM and RF classifiers to address Aspect Based Sentiment Analysis for Crime datasets. We conduct an empirical and analytical study at the level of:

1) The crime implicit aspect sentence Detection IASD phase, where experiments are conducted according to three points: (1) the use of TF-IDF versus TF-ICF for document representation (2) The integration of WordNet Synonym relations of adjectives and verbs in document representation model and (3) The integration of the best WN subsets of adjectives and verbs in document representation.

2) The crime implicit aspect identification IAI phase, where experiments are carried out according to two points: (1) The use of adjectives and verbs for training data enhancement and (2) The Absence of WN terms of adjectives and verbs in document representation.

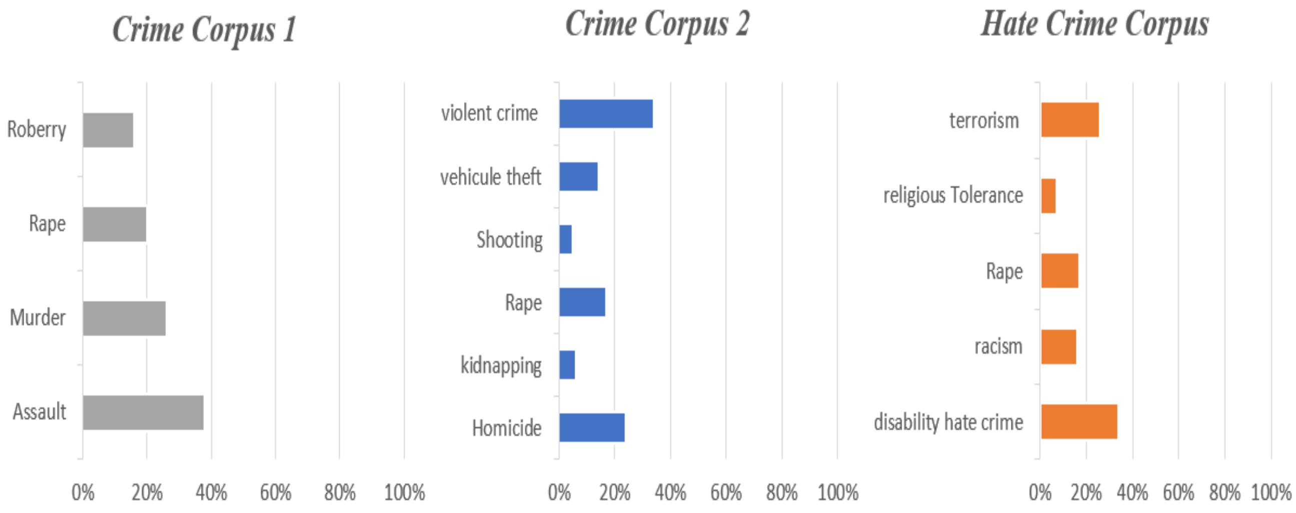


Fig. 8. Percentage of Implicit Aspects of Crime for the Three Crime Datasets.

The major findings of the work include:

- For the three imbalanced crime datasets, using TF-ICF is shown to help the three classifiers achieve better performances for IASD than using TF-IDF. This is true with and without the integration of WordNet terms.
- Using the synonyms relations for adjectives and verbs are shown to support better classifiers for IASD phase.
- Using an appropriately selected WN semantic relations for adjectives and verbs (Best WN subsets) improves training data for crime IASD and IAI and thus helps classifiers performing better for these two phases.
- Comparing to adjectives, verbs and their WN extracted terms are empirically proven to be as the key element for training data enhancement that allows classifiers to be more performant for crime implicit aspect identification.

Further work will investigate those findings to deal with the problem of the identification of crimes committed by the same individual or same group which became an important and challenging task of crime prevention systems.

Another interesting future perspective is applying the proposed approach for crime detection from variant resources of data such as weather data which significantly influence crime rates and criminal behavior.

REFERENCES

[1] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[2] Doaa Mohey El-Din, “Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis” International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016.

[3] B. Liu, Sentiment analysis: mining opinions, sentiments, and emotions. New York, NY: Cambridge University Press, 2015.

[4] X. Chen, Y. Cho, and S. Y. Jang, “Crime prediction using Twitter sentiment and weather,” 2015, pp. 63–68.

[5] Jermy Prichard, Paul Watters, Tony KRONE, Caroline Spiranovic, and Helen Cockburn, “Social Media Sentiment Analysis: A New Empirical Tool for Assessing Public Opinion on Crime?,” pp. 217–236, 2015.

[6] Nisal Waduge, “Machine Learning Approaches For Detect Crime Patterns - Data Gathering and Analysing Techniques,” 2017.

[7] P. Burnap et al., “Detecting tension in online communities with computational Twitter analysis,” Technol. Forecast. Soc. Change, vol. 95, pp. 96–108, Jun. 2015.

[8] N. Zainuddin, A. Selamat, and R. Ibrahim, “Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach,” in Intelligent Information and Database Systems, vol. 9621, N. T. Nguyen, B. Trawiński, H. Fujita, and T.-P. Hong, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 151–160.

[9] Hissah AL-Saif and Hmood Al-Dossari, “Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques” International Journal of Advanced Computer Science and Applications(IJACSA), 9(10), 2018.

[10] B. Keith, E. Fuentes, and C. Meneses, “A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews,” N. S., p. 10, 2017.

[11] K. Schouten, P. O. Box, and D. Rotterdam, “Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data,” IEEE Trans. Cybern., p. 13, 2017.

[12] V. S. Jagtap and K. Pawar, “Sentence-Level Analysis of Sentiment Classification,” Natl. Conf. Emerg. Trends Eng. Technol. Archit., p. 6, 2013.

[13] K. Gull, S. Padhye, and D. S. Jain, “A Comparative Analysis of Lexical/NLP Method with WEKA’s Bayes Classifier,” Int. J. Recent Innov. Trends Comput. Commun., vol. 5, no. 2, p. 7, 2017.

[14] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[15] A. Alghunaim, M. Mohtarami, S. Cyphers, and J. Glass, “A Vector Space Approach for Aspect Based Sentiment Analysis,” 2015, pp. 116–122.

[16] G. Abdulsattar A. Jabbar Alkubaisi, S. Sakira Kamaruddin, and H. Husni, “Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers,” Int. J. Eng. Technol., vol. 7, no. 2.14, p. 57, Apr. 2018.

[17] V. N. Patodkar and S. I.R, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” IJARCCE, vol. 5, no. 12, pp. 320–322, Dec. 2016.

[18] S. Rosenthal, N. Farra, and P. Nakov, “SemEval-2017 Task 4: Sentiment Analysis in Twitter,” p. 17, 2017.

[19] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and As. Perera, “Opinion mining and sentiment analysis on a Twitter data stream,” 2012, pp. 182–188.

[20] M. Ishtiaq, “Sentiment Analysis of Twitter Data Using Sentiment Influencers,” vol. 6, no. 1, p. 9, 2015.

[21] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” p. 6, 2012.

[22] Y. Liu, H. T. Loh, and A. Sun, “Imbalanced text classification: A term weighting approach,” Expert Syst. Appl., vol. 36, no. 1, pp. 690–701, Jan. 2009.

[23] El Hannach, H. and Benkhalifa , M., “Using Synonym and Definition WordNet Semantic relations for implicit aspect identification in Sentiment Analysis,” Pap. Present. 1st Int. Conf. Netw. Inf. Syst. Secur. NISS 2018 Conf. Tangier Morocco., p. 8, 2018.

[24] Junseok Song, Kyung Tae Kim, Byungjun Lee, Sangyoung Kim, and Hee Yong Youn, “A novel classification approach based on Naïve Bayes for Twitter sentiment analysis,” KSII Trans. Internet Inf. Syst., vol. 11, no. 6, Jun. 2017.

[25] R. Sergienko, M. Shan, and A. Schmitt, “A Comparative Study of Text Preprocessing Techniques for Natural Language Call Routing,” in Dialogues with Social Robots, vol. 427, K. Jokinen and G. Wilcock, Eds. Singapore: Springer Singapore, 2017, pp. 23–37.

[26] G. Loosli, S. Canu, and L. Bottou, “Training Invariant Support Vector Machines using Selective Sampling,” p. 26, 2005.

[27] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to Platt’s SMO Algorithm for SVM Classifier Design,” Neural Comput., vol. 13, no. 3, pp. 637–649, Mar. 2001.

[28] B. Xu, X. Guo, Y. Ye, and J. Cheng, “An Improved Random Forest Classifier for Text Categorization,” J. Comput., vol. 7, no. 12, Dec. 2012.

[29] LEO BREIMAN, “Random Forests,” Machine Learning, The Netherlands, pp. 45, 5–32, 2001.

[30] C. J. van RIJSBERGEN, “Information Retrieval,” 2nd ed. Butterworth-Heinemann, 1979.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” ACM SIGKDD Explor. Newsl., vol. 11, no. 1, p. 10, Nov. 2009.

3D Mapping based-on Integration of UAV Platform and Ground Surveying

Muhammad Yazid Abu Sari[1], Asmala Ahmad[7],
Rozilawati Dollah[8]
Centre of Advanced Computing Technology, Faculty of
Information Technology and Communication, Universiti
Teknikal, Malaysia[1, 7]
School of Computing Faculty of Engineering, UTM Johor
Bahru, Johor, Malaysia[8]

Abd Wahid Rasib[2], Hamzah Mohd Ali[3], Abdul Razak
Mohd Yusoff[4], Muhammad Imzan Hassan[5],
Khairulnizam M.Idris[6]
Tropical Map Research Group
Faculty of Built Environment and Surveying, UTM
Johor Bahru, Johor, Malaysia

***Abstract*—Development in aerial photogrammetry technology has contributed a notable impact to the area of large-scale mapping. Nowadays, unmanned aerial vehicle (UAV) platform has become a significant tool in aerial mapping. Generating 3D mapping using photos acquired from UAV is more preferable due to its low cost and flexible operation. Therefore, this study aims to develop a technique for 3D mapping with an integration of UAV aerial photos and detailed ground survey. The produced 3D mapping has RMSE(x) = 0.279, RMSE(y) = 0.215, and RMSE (z) = 1.341 using 25 randomly selected sample points. Besides that, the result shows the location parameters i.e. x, y and z were also positively correlated, t-test(x) = 0.961, t-test(y) = 0.250 and t-test (z) = 1.885, respectively.**

***Keywords*—3D mapping; UAV platform; ground survey; aerial photo**

I. Introduction

The rapid advancement in aerial photogrammetry technology, particularly in the area of digital scanning has produced a new means to create a photorealistic 3D model. 3-dimensional image or also known as 3D image is commonly understood when an image contains the information regarding the depth or height of objects within the image. In realising 3D image production using aerial photos, flight planning parameters and ground control points (GCP) need to be taken into account. GCP is used for georeferencing, which is linking photos with spatial locations. These photos need to further undergo stitching process in order to produced seamless aerial photos of an area, known as aerial orthophoto. The orthophoto can be further processed to generate digital surface model (DSM) and digital terrain model (DTM) in which to be used in producing the 3D mapping. A number of studies related to 3D mapping based on aerial can be found however, accuracy is still an issue. There are effort to look into this issue by researchers in manufacturing-based countries such as USA, China, Japan and UK. Nevertheless, not much effort have been carried out in developing countries such as Malaysia despite having different condition in terms of climate, terrains and land covers in which may have effects on the accuracy. Therefore, this study attempts to further look into the accuracy issue by making use of an approach where aerial photos and ground details are integrated in producing 3D maps.

Nowadays, 3D city mapping can be easily implemented with the advancement of low cost unmanned aerial vehicle (UAV). A 3D city mapping represents an urban environment with three-dimensional geometry of structures and urban objects, with the building as the most prominent features [1-4]. A typical 3D city mapping is derived from various data acquisition techniques, for instance, photogrammetry and laser scanning [5–8], extrusion from 2D footprints [9,10], synthetic aperture radar [11–15], architectural models and drawings [16–18], handheld devices [19,20], procedural modelling [21–26], and volunteered geoformation [27–29].

Meanwhile, the acquisition of aerial photos using UAV platform becomes an efficient optional compared to other methods. Here, UAV allows a higher degree of automation in aerial photos collection [30]. Furthermore, UAV offers a low-cost alternative and a real-time application compared to other classical manned aerial photogrammetry [31]. In addition, UAV photos are also have been proved that applicable in generating high-quality 3D view [31-33]. Subsequently, UAV flown at low altitude (20-30 meters) can produce a ground pixel less than 5 x 5 cm which gives a better resolution for low-cost mapping approach [34].

Numerous research studies have been conducted in term of UAV photogrammetry [35-40] for 3D mapping and modelling. For example, Hudzietz and Saripalli [41] have successfully employed the structure from motion (SfM) techniques for the reconstruction of aerial landscapes imagery. Zhiguang Ding and other scholar verified the superiority of UAV for rapid modelling compared to artificial modelling where the aerial photos were ultimately converted into a 3D image of terrain [42]. The researchers proved that UAV has been used as a platform for aerial photo acquisition. This offers effective and accurate large-scale terrain modelling with cost-efficient solution. Therefore, this study aims to produce a 3D map based on the integration of UAV platform and ground detail survey at 725 acres of the main campus in UTeM, Melaka, Malaysia.

II. Issues

Unmanned aerial vehicle (UAV) is a flexible aerial photography platform widely used in various applications such as mapping, agriculture, 3D documentation and more. Due to benefits provided by UAV, there is no doubt that UAV can be used as tools and sources for data collection. However, despite the numerous benefits, UAV photos are exposed to degradation resulted from issues related to imaging stability. Generally, ordinary UAV considered as lightweight platform which highly influenced by environmental conditions, such as wind [43]. During flight, altitude instability can be caused by wind and other weather factors may lead to a large rotation angle for UAV platform. Large rotation angle from UAV will decrease aerial photo quality and results in photo deformation that will affect their overlaps during stitching process. The insufficient feature overlap between photos may lead to the failure of stitched image reconstruction [44]. Based on the mentioned issue, this will greatly affect the accuracy of the end product that are produced for mapping and 3D documentation. Thus, this study attempts to show the accuracy of 3D mapping that was developed using aerial photos from UAV.

III. Study Area

The location of the study area is UTeM main campus as shown in Figure 1, located in Durian Tunggal, Malacca (Latitude, 2.309980, Longitude 102.317672). The campus coverage area approximately 725 acres and situated near to the bustling historical city of Malacca. UTeM campus which consists of 23 main building and 7 of it is faculty building. For this study, the built-up area at the center of UTeM campus was the main focus where the major buildings are chancellery building, student activity center, mosque, and main hall respectively. The tallest building that recorded for this study is the main hall of UTeM which is 20 meters above mean sea level (MSL). UTeM campus can be easily accessed through Ayer Keroh Toll exit and approximately 20 minutes away from the main city.

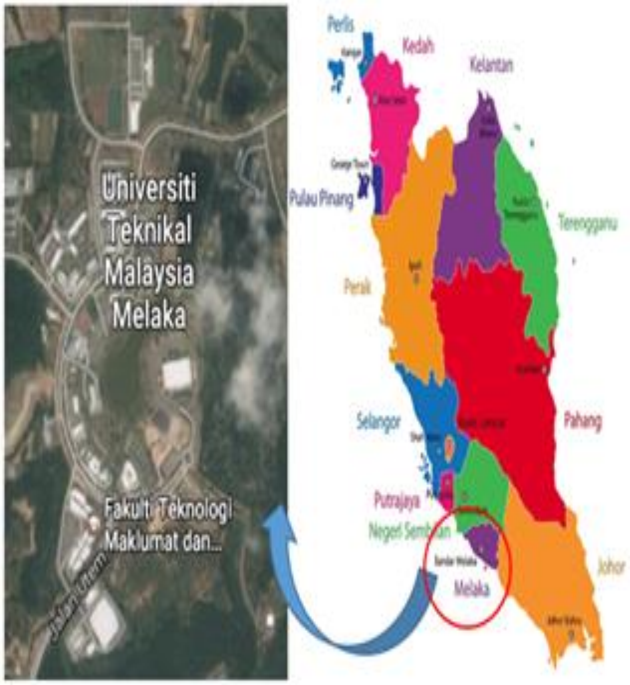


Fig. 1. Area of Study (Source Google Earth).

IV. Methodology

This part describes the workflow for this study. During the early stages of the study, a set of the flowchart is created to make sure the study conducted properly and efficiently. based on Figure 2, the research methodology consists of four (4) main phases which are preliminary study, data acquisition, data processing and results. The first phase is the preliminary study. This phase generally shows the planning process and how the study will be conducted as shown in Figure 2. Preliminary study is essential before conducting any research studies.

In Phase 2, data acquisition focuses mainly the process in obtaining the required data for the study which is ground survey and aerial photos using UAV platform. In ground survey, first, the establishment of ground control point will be conducted followed by detailed survey as the control point will provide referral coordinates for the detailed survey. The flight planning for UAV is created using photogrammetry system and camera calibration for the non-matrix camera is carried out in order to reduce the bundle of error for the orthophoto generation.

Phase 3 describes the process of data processing for ground survey data and aerial photo in order to obtain DSM, DTM, 3D vector and orthophoto respectively. Lastly, Phase 4 is the result and analysis. The result of the study is a 3D orthophoto and 3D vector map of the study area. For accuracy assessment, calculation of RMSE and t-test are implemented.

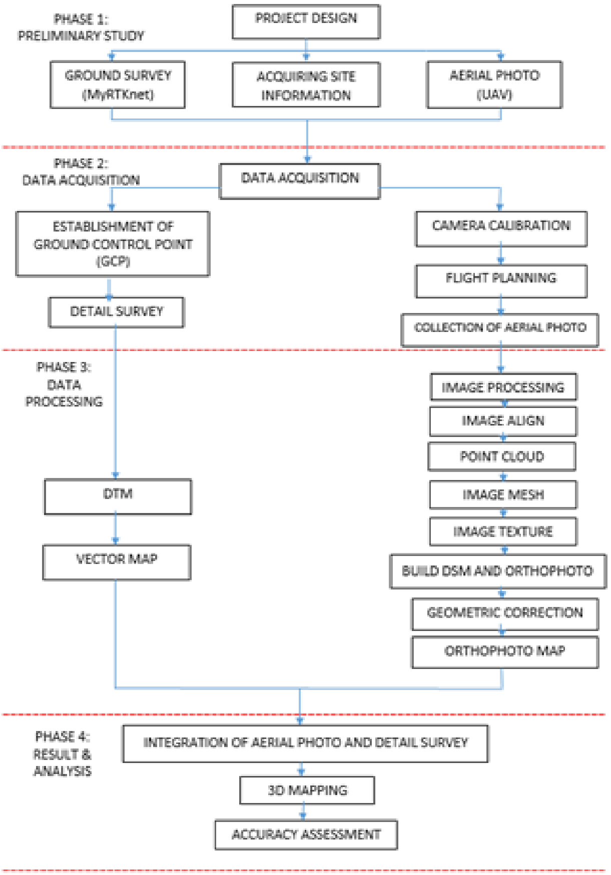


Fig. 2. Flowchart of Study.

A. Data Acquisition

The process of data collection is conducted based on the project frameworks that were collected in the preliminary study. This is to ensure the collected data met the scope of this study. Data collection involves two main parts; primary data and secondary data. For primary data, it involves the collection of aerial photo using non-matrix camera payload UAV platform. Whereas secondary data is ground survey is using GPS and total station.

The data acquisition begins with the establishment of ground control point GCP'S followed by detail survey. The establishment of GCP is done by using GPS Trimble GR-5 at the selected control station. The collection of detailed ground survey data consists several survey tools such as reflectorless total station, handheld GPS and Topcon GPS GR5 respectively in order to obtain the coordinates and elevation (x, y, z) of the feature point.

For the aerial photo, the flight planning of the UAV is conducted and installed in the laboratory by using open source software. In addition, camera calibration has also been conducted to photo distortion. UAV XR Q30 Pro is used for the whole process for collecting the photos.

B. Establishment of Ground Control Point

The establishment of GCPs is a very important stage in the photogrammetric mapping [45]. In this study, ten (10) ground control points have been established. In order to complete the observation of every ground control point, GPS static observation method is applied within 30 minutes' observation for every station. This method is suitable to establish control point at the wide area with sub-centimeter accuracy. GPS observation for control point shall be carried out radially. The concept is known to coordinate at base station and it has been used to compute carrier phase correction in observation. The data observation will be processed using Trimble Total Control software and all coordinates subsequently transformed from WGS84 to local coordinate GDM2000 RSO as shown in Table 1.

TABLE I. COORDINATES OF GCP'S

Station	**WGS84**		**GDM2000 RSO**	
1	102.323	2.312	480602.317	255768.550
2	102.323	2.311	480600.904	255718.090
3	102.320	2.309	480347.303	255457.218
4	102.320	2.309	480351.282	255509.274
5	102.318	2.310	480122.095	255595.380
6	102.318	2.310	480127.291	255585.059
7	102.319	2.311	480184.663	255663.234
8	102.319	2.311	480244.406	255653.314
9	102.319	2.312	480176.377	255818.904
10	102.319	2.312	480154.621	255767.354

C. Primary Data

Primary data involves the acquisition of aerial photo using Unmanned Aerial Vehicle (UAV). UAV refers to an aircraft that fly without an onboard human pilot. UAV can be remotely controlled from the ground station or flown autonomously based on the pre-programmed flight planning that installed before the UAV's flown. For this study, a model of UAV XR Q350 Pro that weight 3 kg is used as shown in Figure 3 and the specification of the UAV is shown in Table 2. Furthermore, UAV can maintain a flight time for 25 minutes. UAV is mounted with non-matrix camera model Canon Powers hot XS260 with focal length 4.5-90.0 mm as shown in Figure 4 and the specifications are mentioned in Table 3.

The UAV maneuvers at altitude average of 100 m above ground in this study. Each photo have 50% overlap, 65% side lap and a total of 7228 photos were processed using the photogrammetric system. Figure 5 shows a sample aerial photo. The photos then underwent stitching process in order to produce the orthophoto of the study area. In addition, the digital surface model (DSM) and a digital terrain model (DTM) were also produced. The whole process is conducted using i7 processor with 16 GB RAM.

Fig. 3. XR Q30 Pro.

TABLE II. SPECIFICATIONS OF XR Q350 PRO UAV

Criteria	**Specification**
Main Rotor Diameter	556mm
Main Rotor Blade Length	206mm
Length	289mm
Width	289mm
Height	200mm
Brushless Motor	WK-WS-28-008A
Brushless ESC	WST-15A(G/R)
Receiver	RX703
Flight Time	25 minutes

Fig. 4. Canon Powershot XS260 Camera.

TABLE III. SPECIFICATIONS OF CANON POWERSHOT XS260 CAMERA

Criteria	**Lens**
Focal Length	4.5 – 90.0 mm (35 mm equivalent: 25 – 500 mm)
Zoom	Optical 20x Zoom Plus 39x Digital Approx. 4.0x (with Digital Tele-Converter Approx. 1.5x or 2.0x and Safety Zoom[1]). Combined Approx. 80x
Maximum f/number	f/3.5 – f/6.8
Construction	12 elements in 10 groups (1 UA lens, 2 double-sided a spherical lens)
Image Stabilization	Yes (lens shift-type), 4-stop. Intelligent IS
Effective Pixels	Approx. 12.1M

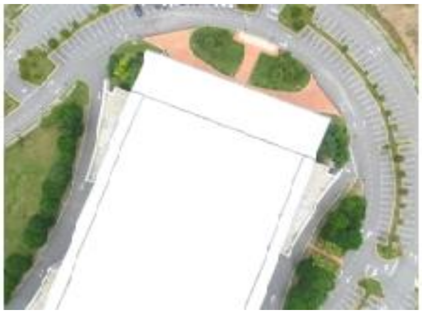


Fig. 5. Aerial Photo of the Study Area.

D. Secondary Data

Secondary data involves the data acquisition of ground survey. In this study, the RTK GPS's surveying technique is used for detailed ground survey. Real-Time Kinematic (RTK) technique is used to improve the precision of the position data derived from satellite-based positioning system. In addition, RTK techniques enable the study to obtain centimeter-level positioning which considered relevant for this study. The system used in this study is Malaysian Real-Time Kinematic GNSS Network (MyRTKnet). MyRTKnet is a system based on a network of seventy-eight (78) global navigation satellite system (GNSS) reference station continuously connected via internet protocol virtual private network (IPVPN) to the control center. At the control center, the computer processor will gather the information from all GPS receiver and creates a living database of Regional Area Connection.

The collections of detailed data were such as spot height, road junctions and other types of data utility. The detailed survey will be processed using CDS (civil design and survey) and AutoCAD. This system was employed to digitize and join the detailed survey data which is obtained using GPS devices and handheld GPS as shown in Figure 6 and Figure 7, respectively. The GPS devices used for this study is Topcon GPS GR-5 and the specification is mentioned in Table 4. Whereas, reflectorless total station shown in Figure 8 will be used to collect the detailed survey of UTeM such as building heights and building edges. The specification of the reflectorless total station as stated in Table 5. Topography map for this study is produced based on the compilation of MyRTKnet and reflectorless total station data. While 3D vector data have been generated from this ground survey.

Fig. 6. Topcon GPS GR-5.

Fig. 7. Handheld GPS.

TABLE IV. TOPCON GPS GR-5 SPECIFICATION

GNSS	
Signals Tracked	GPS: L1, L1C, L2, L2C, and L5 GLONASS: L1, L2 Galileo*: E1, E5a, E5b, AltBOC BeiDou: B1, B2 SBAS L1 C/A WAAS/MSAS/EGNOS QZSS L1 C/A, L1C, L2C
Number of Channels	226-Channel Vanguard Technology with Universal Tracking Channels capable of All-in-View tracking
Antenna Type	Integrated Fence Antenna (1) with Ground Plane
Accuracy	
Static	H: 3mm + 0.1ppm (2) V: 3.5mm + 0.4ppm (2)
RTK	H: 5mm + 0.5ppm V: 10mm + 0.8ppm

Fig. 8. Topcon Total Station FS 105.

TABLE V. TOPCON TOTAL STATION FS 105 SPECIFICATION

Specification	
Display	Dual backlit LCD (ES-107 Single Display)
Battery Operation	Up to 36 hours
Wireless Connection	Bluetooth® Class 1
Operating Temperature	-20ºC to +60ºC

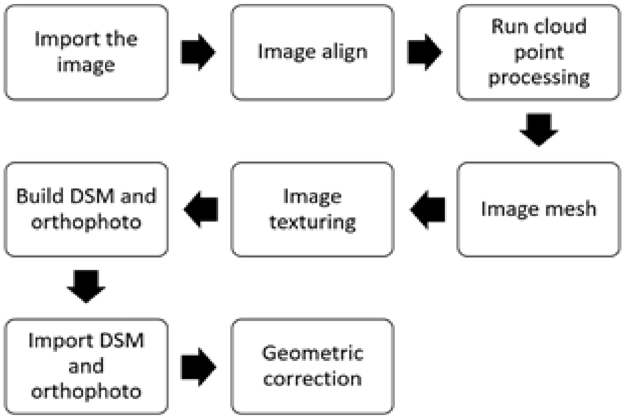


Fig. 9. Flowchart of Image Processing.

E. Data Processing

The aerial photo from UAV will undergo several processes in order to produce the orthophoto of the study area. The first step in producing orthophoto is photo alignment which shows the position and orientation of the photos. In addition, the sparse cloud that produced in this step will reveal the whole area of the study. After the photo alignment, the reconstruction of dense point cloud will be done. It enables the orthophoto that produces in higher detail. Furthermore, the photo alignment is depending on digital matching technique and space intersection using the system.

The following steps are used in photo meshing and texturing. Photo meshing is a process to turn discrete point cloud data into continuous 3D surface and it is constructed in the form of triangulated irregular network (TIN) [46]. In addition, texturing is a process of projecting the texture from the original photo to the model surface [46]. Finally, orthophoto and digital surface model (DSM) were produced. The whole process is shown in Figure 9. The produced orthophoto will undergo geometric corrections in order to transform the coordinates from WGS84 to local coordinate (GDM2000 RSO).

The ground survey data will be processed using Trimble Total Center (TTC) system. Whereby a detailed survey will be processed by using CDS (civil design & survey) and AutoCAD, respectively. The raw data that obtained from the reflectorless total station will be imported to CDS in order to generate the contour line and ground features details. After that, import the data into AutoCAD to join the details. Through this process vector map of the study area will be produced.

F. Generating 3D from Aerial Photo and Detail Survey

The generation of 3D from aerial photo involves two main components which are orthophoto and Digital Surface Model (DSM). The produced orthophoto will provide the location of the object in term of x and y positioning. Subsequently, the digital surface model (DSM) will provide the z values which is the height of an object. By using the module that existed in a GIS system, the DSM value will be referred as the base height and been applied to the orthophoto (2D) in order to produce a 3D model of an object as shown in the results' section.

The process is repeated in order to produce a 3D vector map. While compared to aerial photo the vector drawing that produces from AutoCAD contain the values of z. Thus using the module in the GIS system, a 3D vector map can be produced from 2D vector map.

V. Results

Based on the obtained aerial photo, the orthophoto and DSM for the study area can be produced. Hence, these data can be used to produce a 3D model of the study area. For the obtained featured points through ground detailed survey, it can be used to produce a 3D vector of the study area.

A. 3D Modelling and Orthophoto

Through the data processing, vector map and orthophoto of UTeM main campus are shown in Figure 10 and Figure 11, respectively. In addition, using CDS and AutoCAD system, 3D vector map can be generated as shown in Figure 12. The generated of DSM from the aerial photo will be used to provide height value for orthophoto in order to produce a 3D map based on aerial photo for UTeM main campus (refer to Figure 13). Whereas, Figure 14, Figure 15 and Figure 16 shows the images and 3D visualization of the selected buildings.

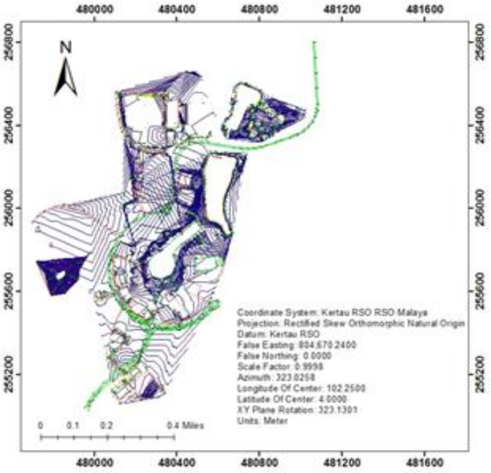

480000
480400
480800
481200
481600
256800
256400
256000
255600
255200
N
Coordinate System: Kertau RSO RSO Malaya
Projection: Rectified Skew Orthomorphic Natural Origin
Datum: Kertau RSO
False Easting: 804,670.2400
False Northing: 0.0000
Scale Factor: 0.9998
Azimuth: 323.0258
Longitude Of Center: 102.2500
Latitude Of Center: 4.0000
XY Plane Rotation: 323.1301
Units: Meter
0 0.1 0.2 0.4 Miles

Fig. 10. Vector Map of UTeM Main Campus.

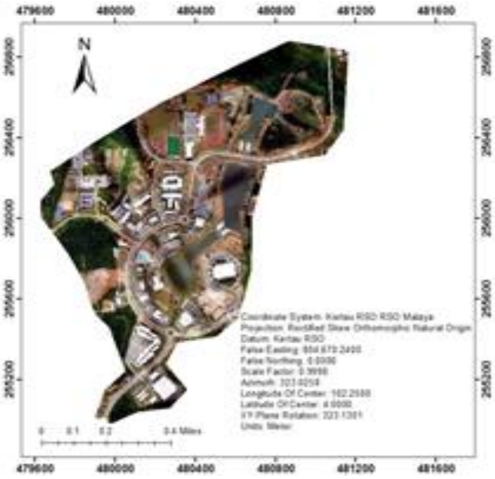

479600
480000
480400
480800
481200
481600
256800
256400
256000
255600
255200
N
Coordinate System: Kertau RSO RSO Malaya
Projection: Rectified Skew Orthomorphic Natural Origin
Datum: Kertau RSO
False Easting: 804,670.2400
False Northing: 0.0000
Scale Factor: 0.9998
Azimuth: 323.0258
Longitude Of Center: 102.2500
Latitude Of Center: 4.0000
XY Plane Rotation: 323.1301
Units: Meter
0 0.1 0.2 0.4 Miles

Fig. 11. Orthophoto of UTeM Main Campus.

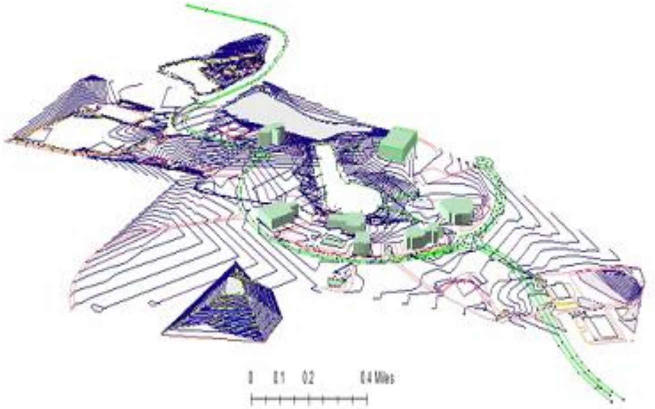


Fig. 12. 3D Vector Map of UTeM Main Campus.

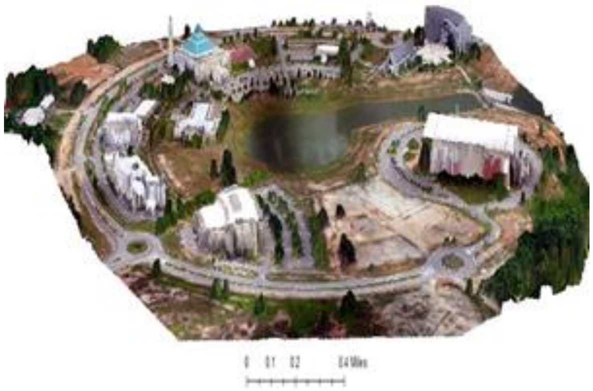
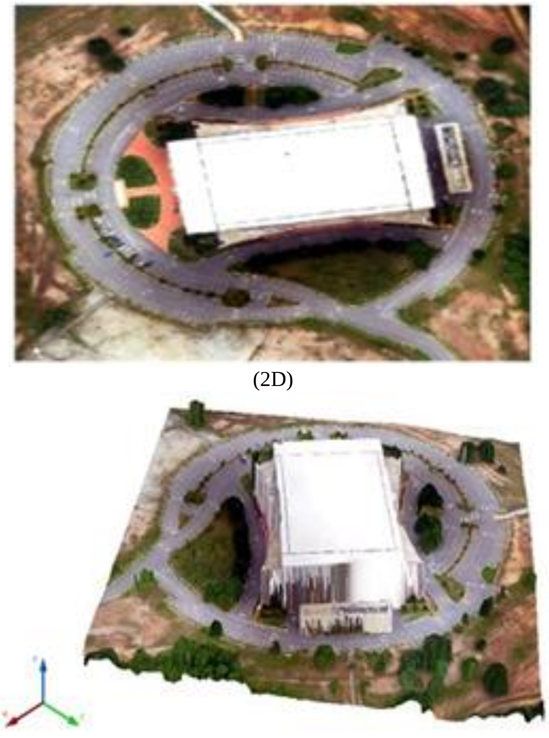


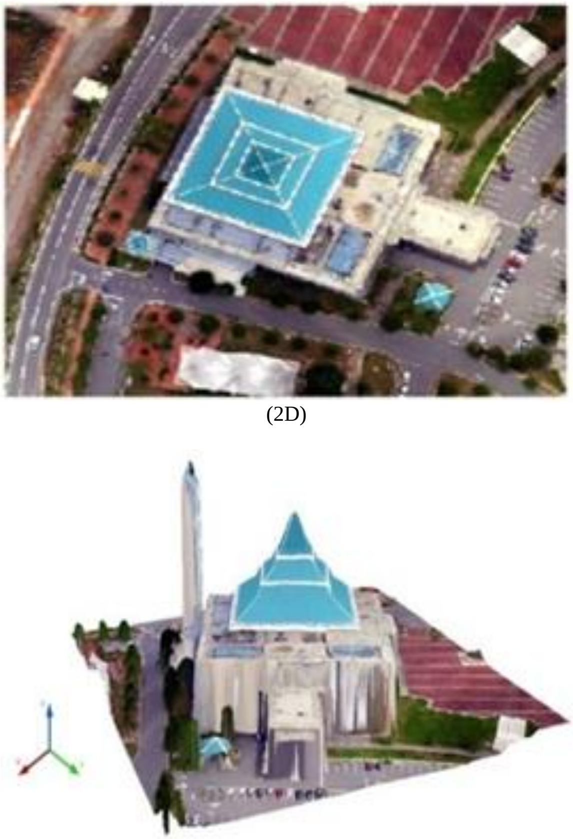
Fig. 13. 3D Visualization of the Study Area.



(2D)

(3D)

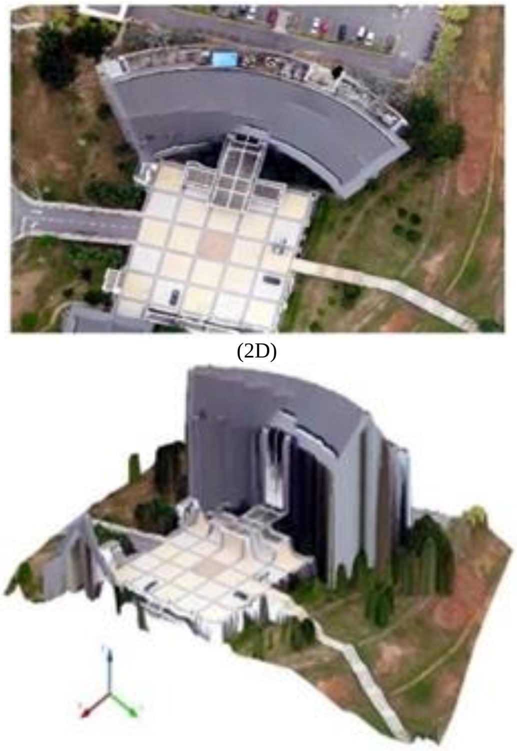
Fig. 14. 2D and 3D Map of the Main Hall.



(2D)

(3D)

Fig. 15. 2D and 3D Map for Mosque.



(2D)

(3D)

Fig. 16. 2D and 3D Map for Chancellery Hall.

VI. ANALYSIS

In this section, the quantitative analysis was carried out to show the accuracy for this study. Quantitative analysis is a technique that used to understand the behaviour of the data through mathematical or statistical modelling. Quantitative assessment is implemented by calculating the root mean square error (RMSE) from two different data sources. This study will show the quantitative assessment in term of location or positions which are (x and y) and height (z) from detailed ground survey and raster images from UAV respectively.

$$\text{RMSE}= \sqrt{\frac{\Sigma(\text{n1}-\text{n2})^2}{N}} \qquad (1)$$

Where

$n1$-$n2$= difference between two parameters

N = total number of points

TABLE VI. TWENTY-FIVE (25) SAMPLE POINTS FROM VECTOR MAP

Point	Vector(x)	Vector(y)	Vector(z)
1	480,197.70	255,668.09	57.53
2	480,221.08	255,676.68	57.53
3	480,195.39	255,635.18	57.53
4	480,210.32	255,635.77	57.53
5	480,244.16	255,653.55	57.53
6	480,117.81	255,605.85	67.08
7	480,133.18	255,575.90	67.08
8	480,168.58	255,597.98	67.08
9	480,156.69	255,621.80	67.08
10	480,334.94	255,458.03	66.66
11	480,314.16	255,482.27	66.66
12	480,392.47	255,487.70	66.66
13	480,351.39	255,534.90	66.66
14	480,608.37	255,726.83	56.98
15	480,602.09	255,769.41	56.98
16	480,611.92	255,769.64	56.98
17	480,151.53	255,546.07	56.46
18	480,169.13	255,554.52	56.46
19	480,031.48	255,570.24	58.84
20	480,010.63	255,573.74	58.84
21	480,005.62	255,589.43	58.84
22	480,021.15	255,601.94	58.84
23	480,454.55	255,997.24	69.87
24	480,399.01	256,037.18	69.87
25	480,396.22	256,031.12	69.87

TABLE VII. TWENTY-FIVE (25) SAMPLE POINTS FROM ORTHOPHOTO

Point	Orthophoto(x)	Orthophoto(y)	Orthophoto(z)
1	480,197.37	255,668.20	55.05
2	480,221.33	255,676.90	55.05
3	480,195.26	255,634.92	55.05
4	480,210.19	255,635.65	55.05
5	480,244.34	255,653.46	55.05
6	480,117.83	255,605.90	67.81
7	480,133.15	255,575.81	67.81
8	480,168.69	255,597.84	67.81
9	480,156.71	255,621.92	67.81
10	480,334.96	255,458.21	66.99
11	480,314.27	255,482.29	66.99
12	480,392.53	255,487.95	66.99
13	480,351.48	255,535.02	66.99
14	480,608.41	255,726.50	55.07
15	480,602.29	255,768.68	55.07
16	480,611.88	255,769.96	55.07
17	480,151.56	255,546.07	56.52
18	480,169.25	255,554.55	56.52
19	480,031.41	255,570.13	58.92
20	480,010.62	255,573.81	58.92
21	480,005.55	255,589.44	58.92
22	480,021.16	255,601.98	58.92
23	480,454.53	255,997.49	70.34
24	480,398.34	256,037.14	70.34
25	480,395.15	256,030.97	70.34

Twenty-five (25) points that randomly selected distributed around the study area were used as the sample point for RMSE calculation (refer to Table 6 and Table 7). By using equation (1) this study managed to obtain RMSE value for coordinates (x and y) and height (z) which are RMSE(x) = 0.279, RMSE(y) = 0.215, and RMSE(z) = 1.341, respectively. This study also conducts pair sample t-test in order to produce a concrete result. Paired sample t-test is a statistical procedure that used to determine whether the mean differences for two set of selected observation is zero. The parameters for paired sample t-test are coordinates (x and y) and height (z) that obtain from 3D raster (refer to Figure 13) and 3D vector (refer to Figure 12). Table 8 and Table 9, respectively show the results for the pair sample correlation and pair sample test respectively.

TABLE VIII. PAIR SAMPLE CORRELATIONS

	N	Correlation
X_vector &X_raster	25	1.000
Y_vector &Y_raster	25	1.000
Z_vector & Z_raster	25	0.990

TABLE IX. PAIR SAMPLE TEST

	Mean	95% Confidence Interval of the Difference		t
		Lower	Upper	
X_vector - X_raster	0.053	-0.061	0.168	0.961
Y_vector - Y_raster	0.010	-0.079	0.101	0.250
Z_vector - Z_raster	0.481	-0.045	1.008	1.885

Based on Table 8 above, the result shows that the parameter obtained from the 3D vector and 3D raster for coordinates (x and y) and height (z) were positively correlated which are x = 1.000, y = 1.000 and z = 0.990 respectively. While Table 9 illustrated the pair sample test, indicates x-coordinates for the 3D vector are 0.053, higher than 3D raster with 95% confidence interval [-0.061, 0.168]. Meanwhile, the test shows y-coordinates for 3D vector is 0.010 higher than 3D raster with 95% confidence interval [-0.079, 0.101]. Lastly, the height (z) from 3D vector was 0.4815 higher than 3D raster with 95% confidence interval [-0.045, 1.008].

VII. Conclusion

This paper has presented the use of the unmanned aerial vehicle (UAV) in generating 3D mapping for UTeM, Malacca. This study shows that aerial UAV photos can be used to generate 3D models of features within the selected study area. Even though the generated 3D models do not look alike in the real world, but it manages to outline the shape of the features, for example, the chancellor hall and mosque with good planimetric accuracy rates. However, this study concluded that the accuracy can be improved especially for the height (z) by changing the aerial flight type to oblique photograph instead of vertical photograph or nadir as used in this study (refer to Figure 17).

The implementation of the UAV platform in this study can be described as a low-cost method or approach using close-range photogrammetry in generating large-scale of 3D map compared to other type of conventional methods. Thus, the integration of UAV platform aerial photo and ground detailed survey can be used to produce a large-scale of 3D mapping.

In addition, the result of this study is well supported using accuracy assessments which are RMSE and paired sample t-test. The obtained RMSE values for x, y and z from 25 distributed points are $RMSE_X$= 0.279, $RMSE_Y$= 0.215 and $RMSE_Z$= 1.341 respectively. For paired sample t-test, the coordinates (x and y) and height (z) from both data were positively correlated.

Furthermore, the study also managed to fulfill the criteria to produce a 3D city model. According to cityGML (Architectural models) there are five (5) types of Level of Detail that used to facilitate efficient visualization and data analysis which are Level of Detail 0 (LOD0), Level of Detail 1 (LOD1), Level of Detail 2 (LOD2), Level of Detail 3 (LOD3) and Level of Detail 4 (LOD4). The positional and height accuracy for each Level of Detail is LOD0 (less than LOD1), LOD1 (5m or less), LOD2 (2m or better), LOD3 (0.5m) and LOD4 (0.2m or less) [47].

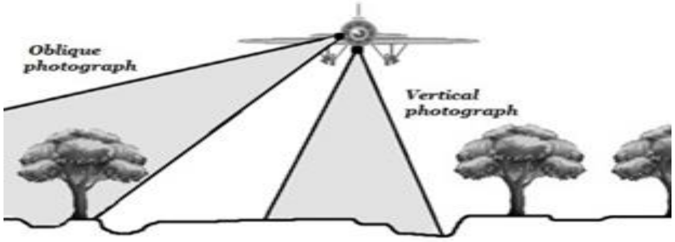


Fig. 17. Type of Aerial Photography.

Referring to cityGML (architectural models), this study has to fulfill the criteria of LOD2 successfully in order to show the positional and height accuracy at 2m or better. Subsequently, the building has differentiated roof structures and thematically differentiated boundary surfaces (refer to Figure 13) which is the criteria of LOD2 [47]. Last but not least, this study managed to show that by using the integration of UAV and detailed surveying method, this study able to produce a 3D campus map for UTeM, Malacca.

Acknowledgment

The authors would like to express their profound gratitude to Faculty of Built Environment and Surveying (FABU), Universiti Teknologi Malaysia (UTM) for all the support that been provided. The authors also would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for data collection (UAV aerial photo) and technical assistance in this study. Last but not least, the authors would like to thank UTM for providing the financial support through FRGS VOT 4F336, GUP-Tier 1 grant VOT 14H44, GUP-Tier 1 grant VOT 20H01, GUP Tier 2 grant VOT 12J47 and UTM-TDR 6.4 (T2).

References

[1] Billen, R.; Cutting-Decelle, A.F.; Marina, O.; de Almeida, J.P.; Matteo, C.; Falquet, G.; Leduc, T.; Métral, C.; Moreau, G.; Perret, J.; et al. 3D City Models and urban information: Current issues and perspectives. In 3D City Models and Urban Information: Current Issues and Perspectives—European COST Action TU0801; EDP Sciences: Les Ulis, France, 2014; pp. 1–118.

[2] Zhu, Q.; Hu, M.; Zhang, Y.; Du, Z. Research and practice in three-dimensional city modeling. Geo-Spat. Inf. Sci. 2009, 12, 18–24.

[3] Döllner, J.; Baumann, K.; Buchholz, H. Virtual 3D City Models as Foundation of Complex Urban Information Spaces. In Proceedings of the 11th International Conference on Urban Planning and Spatial Development in the Information Society, Vienna, Austria, 13–16 February 2006.

[4] Lancelle, M.; Fellner, D.W. Current issues on 3D city models. In Proceedings of the Proceedings of the 25th International Conference in Image and Vision Computing, Queenstown, New Zealand, 8–9 November 2010; pp. 363–369.

[5] Suveg, I.; Vosselman, G. Reconstruction of 3D building models from aerial images and maps. ISPRS J. Photogramm. Remote Sens. 2004, 58, 202–224.

[6] Haala, N.; Kada, M. An update on automatic 3D building reconstruction. ISPRS J. Photogramm. Remote Sens. 2010, 65, 570–580.

[7] Tomljenovic, I.; Höfle, B.; Tiede, D.; Blaschke, T. Building extraction from airborne laser scanning data: An analysis of the state of the art. Remote Sens. 2015, 7, 3826–3862.

[8] Blaschke, T. Object-based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 2010, 65, 2–16.

[9] Ledoux, H.; Meijers, M. Topologically consistent 3D city models obtained by extrusion. Int. J. Geogr. Inf. Sci. 2011, 25, 557–574.

[10] Arroyo Ohori, K.; Ledoux, H.; Stoter, J. A dimension-independent extrusion algorithm using generalised maps. Int. J. Geogr. Inf. Sci. 2015, 29, 1166–1186.

[11] Shahzad, M.; Zhu, X.X. Robust reconstruction of building facades for large areas using spaceborneTomoSAR point clouds. IEEE Trans. Geosci. Remote Sens. 2015, 53, 752–769.

[12] Zhu, X.X.; Shahzad, M. Facade reconstruction using multiviewspaceborneTomoSAR point clouds. IEEE Trans. Geosci. Remote Sens. 2014, 52, 3541–3552.

[13] Schmitt, M. Reconstruction of urban surface models from multi-aspect and multi-baseline interferometric SAR. Ph.D. Thesis, TechnischeUniversitätMünchen, München, Germany, 2014.

[14] Still, U.; Soergel, U.; Thoennessen, U. Potential, and limits of InSAR data for building reconstruction in built-up areas. ISPRS J. Photogramm. Remote Sens. 2003, 58, 113–123.

[15] Thiele, A.; Wegner, J.D.; Soergel, U. Building reconstruction from multi-aspect InSAR data. In Remote Sensing and Digital Image Processing; Soergel, U., Ed.; Springer: Dordrecht, The Netherlands, 2010; pp. 187–214.

[16] Donkers, S.; Ledoux, H.; Zhao, J.; Stoter, J. Automatic conversion of IFC datasets to geometrically and semantically correct CityGML LOD3 buildings. Trans. GIS 2015, doi:10.1111/tgis.12162.

[17] Yin, X.; Wonka, P.; Razdan, A. Generating 3D building models from architectural drawings: A survey. IEEE Comput. Graph. Appl. 2009, 29, 20–30.

[18] Lewis, R.; Séquin, C. Generation of 3D building models from 2D architectural plans. Comput.-Aided Des. 1998, 30, 765–779.

[19] Sirmacek, B.; Lindenbergh, R. Accuracy assessment of building point clouds automatically generated from iphone images. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2014, XL-5, 547–552.

[20] Rosser, J.; Morley, J.; Smith, G. Modelling of building interiors with mobile phone sensor data. ISPRS Int. J. Geo-Inf. 2015, 4, 989–1012.

[21] Besuievsky, G.; Patow, G. Recent advances on LoD for procedural urban models. In Proceedings of the 2014 Workshop on Processing Large Geospatial Data, Cardiff, UK, 8 July 2014.

[22] Tsiliakou, E.; Labropoulos, T.; Dimopoulou, E. Procedural modeling in 3D GIS environment. Int. J. 3-D Inf. Model. 2014, 3, 17–34.

[23] Müller, P.; Wonka, P.; Haegler, S.; Ulmer, A.; van Gool, L. Procedural modeling of buildings. ACM Trans. Graph. 2006, 25, 614–623.

[24] Smelik, R.M.; Tutenel, T.; Bidarra, R.; Benes, B. A Survey on procedural modelling for virtual worlds. Comput. Graph. Forum 2014, 33, 31–50.

[25] Biljecki, F.; Ledoux, H.; Stoter, J. Error propagation in the computation of volumes in 3D city models with the Monte Carlo method. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2014, II-2, 31–39.

[26] Martinovi´c, A. Inverse Procedural Modeling of Buildings. Ph.D. Thesis, KU Leuven, Leuven, Belgium, 2015.

[27] Over, M.; Schilling, A.; Neubauer, S.; Zipf, A. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. Comput. Environ. Urban Syst. 2010, 34, 496–507.

[28] Goetz, M. Towards generating highly detailed 3D CityGML models from OpenStreetMap. Int. J. Geogr. Inf. Sci. 2013, 27, 845–865.

[29] Goetz, M.; Zipf, A. Towards defining a framework for the automatic derivation of 3D CityGML models from Volunteered Geographic Information. Int. J. 3-D Inf. Model. 2012, 1, 1–16.

[30] Rebelo C., Rodrigues A.M., Tenedório J.A., Goncalves J.A., Marnoto J. (2015) Building 3D City Models: Testing and Comparing Laser Scanning and Low-Cost UAV Data Using FOSS Technologies. In: Gervasi O. et al. (eds) Computational Science and Its Applications -- ICCSA 2015. ICCSA 2015. Lecture Notes in Computer Science, vol 9157.

[31] Eisenbeiß, H., 2009. UAV Photogrammetry, Dissertation Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland.

[32] Remondino, F., Barazzetti, L., Nex, F., Scaioni, M., Sarazzi, D., 2011, UAV photogrammetry for mapping and 3Dmodeling– current status and future perspectives, in: H. Eisenbeiss, M. Kunz, H. Ingensand (Eds.), Proceedings of the International Conference on Unmanned Aerial Vehicle in Geomatics (UAV-g) 2011, Zurich, Switzerland.

[33] Zongjian, L.I.N., 2008, UAV for mapping—low altitude photogrammetric survey, International Archives of Photogrammetry and Remote Sensing, Beijing, China.

[34] Ouédraogo, M.M., Degré, A., Debouche, C., Lisein, J., 2014. The evaluation of unmanned aerial system-based photogrammetry and terrestrial laser scanning to generate DEMs of agricultural watersheds. Geomorphology, 214, pp. 339–355.

[35] H. Eisenbeiß, UAV Photogrammetry, Dissertation Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland, 2009.

[36] F. Remondino, L. Barazzetti, F. Nex, M. Scaioni, D. Sarazzi, UAV photogrammetry for mapping and 3Dmodeling–current status and future perspectives, in: H. Eisenbeiss, M. Kunz, H. Ingensand (Eds.), Proceedings of the International Conference on Unmanned Aerial Vehicle in Geomatics (UAV-g) 2011, Zurich, Switzerland, September 2011.

[37] L.I.N. Zongjian, UAV for mapping—low altitude photogrammetric survey, International Archives of Photogrammetry and Remote Sensing, Beijing, China, 2008.

[38] B.P. Hudzietz, S. Saripalli, An experimental evaluation of 3d terrain mapping with an autonomous helicopter, in: H. Eisenbeiss,M. Kunz, H. Ingensand (Eds.), Proceedings of the International Conference on Unmanned Aerial Vehicle in Geomatics (UAV-g) 2011, Zurich, Switzerland, September 2011.

[39] D. Bulatov, P. Solbrig, H. Gross, P. Wernerus, E. Repasi, C. Heipke, Context-based urban terrain reconstruction from UAV-videos for geoinformation applications, in: H. Eisenbeiss,M. Kunz, H. Ingensand (Eds.), Proceedings of the International Conference on Unmanned Aerial Vehicle in Geomatics (UAV-g) 2011, Zurich, Switzerland, September 2011.

[40] F. Neitzel, J. Klonowski, Mobile 3D mapping with a low-cost UAV system, Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 38 (2011) 1–6.

[41] Zhiguang Ding, Xinsheng yan, Huiguang Chen. Study on the Rapid 3D Modeling of Jiangmen City Based on Tilting Skill of Unmanned Aerial Vehicle [J]. Urban Survey, 2016 (4): 72-78.

[42] B.P. Hudzietz, S. Saripalli, An experimental evaluation of 3d terrain mapping with an autonomous helicopter, in: H. Eisenbeiss,M. Kunz, H. Ingensand (Eds.), Proceedings of the International Conference on Unmanned Aerial Vehicle in Geomatics (UAV-g) 2011, Zurich, Switzerland, September 2011.

[43] Eisenbeiss H. 2011. The potential of unmanned aerial vehicles for mapping. In *Photogrammetric Week '11*, Fritsch D (ed.). Wichmann: Berlin/Offenbach; 135–145.

[44] Gatziolis, D., J. F. Lienard, A. Vogs, and N. S. Strigul. 2015. "3D Tree Dimensionality Assessment Using Photogrammetry and Small Unmanned Aerial Vehicles." PLoS One 10 (9): e0137765.

[45] Tahar, K.2015. Height accuracy based on different RTK GPS method for Ultra-Light Aircraft Images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43, 1-6.

[46] Hidayat H and Cahyono A B 2016 Combined Aerial and Terrestrial Images For Complete 3D Documentation Of Singosari Temple Based On Structure From Motion Algorithm *IOP Conference Series: Earth and Environmental Science* 47.

[47] Biljecki, I. F. 2013. "The Concept of Level of Detail in 3D City Models." PhD Research Proposal GISt Report No. 62.

Triangle Shape Feature based on Selected Centroid for Arabic Subword Handwriting

Nur Atikah Arbain[1], Mohd Sanusi Azmi[2], Azah Kamilah Muda[3], Amirul Ramzani Radzid[4]
Faculty of Information and Communication Technology,
University Teknikal Malaysia Melaka, 76100 Durian
Tunggal, Melaka, Malaysia

Azrina Tahir[5]
Department of Information & Communication Technology,
Politeknik Ungku Omar,
Jalan Raja Musa Mahadi,
31400 Ipoh, Perak, Malaysia

***Abstract*—Features are normally modelled based on color, texture and shape. However, some features may have different constraints based on types, styles and pattern of an image. The Arabic subword handwriting, for example, cannot be recognized by color and not suitable to be characterized based on texture. Therefore, features based on shape are suitable to be used for recognizing Arabic subword handwriting since each of the character has various characteristics such as diacritics, thinning and strokes. These characteristics can contribute to particular a shape that is unique and can represent Arabic subword handwriting. Currently, geometry shape such as triangle has been adopted to extract useful features based on triangle properties without implicating any triangle form. In order to increase classification accuracy, these properties have been categorized into several zones where the number of features produced is directly proportional to the number of zones. Nevertheless, shape representation does not implicate any triangle properties such as ratio of side, angle and gradient. By using shape representation, it helps in reducing the number of features. Thus, this paper presents feature based on triangle shape that can represent the identity of Arabic subword handwriting. The method based on triangle shape identifies three main coordinates of triangle formed based on selected centroids. The AHDB dataset is used as a testing data. The Support Vector Machine (SVM) and Random Forest (RF), respectively were used to measure the accuracy of the proposed method using triangle shape as a feature. The accuracy results have shown better outcome with 77.65% (SVM) and 76.43% (RF), which prove the feature based on triangle shape is applicable for Arabic subword handwriting recognition.**

Keywords—Arabic subword; feature extraction; random forest; support vector machine; triangle geometry

I. Introduction

Subword handwriting is one of the popular handwriting studies that have been actively explored for many years due to challenges in identifying the styles, patterns, and signatures of subword handwriting. The recognition of the handwritten words is based on the recognition of segmented characters from subword. Images of handwritten documents will be processed from imaging from pages to lines, and words to subwords.

Due to the challenging task in subword handwriting, there have been intensive responses and encouragement by numerous researchers to develop or improvise the existing recognition methods and systems. Based on [1], work in Arabic character recognition is limited. However, Arabic handwritten character recognition systems have achieved much improvement over the years. Since many languages such as Farsi, Curds, and Urdu used Arabic characters in their writing, it makes tasks more challenging due to different words used, strength and sequential order of the writing.

Over the past decades, a lot of handwritten subwords databases [1]–[3], which contain images were developed. Image processing is required to process the images, for example, to convert the image into binary pixels. Image processing is one of the vital elements that are widely used in a research area within engineering and computer science disciplines. Arabic handwritten documents currently exist in a big number of resources in physical and web form, providing a challenge for the word recognition process. The features extraction process will play an important role before the classification process. The problem with segmentation to the single characters is that the characters may be overlapped and some of the characters share the same shape, for example, characters: "ب" ,"ت" and "ث".

The features for recognizing Arabic subword have been introduced, which led to in-depth studies due to variation of style, pattern, characteristic and type of Arabic subword handwriting. Thus, the generated features must have hallmarks that can differentiate them from another subword. Two groups namely analytical and holistic are used as a recognition method in handwritten text. In the analytical group, a word segmented into components such as character or subword, a feature is extracted from each other, and a general vector is obtained from each word. Besides, the character segmentation is needed, and the errors may occur in the recognition step. In the holistic group, a feature vector is extracted from the whole word image without any need to image segmentation.

In this study, a holistic group is applied in producing novel features for Arabic subword handwriting. The feature based on triangle shape is proposed using three main coordinates of triangle formed based on selected centroids. This paper is organized as follows. In Section II, the related work is discussed. The proposed method is discussed in Section III. Next, the experiment and evaluation of study is presented in Section IV. Finally, Section V concludes this paper.

II. Related Works

A huge number of pre-modern texts have been scanned as subword images to remain against aging. Nevertheless, the ability of researchers to handle with the images were limited due to the difficulty to handle certain tasks such as query search [4]. According to [4], it is important to provide researcher with algorithms for automatic transliteration and transcription of scanned images, which would extract the textual content of the image and reproduce into an editable text file.

With the advanced technology, many approaches and methods have been introduced for Arabic handwritten text recognition. The feature learning framework had been proposed by [5] using a Bag-of-Feature (BoF) paradigm for Arabic handwritten text recognition. Besides that, scale invariant feature transform (SIFT) descriptor was used by [6] to represent the object in detail to reduce the computation cost. However, [6] has stated that the complexity of testing image cannot be too high when performing object recognition and image retrieval on big data. This is because a vector with 128 dimensions represents one feature point and an image will have several feature points. Thus, more time is needed to compare the feature points individually.

Therefore, it is important in selecting suitable features applicable to represent the image. The structural approach is applied in generating triangle shape feature based on selected centroid. Then, the holistic approach is applied where the whole subword image is used without segmenting each character from subword image.

The holistic approach has been applied by [7] in producing features for AHDB subword images using Discrete Cosine Transform (DCT) and histogram of oriented gradient (HOG). The features are produced based on whole subword image without any word segmentation. An array of the best 50 DCT coefficients and 324 of HOG features are produced as the parameters of the features for subword images. Besides, a study in [5] also has used holistic approach in producing features based on Bag-of-Feature (BoF) paradigm. The BoF framework is exploited by [5] as to learn robust feature representations for Arabic handwriting recognition. Several approaches have been implemented as there are few stages in the framework that will use different approaches. The Harris detector and dense sampling have been applied for selecting representative image regions. Then, Principle Component analysis (PCA) is applied to reduce Scale-Invariant Feature Transform (SIFT) descriptors to 64-D vectors. In a study by [8], a holistic group approach also has been applied in generating novel features for recognition of Persian/Arabic handwritten words. The generated feature is proposed based on a geometric attribute of components forming the word. The number, angle, location and size of a line are the parameters that represent the features in [8].

The geometry features have been adopted in object recognition, which is especially used for identifying the style and pattern of writing, font, authors, and number of authors, place of writing and originality of the documents. Apart from that, these features also have been extensively used for recognizing the type of writing and calligraphy in existing documents especially for ancient manuscripts [9]. The geometry features can be produced based on geometry shapes such as polygons including triangles, squares and pentagons. These polygons have respective properties that can be used in object recognition.

Most of the properties, for example, triangle properties have been used by researchers to produce proposed features for image classification [9], [10]. The properties are extracted after the polygon is formed. The geometry method also has been broadly used in various domains such as face recognition [11]–[13], fingerprint recognition [14]–[16], vehicle detection [17], intrusion [18] and digit recognition [9], [10]. Each of the domain has a special form that uses an indicator to determine the corner points of the formed geometry shape.

In face recognition, eyes and nose are face elements that are used as indicator to determine the points on the face. The minutiae, ridges and valleys were used as indicator in fingerprint recognition. In vehicle detection, flat road assumption has been used as indicator to search for vehicles that are located on the ground. Besides that, geometry method was also used in recognizing digit recognition and calligraphy [9], [10], [19]. A local foreground image was applied to construct triangle points based on the size of image. The author of [9] proposes new features based on triangle properties. The triangle is formed based on three triangle points of corners A, B and C. The determination of the three triangle points of corners plays a big role in triangle formation. Any fault in determining the exact coordinates of triangle points can affect the triangle formation. The midpoint of triangle is important to determine the position of triangle's point of A and B.

However, the current algorithm to extract features from face and fingerprint recognition respectively cannot be implemented in recognizing subword images. The limitation of elements used such as eyes, mouth, nasal tip, ear hole and corner of mouth in face recognition as well as minutiae in fingerprint recognition cannot be applied due to the aforementioned non-elements that exists in subword images. Thus, elements from both face and fingerprint recognition cannot be used as new feature parameter based on triangle geometry for subword images. Nevertheless, the current algorithm using triangle geometry in digit recognition is possible to be applied on subword image. However, there are constraints where every feature must be produced for all the 33 zones, which eventually lead to the increasing number of features into 297. The algorithm has increased training time in feature extraction process concomitantly with big data image used. Thus, a research on feature based on triangle shape is needed to be extended in order to facilitate subword image.

III. Proposed Method

A. Pre-processing

Before proposed features are produced, binarization process is performed in pre-processing stage for selecting adequate threshold of grey level for extracting objects from image background. Thus, Otsu thresholding method [20] is applied to convert subword image into binary form. The binarization process will transform image into binary form where ‘0’ represents foreground of image while ‘1’ represents background of image as shown in Fig. 1.

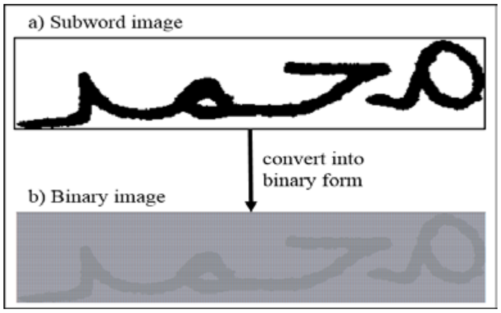


Fig. 1. Subword Image is Converted into Binary Form.

B. Feature Extraction

1) Zoning method: In this stage, the zoning method is applied to divided image into several zones, which contains useful information that can be extracted as the features. The zoning method is known as one of handwriting recognition method where handwriting image will be divided into several zones that provide regional information according to feature needs. There are four types of zoning method applied namely Cartesian plane zone, horizontal zone, vertical zone and 45-degree zone. These zoning methods also have been used in digit recognition [10], [19]. TABLE I shows the summary of zoning method information while Fig. 2 illustrates the image output from Cartesian plane zone method. Based on Fig. 2, binary image is divided into five zones including main image using Cartesian plane zone method. The binary image is measured based on height and width of the image. The height and width of binary image is obtained based on the number of binary pixels including ‘0’ and ‘1’.

TABLE I. SUMMARY OF ZONING METHOD

Zoning method	Number of zones
Cartesian Plane Zone	5 including main image
Horizontal Zone	6
Vertical Zone	14
45-Degree Zone	8
Total	**33**

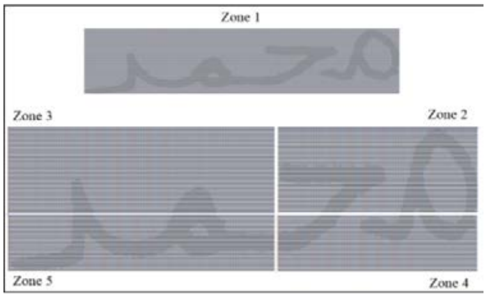


Fig. 2. Output Image after using Cartesian Plane Zone Method.

2) Geometry Method: After applying zoning method, the features can be extracted from each of zones using geometry method. In this study, triangle geometry method has been applied to generate features where the triangle shape is formed inside divided zones. There are 33 triangle shapes formed based on a total number of zones for all types of zoning method (refer to TABLE I). The features are generated based on triangle shape where the three main coordinates of triangle are formed based on selected centroids. There are six types of possible centroids that form the triangle shape as shown in TABLE III. The algorithm to obtain three main triangle coordinates is shown in Fig. 3.

```
Input: binary image of zone
Output: triangle shape points (A, B, C)
Begin
    •   Read image I from dataset
    •   $N$ ← total number of pixels at $x$-axis
    •   Get point $C$ (centroid)
    •   h ← centroid height of zone,
        w ← centroid width of zone
    •   Get point $A$.
        Find $Ax = Cx\ until\ Ax <= N - 1$
    •   Get point $B$. Find $Bx = 0\ until\ Bx <= Cx$

End
```

Fig. 3. Algorithm for Triangle Shape Coordinates.

After identifying three main triangle coordinates based on selected centroids, the coordinates are used to extract the features based on triangle shape. The number of features based on triangle shape produced 99 features (3 features × 33 zones). The description of triangle shape features based on three main triangle coordinates from selected centroids is shown in TABLE II.

TABLE II. DESCRIPTION OF TRIANGLE SHAPE FEATURES

No	Triangle shape features	Formula
1	Length of side a	$a = \sqrt{b^2 + c^2 - 2bc.cosA^{\circ}}$
2	Length of side b	$b = \sqrt{a^2 + c^2 - 2ac.cosB^{\circ}}$
3	Length of side c	$c = \sqrt{a^2 + b^2 - 2ab.cosC^{\circ}}$

I. EXPERIMENT AND EVALUATION

In this study, Arabic subword handwriting from AHDB database is used. This dataset contains more than 2000 images of Arabic words and texts written by a hundred different writers where 70% data is used as training data while 30% is used as testing data. As to evaluate the data, Support Vector Machine (SVM) and Random Forest (RF) are applied to measure the data based on accuracy. As known, the SVM is one of most popular approach that has been used in measuring classification accuracy for handwriting recognition. Thus, the libSVM is required to gain the highest cross-validation (CV) accuracy for each of the SVM parameter. The Gaussian kernel is applied to search the best grid point of cost and gamma with highest cross-validation. Then, the best value of cost and gamma are used to train the dataset using SVM. With the best value of cost and gamma, a good accuracy is achieved accordingly to the dataset nature and characteristic. The cost and gamma value for proposed method in [9] and our proposed method respectively is shown in TABLE IV.

TABLE III. DESCRIPTION OF TRIANGLE SHAPE BASED ON SELECTED CENTROID

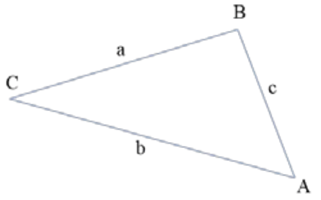
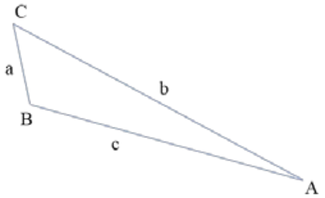
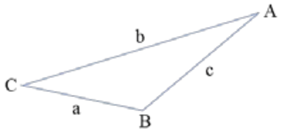
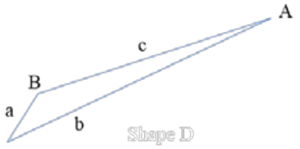
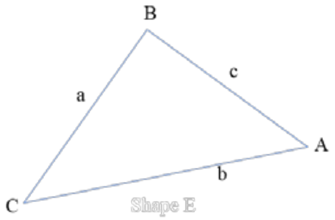
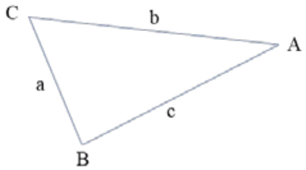
Shape types	**Rules for centroid**	**Triangle output**
A	$\boldsymbol{yA \geq yC \geq yB}$	
B	$\boldsymbol{yA \geq yB \geq yC}$	
C	$\boldsymbol{yA \leq yC \leq yA}$	
D	$\boldsymbol{yA \leq yB \leq yC}$	
E	$\boldsymbol{yA \geq yB \leq yC}$	
F	$\boldsymbol{yA \geq yC \leq yB}$	

TABLE IV. COST AND GAMMA RESULTS USING LIBSVM FUNCTION

Proposed Method	**Cost (c)**	**Gamma (γ)**
M. S. Azmi (2013) [9]	32.0	0.001953125
Our proposed method	32.0	0.03125

TABLE V. CLASSIFICATION ACCURACY RESULTS BASED SVM

Proposed Method	**Number of features**	**Accuracy (%)**
M. S. Azmi (2013) [9]	297	76.122
Our proposed method	99	77.653

The results of accuracy based on SVM classifier are compared between prior method [9] and the proposed method. Based on TABLE V, the accuracy result for the proposed method has shown better outcome by obtaining 77.653% compared to proposed method by proposed method of [9] which obtained only 76.122%. The results based on SVM showed that the proposed method has achieved target to apply minimum number of features by using triangle shape feature. Number of features is possible to be reduce from 297 to 99 by using different approaches of triangle shape features types. Furthermore, triangle shape feature can differentiate triangle shape from another triangle shape types.

Besides that, the accuracy results are also compared using other classifier based on different features used on AHDB dataset. Based on TABLE VI, the accuracy result based on random forest has shown good result for our proposed method by increasing about 7% compared to the proposed method by prior method [21]. It has shown that the proposed features using triangle shape is efficient and applicable to be used in Arabic handwritten text recognition. However, the handwriting styles, pattern and types may influence in producing the features, which made recognizing the Arabic handwriting text more challenging.

TABLE VI. CLASSIFICATION ACCURACY RESULTS BASED RF

Proposed Method	Features	Accuracy (%)
J. Salem (2017) [21]	MOMENTS	68.750
Our proposed method	Triangle shape	76.417

II. CONCLUSION

This paper presents a feature based on triangle shape that formed three main coordinates using selected centroids. The proposed feature based on triangle shape has been proven applicable to be used as a feature for recognizing Arabic subword handwriting. The results based on SVM and RF have shown good result for the proposed method compared to prior methods. The further research can be extended where other geometry shapes can be applied as a feature.

ACKNOWLEDGMENT

The authors would like to express their appreciation to Universiti Teknikal Malaysia Melaka for providing a scholarship of UTeM Zamalah Scheme. Besides, thank you to Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

[1] S. Al-Ma'adeed, D. Elliman, and C. Higgins, “A database for Arabic handwritten text recognition research,” Int. Arab J. Inf. Technol., vol. 1, no. 1, pp. 117–121, 2004.

[2] J. J. Hull, “A Database for Handwritten Text Recognition Research,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 5, pp. 550–554, 1994.

[3] H. C. Fernando, N. D. Kodikara, and S. Hewavitharana, “A database for handwriting recognition research in Sinhala language,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2003–Janua, no. Icdar, pp. 1262–1264, 2003.

[4] Y. Chherawala, R. Wisnovsky, and M. Cheriet, “TSV-LR: topological signature vector-based lexicon reduction for fast recognition of pre-modern Arabic subwords,” Proc. 2011 Work. Hist. Doc. Imaging Process., pp. 6–13, 2011.

[5] M. O. Assayony and S. A. Mahmoud, “An Enhanced Bag-of-Features Framework for Arabic Handwritten Sub-words and Digits Recognition,” J. Pattern Recognit. Intell. Syst., vol. 4, no. 1, pp. 27–38, 2016.

[6] C. Wu, C. Te Chiu, and Y. S. Hsu, “Object recognition using bag of words with kernels for big data,” Dig. Tech. Pap. - IEEE Int. Conf. Consum. Electron., pp. 89–90, 2014.

[7] M. S. Kadhm and A. K. A. Hassan, “Arabic handwriting text recognition based on efficient segmentation, DCT and HOG features,” Int. J. Multimed. Ubiquitous Eng., vol. 11, no. 10, pp. 83–92, 2016.

[8] R. Tavoli, M. Keyvanpour, and S. Mozaffari, “Statistical geometric components of straight lines (SGCSL) feature extraction method for offline Arabic/Persian handwritten words recognition,” IET Image Process., vol. 12, no. 9, pp. 1606–1616, 2018.

[9] M. S. Azmi, “Fitur Baharu Dari Kombinasi Geometri Segitiga Dan Pengezonan Untuk Paleografi Jawi Digital,” Doctoral dissertation, Universiti Kebangsaan Malaysia, 2013.

[10] M. S. Azmi, N. A. Arbain, A. K. Muda, Z. Abal Abas, and Z. Muslim, “Data Normalization for Triangle Features by Adapting Triangle Nature for better Classification,” 2015 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol., pp. 1–4, 2015.

[11] M. M. M. Tin, M. M. Sein, and H. Township, “Multi Triangle Based Automatic Face Recognition System By,” I2MTC 2009 - Int. Instrum. Meas. Technol. Conf., no. May, pp. 5–7, 2009.

[12] J. Zheng, Y. Gao, and M.-Z. Zhang, “Fingerprint Matching Algorithm Based on Similar Vector Triangle,” in Image and Signal Processing, 2009. CISP '09. 2nd International Congress, 2009, pp. 1–6.

[13] Z. Zhang, S. Wang, and A. I. Morphing, “Multi-feature facial synthesis based on triangle coordinate system,” Proc. 2nd Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2012, no. 2, pp. 141–145, 2012.

[14] M. Ghazvini, H. Sufikarimi, and K. Mohammadi, “Fingerprint matching using genetic algorithm and triangle descriptors,” 19th Iran. Conf. Electr. Eng., pp. 1–6, 2011.

[15] A. Gago-Alonso, J. Hernández-Palancar, E. Rodríguez-Reina, and A. Muñoz-Briseño, “Indexing and retrieving in fingerprint databases under structural distortions,” Expert Syst. Appl., vol. 40, no. 8, pp. 2858–2871, 2013.

[16] W. Yang, J. Hu, S. Wang, and J. Yang, “Cancelable Fingerprint Templates with Delaunay Triangle-Based Local Structures,” Cybersp. Saf. Secur., pp. 81–91, 2013.

[17] A. Haselhoff and A. Kummert, “A Vehicle Detection System Based on Haar and Triangle Features,” in Intelligent Vehicles Sysmposium, 2009, pp. 261–266.

[18] P. Tang, R. Jiang, and M. Zhao, “Feature selection and design olintrusion detection system based on k-means and triangle area support vector machine,” Second Int. Conf. Futur. Networks ICFN'10, pp. 144–148, 2010.

[19] N. A. Arbain, M. S. Azmi, S. S. S. Ahmad, I. E. A. Jalil, M. Z. Masud, and M. A. Lateh, “Detection on Straight Line Problem in Triangle Geometry Features for Digit Recognition,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1019–1025, 2016.

[20] P. Smith, D. B. Reid, C. Environment, L. Palo, P. Alto, and P. L. Smith, “A Threshold Selection Method from Gray-Level Histograms,” IEEE Trans. Syst. Man. Cybern., vol. 9, no. 1, pp. 62–66, 1979.

[21] J. R. A. Salem, “Segmentation Methods of Arabic Handwriting Using Neighbourhood Information For Voronoi Diagrams,” Doctoral dissertation, Universiti Kebangsaan Malaysia, 2017.

Efficient Reduction of Overgeneration Errors for Automatic Controlled Indexing with an Application to the Biomedical Domain

Samassi Adama[1], Brou Konan Marcellin[2], Gooré Bi Tra[3], Prosper Kimou[4]
Ecole Doctorale Polytechnique
Institut National Polytechnique (INP-HB)
Yamoussoukro, Côte d'Ivoire

***Abstract*—Studies on MetaMap and MaxMatcher has shown that both concept extraction systems suffer from overgeneration problems. Over-generation occurs when the extraction systems mistakenly select an irrelevant concept. One of the reasons for these errors is that these systems use the words to weight the terms of the concepts. In this paper, an Integer Linear Programming model is used to select the optimal subset of extracted concept mentions covering the largest number of important words in the document to be indexed. Then each concept mentions that this set is mapped to a unique concept in UMLS using an information retrieval model.**

***Keywords*—Concept extraction; concept recognition; automatic controlled indexing; controlled vocabulary; information retrieval**

I. Introduction

With the fast evolution of scientific publications in the biomedical field, the availability of these resources on the Internet requires automatic indexing methods. For this, specialized search engines such as PubMed in the USA and CisMEF in France have been developed. They use the concepts of the MeSH (Medical Subject Headings) thesaurus to index their document resources. This allows users (health professionals, students, patients) to retrieve relevant documents.

Indexing is the representation of a document by terms (keywords, sentences or concepts). It can be done manually or automatically. Manual indexing is efficient but it is costly in terms of time and human resources. On the other hand, automatic indexing is less efficient, but it is faster than manual indexing and requires less time and human resources.

Several automatic indexing approaches have been proposed. They can be classified into three categories: approaches based on the free extraction of terms, approaches based on controlled vocabularies (ontology, thesaurus, dictionary) and hybrid approaches that combine these two approaches. The first category represents the document by the keywords it contains without using controlled vocabulary. The second category such as MaxMatcher [3] and MetaMap [4] uses the terms of a controlled vocabulary to index the document. Approaches in this category can be subdivided into four sub-categories: (1) language-based approach, (2) machine learning approach, (3) statistic-based approach and (4) approach-based on searching in a dictionary. Language rule-based approaches define particular rules for describing terms that designate concepts. These rules are set manually by experts and depend on the characteristics of the language used.

Approaches based on machine learning use corpora manually annotated to train classifiers who consider several characteristics of textual instances to associate technical terms with predefined classes. However, these approaches are dependent on the availability of training data.

Most approaches based on statistical measures combine statistical information such as the frequency of the terms (TF), the inverse of the document frequency (IDF).

Dictionary-based approaches use terminological resources to compare textual instances with the entries (terms) of concepts in the dictionary. This search is based on the exact or partial matching between the textual fragments of the document and the entries (terms designating concepts) of the dictionary. However, exact matching leads to sub-generation problems. The partial matching leads to over-generation problems. Over-generation occurs when the system mistakenly considers a concept to be relevant because it contains a word with a high weight. This is the result of the use of simple words in the rough comparison between dictionary concepts and noun groups identified in the text as candidate concepts. The sub-generation is linked to ignorance of relevant concepts. This is the result of the strict comparison between the nominal groups of the text and the concepts of the dictionary.

In this article, the authors focus on dictionary-based approaches. The goal is to reduce the over-generation errors of concepts extraction systems based on search in a dictionary. The proposed approach is based on recent methods [1,2] of solving over-generation errors.

According to Boudin et al. [2], one of the reasons for over-generation errors is that in extraction systems, candidate concepts are selected based on the weight of their constituent words as in Zhou et al. [3]. As a result, an irrelevant concept containing a significant word can be selected. The selection of the candidate terms according to their constitutive words makes it possible to reduce the over-generation errors provided that the weight of each word is calculated only once in the set of extracted terms [2]. Thus, Boudin et al. [2] proposed an integer linear programming model for extracting key terms. This model reduces over-generation errors by weighting candidate terms as a set rather than independently. In this model, key terms are selected based on their constituent words and the weight of each unique word. The main contribution of this paper is the proposal of a method for reducing the over-generation errors of extraction systems based on the search in a dictionary. It is summarized in these points:

- Identification of some problems related to the extraction of concepts based on the search in a dictionary.
- Identification of methods of literature that can provide solutions to these problems.
- Proposal of a method for reducing the over-generation errors in an extraction system.

The rest of the article is organized as follows. In Section 2 the authors present MetaMap and MaxMatcher, two state-of-the-art concept retrieval systems based on research in a controlled vocabulary. Next, they present two methods of solving the over-generation errors identified in the results of the extraction of the two previous systems. In Section 3 they describe their approach for reducing over-generation errors of search-based retrieval approaches in a dictionary. Section 4 presents the discussion. They conclude and present some ideas for future work in Section 5.

II. Related Work

In this section, the authors first introduce MetaMap and MaxMatcher, two state-of-the-art retrieval approaches based on search in a dictionary of terminology concepts. Next, they present two methods of solving the over-generation errors identified in the results of the extraction of the two previous systems.

A. MetaMap

MetaMap [4] is a tool for extracting the concepts of UMLS from biomedical documents. The MetaMap extraction process consists of the following five main steps:

- Identification of nominal groups in the text using an analysis grammatical.
- Generation of variants (synonyms, acronyms, ...) for each group nominal using the SPECIALIST Lexicon resource of the UMLS,
- Selection of candidate concepts: a concept with at least one word found in one of the variants is retrieved (this leads to over-generation and sub-generation problems),
- Concept evaluation: The candidate concepts are compared with the original text using the following four measures: centrality, variation, coverage and consistency. The candidate concepts are finally ordered according to the final score.
- Correspondence construction: for each document, the concepts are assigned according to their similarity score with it.

Among the disadvantages of MetaMap we have the over-generation problems, the under-generation issues and the data processing time.

1) Over-generation errors: Over-generation occurs when the system mistakenly selects an irrelevant concept. This is the result of the use of simple words in the comparison between dictionary concepts and noun groups identified in the text as candidate concepts. For example, for the nominal group "ocular complications", MetaMap selects the three concepts "Ocular", "Complications" and "Complications Specific to Antepartum or Postpartum" because they share at least one word.

2) Sub-generation errors: The sub-generation is linked to the non-selection of relevant concepts. This is the result of the strict comparison between each nominal group of the text and the concepts of the dictionary. For example, for the expression "gyrb and p53 protein", MetaMap can't identify the word "gyrb" as a protein because it is registered in the UMLS as "gyrb protein".

3) Data processing time: Another disadvantage of MetaMap is its data processing time. Indeed, this tool uses a set of sophisticated linguistic methods such as grammatical analysis, the generation of variants, the search in the whole of the Metathesaurus, as well as the calculation of several statistical measures.

B. MaxMatcher

Zhou et al. [3] proposed MaxMatcher, a generic extraction approach based on the approximate search for strings in a dictionary of terms designating concepts. The basic idea of this approach is to index documents with only the most significant words of the UMLS meta-thesaurus concepts.

1) Concept Recognition: For a document, MaxMatcher cuts it into sentences and then identifies biological concept names (terms). For a given text, a set of rules to identify the boundary of a biological concept name. A biological concept term should begin with a noun, a number, or an adjective while ending with a noun or a number. It can not contain any boundary words including: punctuations (except hyphen, period, and single quote), verbs, and conjunctions and prepositions (except “of”). Whenever a boundary word is encountered, a candidate concept term reaches its end and it is then extracted.

2) Concept Normalization: The task of mapping a biological term to a concept in a controlled vocabulary, typically to the standard thesaurus in the Unified Medical Language System (UMLS), is known as medical concept normalization.

After the concept recognition step, MaxMatcher identifies the extracted terms that correspond to concept entries (terms) in a dictionary of biological concept terms.

Let t = {w_1, w_2, ..., w_m} be a candidate term (extracted from the text) consisting of a set of simple words, N(w) the number of concepts whose variant names contain word w, w_{ji} the i-th word in the j-th variant name of the concept. The similarity between each of its words w_i and a concept c of UMLS, denoted by a set of n variant names (terms) $\{v_1, v_2, ..., v_n\}$, is defined in [3] as follows :

$$I(w_i, c) = \max\{I(w_i, v_j) | j \leq n\} \quad (1)$$

where :

$$I(w, v_j) = \left\{ \begin{matrix} \sum_i \frac{1/N(w)}{1/N(w_{ji})} & \text{if } w \in v_j \\ 0 & \text{else} \end{matrix} \right\} \quad (2)$$

According to equation (1), candidate concepts are selected based on the weight of their constituent words. As a result, an irrelevant concept containing a significant word can be selected. This is an over-generation error. It is the result of partial matching method used by MaxMatcher. However, according to Boudin et al. [2], the selection of the candidate terms according to their constitutive words makes it possible to reduce the over-generation errors provided that the weight of each word is calculated only once in the set of extracted terms.

C. *Automatic Keyphrase Extraction Approaches*

Keyword extraction is the task of automatically identifying a set of terms that best describe a text document [10]. Automatic keyword extraction has been found to be useful for many natural language processing applications such as information retrieval, automatic indexing and classification of text documents, automatic summarization [11,12]. However, state-of-the-art keyword extraction systems suffer from over-generation errors.

According to **Boudin et al**. [2], the selection of key terms according to their constituent words makes it possible to reduce over-generation errors, provided that the weight of each word is calculated only once in the set of these terms. The key-term extraction model they propose has three steps: (1) Extraction of candidate terms using heuristic rules (2) weighting of words using supervised or unsupervised methods (3) optimal subset of key terms by integer linear programming.

Jia et al. [1] proposed an unsupervised method of extracting key terms. According to these authors, unsupervised methods for extracting existing key-words suffer from over-generation error because they generally identify keywords and then return as keywords the terms of the text containing these keywords. In other words, key word extraction systems first assign scores to the words, then rank the candidate key terms based on the sum of the weights of their constituent words. To overcome this problem, Jia et al. proposed a weighting scheme that is applied directly to candidate key terms by exploiting some of their properties such as informativeness and positioning preference.

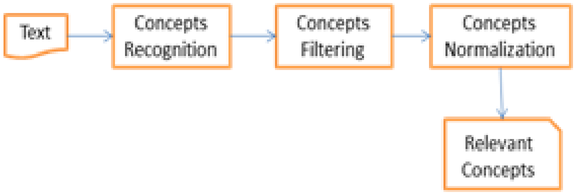


Fig. 1. The Proposed Approach.

III. Proposed Approach

The concept extraction approach proposed in this section is based on three steps (Figure 1): (1) Concepts Recognition; (2) Concepts Filtering; (3) Concept Normalization.

A. *Concepts Recognition*

Concepts recognition consists in the identification of the mentions (terms) of biological concepts in a textual document. For a given text, we used a set of rules to identify the boundary of a biological concept term (concept name) as in [3]. A biological concept name should begin with a noun, a number, or an adjective while ending with a noun or a number. It can not contain any boundary words including: punctuations (except hyphen, period, and single quote), verbs, and conjunctions and prepositions (except "of"). In other words, whenever a boundary word is encountered, a candidate concept mention reaches its end.

Let V = $\{T_1, T_2, ..., T_P\}$ be the set of recognized concept mentions extracted from the document. All these concepts are not relevant with respect to this document. We must select the optimal subset of these mentions covering the largest number of important words in the document.

B. *Concepts Filtering*

Filtering concepts consists in selecting of the optimal subset of concept mentions covering the largest number of important words in a document. The authors used the model of Boudin et al. [2] to find the optimal subset of concept mentions. The model is defined as

$$\max \sum_i \omega_i x_i - \lambda \sum_j \frac{(l_j - 1)c_j}{1 + substr_j} \quad (3)$$

$$s.t \sum_j c_j \leq N \quad (4)$$

$$c_j Occ_{ij} \leq x_i, \quad \forall i, j \quad (5)$$

$$\sum_j c_j Occ_{ij} \geq x_i, \quad \forall i \quad (6)$$

$$x_i \in \{0,1\} \quad \forall I \qquad\qquad c_j \in \{0,1\} \quad \forall j$$

where w_i (computed using Equations (7,8) is the weight of a word i, x_i and c_j are two binary variables indicating the presence of word i and candidate concept j in the set of extracted concepts, l_j is the size of concept j, $substr_j$ is the number of times concept j appears as a substring in the other concepts, Occ_{ij} is an indicator of the occurrence of word i in concept j and N the number of candidate concepts.

1) Word weighting Functions: The performance of the model of Boudin et al. [2] depends on how word weight w_i is estimated. It can be computed using one of the following unsupervised weighting functions: BM25 [7] and TFxIDF [6].

TF.IDF

$$TF \times IDF(t, d) = tf(t, d) \times \log(\frac{N}{n}) \quad (7)$$

Where tf(t, d) is the frequency of the word t in a document d, N is the number of documents in the corpus, n is the number of document containing t.

BM25

$$\mathrm{BM25}(t,d) = \mathrm{tf}(t,d)\frac{\log(\frac{N-n+0.5}{n+0.5})}{\mathrm{tf}(t,d)+k_1((1-b)+b\frac{l_d}{\mathrm{avgl}_d})} \quad (8)$$

where tf(t, d) is the frequency of the word t in a document d, N is the number of documents in the corpus, n the number of documents containing the word t, l_d the length of a document d, n the number of documents containing the word t, avgl_d the average document length (number of words in the document) ; k_1 and b are free parameters.

Once the optimal set of concept mentions is found, each of them needs to be normalized, if possible, with a unique identifier (CUI) from the Unified Medical Language System (UMLS) metathesaurus.

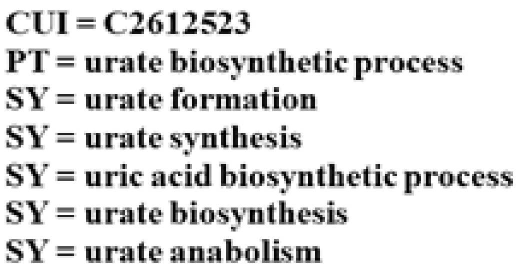
C. Concepts Normalization

The task of mapping a concept mention in a text to a semantically equivalent concept in a biological knowledge base (like UMLS) is known as concept normalization. In this study, each concept mention is mapped to a Concept in the UMLS metathesaurus. This way, a semantic meaning is associated to each of them. In the UMLS each concept is given a Concept Unified Identifier (CUI). Each synonym and abbreviation of this concept is called Term. A term is either Preferred Term (PT) or Synonym (SY) (figure 2).

Concept normalization is challenging because: (1) the same word or term can be used to refer to different concepts, and (2) the same concept can be referred to by different words or terms, (3) the different expressions (terms) of a concept are not necessary all present in the knowledge base.

Let T = {T_1, T_2, ..., T_N} be the optimal subset of the set V(the set of recognized concept mentions) containing the N concept mentions extracted from the document. In the proposed concept normalization method, each concept mention T_i is treated as a query, while the concepts in the UMLS are treated as documents that are searched to find the relevant concepts. So, all the concepts in the UMLS metathesaurus are indexed.

Formally, a concept mention is modeled as a sequence T_i of one or more words {t_1, t_2, ..., t_n}. A concept in UMLS is modeled as a concept-document C_j, which is a sequence of one or more words {t_1, t_2, ..., t_n}.



CUI = C2612523
PT = urate biosynthetic process
SY = urate formation
SY = urate synthesis
SY = uric acid biosynthetic process
SY = urate biosynthesis
SY = urate anabolism

Fig. 2. The Concept Urate Biosynthetic Process Identified by the Unique Identifer C2612523 in UMLS Metathesaurus.

The authors cast the concept normalization task in an information retrieval problem as in [5,8]. All the concepts in the UMLS metathesaurus are indexed. Thus, given a concept mention T_i (a query), retrieve the relevant concept-documents (concepts) C_1, C_2, ..., C_k from this index. Standard Information Retrieval (IR) models can be used on the concept mapping problem. In this paper we used the BM25 to rank the concept-documents for a given concept mention T. Thus, the score of a concept-document (a concept in UMLS) for a concept mention T is defined as

$$\mathrm{BM25}(T,C) = \sum_{t\in(T\cap C)} \mathrm{BM25}(t,C) \quad (9)$$

with :

$$\mathrm{BM25}(t,C) = \mathrm{tf}(t,C)\frac{\log(\frac{N-n+0.5}{n+0.5})}{\mathrm{tf}(t,C)+k_1((1-b)+b\frac{l_C}{\mathrm{avgl}_C})} \quad (10)$$

where tf(t, C) is the frequency of the word t in a concept C, N is the number of concepts in the UMLS metathesaurus, n the number of concepts containing the word t, l_C the length of a concept C, n the number of concepts containing the word t, avgl_C the average concept length (number of words in the concept) ; k_1 and b are free parameters.

Finally, each concept mention T is mapped to the concept C which has the maximum score.

IV. Discussion

Dictionary-based concept extraction is the state-of-the-art approach to biomedical literature indexing. In this work, the authors are interested in reducing the over-generation errors of MaxMatcher [3], which is a concept extraction system based on search in a dictionary. One reason these errors is that this system ranks extracted concepts according to the weights of their component words (equation 1). This approach poses a major problem: if a word is very important then all the terms that contain it will also be considered important. As a result, irrelevant concepts containing a significant word are selected. To reduce the number of these irrelevant concepts, the idea is to search for the optimal subset of extracted concepts covering the largest number of important words in the document. According to Boudin et al. [2], finding this optimal set of concepts is a combinatorial optimization problem, and can be formulated as an integer linear program. However, this approach [2] is supervised. Thus it requires large amounts of labeled training data. At the same time, unsupervised systems like [1] have poor accuracy and do not generalize well.

Jia et al. [1] proposed to directly weight candidate key terms by considering some of their properties such as informativeness and positioning preference. Such approach can be used to reduce over-generation errors. Since ambiguous concept mention can be mapped to multiple concepts in the referenced ontology (UMLS) depending on the context, one of the main challenges in the concept normalization task consists in the disambiguation of these cases [9].

V. Conclusion

In this paper, authors analyzed some works on MetaMap and MaxMatcher, two concept extraction systems based on the search for strings in a dictionary of terms designating concepts. They found that both systems suffer from over-generation errors. So they proposed to use an Integer Linear Programming model to select the optimal subset of extracted concepts that are relevant to the document to index. Then each concept mention of this set is mapped to a unique concept CUI in the UMLS metathesaurus. The authors cast the mapping task in an information retrieval problem, using a concept mention as query and the concepts in ULMS as documents.

In future work, we plan to test our method using the OHSUMED collection. Since ambiguous concept mention can be mapped to multiple concepts in the referenced ontology (UMLS) depending on the context, one of the main challenges in the concept normalization task consists in the disambiguation of these cases.

References

[1] Jia, H., & Saule, E. (2018). Addressing Overgeneration Error: An Effective and Effcient Approach to Keyphrase Extraction from Scientific Papers. In BIRNDL@ SIGIR (pp. 60-73).

[2] Boudin, F. (2015, July). Reducing over-generation errors for automatic keyphrase extraction using integer linear programming. In ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction.

[3] Zhou, X., Zhang, X., & Hu X. (2006). MaxMatcher: Biological concept extraction using approximate dictionary lookup. PRICAI 2006: trends in artificial intelligence, 1145-1149.

[4] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

[5] Mirhosseini, S., Zuccon, G., Koopman, B., Nguyen, A., & Lawley, M. (2014, November). Medical free-text to concept mapping as an information retrieval problem. In Proceedings of the 2014 Australasian Document Computing Symposium (p. 93). ACM.

[6] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28:11–21.

[7] Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In TREC-7, pages 199–210.

[8] Ruch, P. (2005). Automatic assignment of biomedical categories: toward a generic approach.

[9] Leal, A., Martins, B., & Couto, F. (2015). ULisboa: Recognition and normalization of medical concepts. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 406-411).

[10] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

[11] Balaji, J., Geetha, T. V., & Parthasarathi, R. (2016). Abstractive summarization: A hybrid approach for the compression of semantic graphs. International Journal on Semantic Web and Information Systems (IJSWIS), 12(2), 76-99.

[12] Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1262-1273).

Optimizing the Behaviour of Web Users Through Expectation Maximization Algorithm and Mixture of Normal Distributions

R. Sujatha[1]
Department of Mathematics
SSN College of Engineering, Chennai, India

P. Saravanan[3]
Research Scholar, Bharathiyar University
G B Pant Govt. Engineering College, New Delhi, India

D. Nagarajan[2]
Department of Mathematics
Hindustan Institute of Technology & Science, Chennai, India

J. Kavikumar[4]
Department of Mathematics and Statistics, Faculty of Applied Science and Technology
Universiti Tun Hussein Onm Malaysia, Malaysia

***Abstract*—The proposed work is to analyse the user's behaviour in web access. Worldwide, the web users are browsing through different websites every second. Aim of this paper is to identify the behaviour of user's in a time bound using an Expectation Maximization (EM) algorithm and the maximum likelihood estimates of the model parameters. A novel approach based on Mixture normal distribution is used to discuss the percentage of user along with web page frequency.**

***Keywords*—EM algorithm; maximum likelihood; mixture normal distribution; web page frequency**

I. Introduction

The number of accessible web pages grows significantly; it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. Users must either browse through a large hierarchy of concepts to find information or submit a query to a widely available search engine [1]. Therefore, the process of understanding the user's navigation behaviour is challenging but fundamental in improving web query answering, link structure and in simplifying navigation through a large number of individual webpages. The web sites are making great effort to understand user's behaviour and make the web sites easy to access. To achieve this goal, researchers proposed lots of approaches to use web usage data.

Researchers studied this topic from different points of view. A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data or hidden data is presented at various levels of generality [2], [3], [4] and [5]. EM algorithm to retrieve the complete scatterer trajectory matrix is discussed in [6].

Mixture distributions are extensively used to model a wide variety of empirical phenomena, in diverse fields such as biology, anthropology, psychology, economics, and marketing. Overviews of mixture distributions and many examples of their applications are given by [7]. Mixtures of t-distributions and their numerous variants are discussed by [8], [9], [10] and [11]. EM algorithm and finite mixture model is discussed in [12]. The EM-GMM algorithm targets reconfigurable platforms, with five main contributions [13].

In this paper we have studied the web user's behaviour using EM algorithm. The web page access is predicted using mixture normal variate. The remaining of the paper is organized as follows. In section 2 we present the concept of EM algorithm. Section 3 gives the application of EM algorithm to the selected database. In Section 4 we deal with mixture normal variate and its application in predicting web page frequency and finally concluded in Section 5.

A. Data Base

The data is taken from the educational institute of Sri Sivasubramaniya Nadar College of Engineering (SSNCE), Chennai, Tamil Nadu, India.

II. Concept of EM Algorithm

The EM algorithm is a general method, to estimate the parameters using maximum-likelihood estimation.

EM algorithm is used when the data is incomplete, due to the limitations of the observation process. The algorithm consists of two steps. This is diagrammatically shown in Fig.1.

Given a set of parameter estimates the E-step calculates the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates. In this step, using conditional expectation, given the observed data and current estimate, the missing data is estimated. Given complete-data log likelihood, the M-step finds the parameter estimates in order to maximize the complete-data log likelihood from the E-step. These two steps are iterated until convergence. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

The sequences of web users randomly access the SSN educational institute website for various departments. Web user access 6 engineering departments with 4 independent various attributes in each department. For application of EM algorithm

This work is part of the project SR/S4/MS:816/12, Science and Engineering Research Board, Department of Science and Technology, India.

the dataset corresponds to page views of a user. To predict the accessibility of various departments among users EM Algorithm is applied. From the Internet browsing logs, we could gather the following information about a web user the frequency.

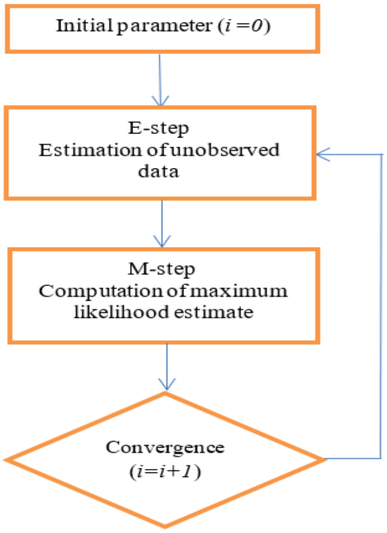

Initial parameter ($i=0$)
E-step
Estimation of unobserved data
M-step
Computation of maximum likelihood estimate
Convergence
($i=i+1$)

Fig. 1. EM Algorithm

III. Application of EM Algorithm

The sessions are grouped based on the user's profile. The sessions are grouped as various departments, namely, EEE, ECE, MECH (MEC), CIVIL (CIV), IT and BME. The considered webpages are Events, Faculty (Fac), Research (Res) and News. The webpages considered is shown in Fig. 2. To determine the proportion of usage of various departments, we determine the likelihood of the webpage access and the sessions by using EM algorithm.

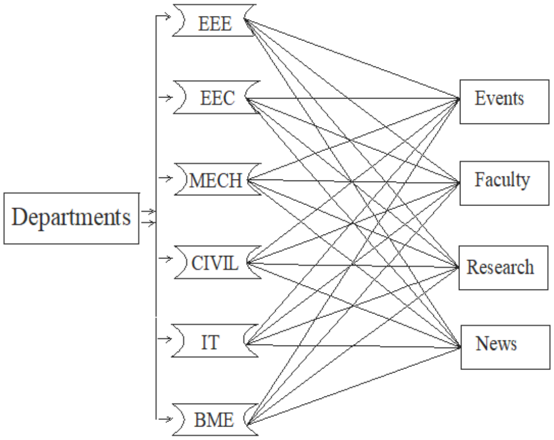


Fig. 2. Grouped webpages

TABLE I. Initial and E-Step of the Web User

Dept.	EEE	ECE	MEC	CIV	IT	BME
Event	0.487	0.390	0.335	0.405	0.361	0.303
Fac	0.252	0.170	0.262	0.227	0.468	0.278
Res	0.162	0.317	0.012	0.025	0.117	0.177
News	0.121	0.121	0.387	0.341	0.053	0.240

TABLE II. Final and M-Step of the Web User

Dept.	EEE	ECE	MEC	CIV	IT	BME
Event	0.527	0.422	0.363	0.405	0.361	0.347
Fac	0.291	0.150	0.264	0.224	0.498	0.214
Res	0.201	0.345	0.114	0.051	0.196	0.200
News	0.147	0.151	0.314	0.342	0.053	0.296

The initial values *i.e.*, Expectation values are depicted in Table I as the values for E-step. By application of the Maximization step, the updated values are shown in Table II. Based on the calculations, the accessibility for each department can be determined and the results are depicted in Fig. 3.

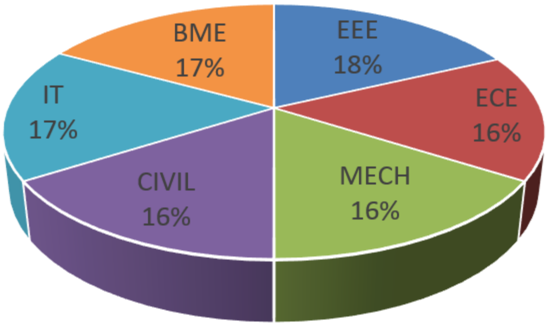


Fig. 3. Web User Activity

IV. Mixture of Normal Variates to Predict Percentage of User and Web Page Frequency

It is essential to predict the percentage of usage and web page frequency to understand the accessibility and popularity of the website among users. In this paper, we have used mixture of normal variates for this purpose.

Mixture of normal variates is used in statistical methods. Random vector x has a normal variate and it can be written as linear combination of variables from vectors x, all the samples of x variables from normal variates. It is independently distributed with zero covariance. The density function of a mixture of two univariate normal distributions is $f(y;w) = pf_1(y;w) + (1-p)f_2(y;w)$, where $f_j(y;w) = \frac{1}{\sigma_j}\phi\left(\frac{y-\mu_j}{\sigma_j}\right)$,$j = 1$ and $\phi(.)$ is the standard normal distribution [14], [15], [16] and [17]. The interpretation of this system consists of mixture of two population and p lies between zero and one. The component of two mixture normal variates $\sum_j \sigma_j^2\, i$ where i is the unit matrix $N\left(x_n/X_n, \sigma_j^2 i\right) = \frac{1}{(2\pi)^{\frac{1}{2}}}\frac{1}{\sigma_j}$.

If $\sigma_j \to 0$ then the term goes to infinity. The variance of mixture components are finite and finite probability to all points. While other components can shrink onto the data point thus contributing the data point increasing additive value to the log likelihood. Two mixture A, B of normal distribution with mean μ_a, μ_b and standard deviations σ_a, σ_b to take mixture of distribution p and q where $0 \leq p \leq 1$, $q = 1 - p$. Therefore the mixture of mean is $\mu_{ab} = (p \times \mu_a) + (q \times \mu_b)$ [18]. The mixture of the resulting normal curve is estimated using MATLAB and the results are shown in Fig. 4. From the graph, shown in Fig. 4 we observe that variance and mean are different. It is an equally weighted average of the bell-shaped probability density function of the two normal distributions. The weights were not equal, the resulting distribution could still be bimodal but with peak of different height and split-up is a linear combination of two normal variates with means 11 and 18; variance 0, 1 and 4, given by *0.5N(11,1)+0.5N(18,1)* and *0.75(11,0)+0.25N(18,4)*.

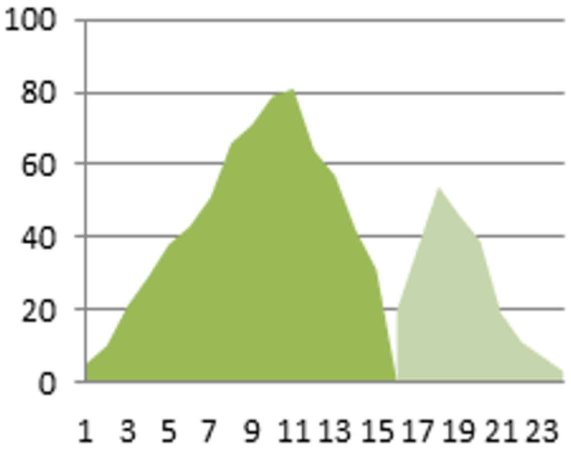


Fig. 4. Percentage user vs web page frequency

V. Conclusion

In this paper we proposed a method of using EM algorithm to predict the accessibility of webpages among users. We have used mixture distribution to identify web page frequency and percentage of users. Based on these the popularity of the web pages among users can be studied. The frequently accessed web pages can be updated. The study reveals that EEE department is popular among the users and is accessed much frequently when compared to the other departments. The study can be extended to centrality of networks.

References

[1] G. Pallis, L. Angelis, A. Vakali and J. Pokorny, “A Probabilistic Validation Algorithm for Web Users Clusters”, IEEE International Conference on Systems, Man and Cybernetics, pp. 4129–4134, 2004.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp.1–38, 1977.

[3] D. Chauveau, “A stochastic EM algorithm for mixtures with censored data”, Journal of Statistical Planning and Inference, vol. 46, pp.1–25, 1995.

[4] R. A. Levine, and G. Casella, “Implementations of the Monte Carlo EM algorithm”, Journal of Computational and Graphical Statistics, vol. 10(3), pp.422–439, 2001.

[5] S. Balakrishnan, M. J. Wainwright and B. Yu, “Statistical Guarantees For The EM Algorithm: From Population To Sample-Based Analysis”, The Annals of Statistics, vol. 45(1), pp.77–120, 2017.

[6] L. Liu, F. Zhou, X. Bai, J. Paisley and H. Ji, “A Modified EM Algorithm for ISAR Scatterer Trajectory Matrix Completion”, IEEE Transactions on Geoscience and Remote Sensing, vol. 56(7), pp. 3953-3962, 2018.

[7] G. J. McLachlan and D. Peel, Finite Mixture Models, Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, 2000, New York.

[8] C. Archambeau and M. Verleysen, “Robust Bayesian clustering”, Neural Networks, vol. 20(1), pp. 129–138, 2007.

[9] C. M. Bishop and M. Svensen, “Robust Bayesian mixture modelling”, Neurocomputing, vol. 64, pp. 235–252, 2005.

[10] J. L. Andrews and P. D. McNicholas, “Model-based clustering, classification, and discriminant analysis via mixtures of multivariate tdistributions”, Statistics and Computing, vol. 22(5), pp. 1021–1029, 2012.

[11] F. Forbes and D. Wraith, “A new family of multivariate heavy-tailed distributions with variable marginal amounts f tail weight: application to robust clustering”, Statistics and Computing, vol. 24(6), pp. 971– 984, 2014.

[12] Y. Li and Y. Chen, “Research on Initialization on EM Algorithm Based on Gaussian Mixture Model”, Journal of Applied Mathematics and Physics, vol. 6, pp. 11-17, 2018.

[13] C. He, H. Fu, C. Guo, W. Kuk and Guangwen, “A fully pipelined hardwared design for Gaussian mixture models”, IEEE Transactions on Computers, vol. 66(11), pp. 1837–1850, 2017.

[14] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, Very Large Data Base Endowment, pp. 487-499, 1994.

[15] J. Xiao and Y. Zhang, “Clustering of Web Users Using Session-based Similarity Measures”, International Conference on Computer Networks and Mobile Computing, pp. 223-228, 2001.

[16] S. Vishwakarma, S. Lade, M. K. Suman and D. Patel, “Web user prediction by integrating markov model with different features”, International Journal of Engineering Research and Science & Technology, vol. 2(4), pp.74-83, 2013.

[17] M. Deshpande and G. Karypis, “Selective Markov Models for Predicting Web-Page Accesses”, ACM Transactions on Internet Technology, vol. 4(2), pp.168-184, 2004.

[18] E. Meijer and J. Y. Ypma, “A Simple Identification Proof for a Mixture of Two Univariate Normal Distributions”, Journal of Classification, vol. 25, pp.113-123, 2008.

Application of Fuzzy Analytical Hierarchy Process based on Geometric Mean Method to Prioritize Social Capital Network Indicators

Vy Dang Bich Huynh[1]
Department of Learning Material
Ho Chi Minh City Open University (HCMCOU)
Ho Chi Minh City, Vietnam

Phuc Van Nguyen[2]
Ho Chi Minh City Open University (HCMCOU)
Ministry of Education and Training (Hanoi)
Vietnam

Quyen Le Hoang Thuy To Nguyen[3]
Office of Cooperation and Research Management
Ho Chi Minh City Open University (HCMCOU)
Ho Chi Minh City, Vietnam

Phong Thanh Nguyen[4]
Department of Project Management
Ho Chi Minh City Open University (HCMCOU)
Ho Chi Minh City, Vietnam

Abstract—**Vietnam is striving to develop dynamically and overcome many human resource challenges. As the economy expands, the demand for jobs and human resource development has become increasingly critical. The pressures from reform and international integration are forcing many changes. New graduates must provide proof of their academic capabilities, while also actively developing their social capital to support their job search process. In fact, social capital is an essential capital for personal development as well as professional development of new graduates. This paper applies Fuzzy Analytical Hierarchy Process based on Geometric Mean Methodology to evaluate factors for measuring the social capital of graduates at Ho Chi Minh City Open University in Vietnam. The research results highlight the important role of social networks for graduates, in which a linking network is the most important, a bonding network is the second most important, and a bridging network is the third most important. In addition, the research shows that trust plays an even more important role than networks; and specific belief is more important than general belief.**

Keywords—Education; human resource; fuzzy analytic hierarchy process, geometric mean; social capital network

I. Introduction

According to the Ministry of Labor, Invalids and Social Affairs, by the end of 2017, about 215,000 Vietnamese people with university degrees were unemployed. The situation shows that after graduation a certain percentage of students cannot find a job. This situation creates a tremendous waste of social resources and thus is an important issue to address. Various researchers have proposed solutions in terms of human capital development, and in particular focused on enhancing students' soft skills. The aim in this approach is to improve the likelihood of success in their job search process and to also obtain a higher starting salary. However, this is just one of the factors contributing to such success.

In the last few decades, interest and attention about social capital from scholars and researchers in various fields have been growing. In 1916, Hanifan, an American educator, introduced the concept of social capital. According to him, social capital refers to goodwill, friendship, sympathy, and social interaction between individuals and families [1]. More recent research all conclude that social capital should be regarded as the fourth necessary capital for individuals and communities, which is in addition to other traditional capital such as natural capital, physical capital and human capital. Therefore, it is necessary to develop an index to measure the social capital of students at Ho Chi Minh City Open University. Such research constitutes the foundation for future micro-research on the impacts of social capital in order to develop a strategy to maximize students' social capital. This paper presents different types of social capital as well as the development of a model to measure social capital of graduates from Ho Chi Minh City Open University.

II. Literature Review

According to Hanifan [1], social capital refers to goodwill, friendship, sympathy, and social interaction between individuals and families. When an individual meets a neighbor, this interaction will acquire social capital to meet various social needs and potentially improve the life standards of the community. Bourdieu [2] expands the concept of social capital by holding that all known networks contribute to the creation of social capital. This point is further clarified by Leonard and Onyx [3] as he points out that social capital is formed and developed by the participation of individuals in organizations and groups of people with the same interest. This is the structural angle of social capital. Paldam [4] approaches social capital from three categories of characteristics: trust, cooperation, and networking. In fact, trust is the foundation for standard behaviors, collaboration and networking, enabling members to act more effectively to pursue common goals.

According to social capital is a person's ability to work voluntarily with others and the pre-condition for this is social standards. Standards are understood as behavioral attitudes shared by most of individuals or groups in the society that are reinforced by sanctions. Such a standard may be a philosophy, religious doctrine or professional standards, or rules of conduct [5]. In addition, for everyone to act in accordance with a standard, trust is required as a condition.

Trust is defined as the willingness of a person to accept other people's actions based on the expectation that the other person will perform an action that is important without supervision or control. [6]. In other words, trust is the belief in others [7]. Trust is a core condition for standard behaviors, which is the source of cooperation. This is an important component in the formation of social capital. Thus, it could be said that social capital is a multidimensional, multidisciplinary concept and, subject to the purpose, approach and investigational method, and researchers can determine the inherent contents of social capital differently.

The concept of social capital is classified by three functions: bonding, bridging and linking [8, 9]. Bonding social capital refers to the trust and relationships within a group, such as among family members, relatives, and close friends [10, 11]. The definition of bonding social capital shows that individuals may be closely bound to and "strongly" believe in other members in the network. From a structural component, this is a strong, close, and unofficially organized network.

Bridging social capital refers to the trust and network of inter-group relationships and networks among different communities. This type of social capital is beneficial in connecting external resources and helping to spread information. This view emphasizes the importance of loose relationships and general trust [9, 12]. This network is often open and formally organized, typically in the form of a voluntary association or group. It is identified as including: (1) community networks, such as groups, associations (charities, clubs, skill clubs, physical training and sport teams, etc.) in the community, and (2) other social organizations and networks.

Linking social capital has the characteristic of having only one vertical link [13, 14]. It refers to inter-class relationships, typically hierarchical relationships and trust in the institution, State, Party, Youth Union, Student Union, etc. Accordingly, this type of social capital is the key leverage to spread resources, concepts, and information among official organizations in the community. This is a very important factor in economic development.

From a cognitive perspective, trust is the basis of normative behavior and interrelationships. It is the source of all cooperation efforts aimed at achieving common goals [15]. Harpham [16] assumed that trust can be classified into two types, corresponding to the functions of bonding, bridging, and linking social capital, including: (1) specific trust: the belief in familiar persons in the closed network (bonding); and (2) generalized trust: the belief is open to strangers (close relationship), organizations and institutions (bridging, linking).

Many organizations and researchers have tried to model a social capital measurement index. Putnam [8] proposes the simplest method to measure social capital, which is the foundation for subsequent studies on social capital. The author analyzed institutional efficiency between northern and southern regions of Italy, which reflects structural components of social capital and did this by looking at participation in voluntary organizations to help explain much of the difference in institutional efficiency. The author also added a trust factor to his social capital measurement model.

Xue [17] has developed a social capital indicator based on a review of micro- and medium-scale social networks (relatives, friends, associations) and contents of the network (size, quality) in his study to explore the role of social networks in terms of the probability of finding a new job for new immigrants in Canada. The author applied regressive models as a fixed-effects logit model, random-effects logit model, and unobserved heterogeneity. Results of the research showed that social capital has a positive impact on the employment of migrant workers in the first four years.

Kanas [18]'s research on immigrants in Germany shows that social capital has a positive impact on immigrants' professional status and income. In this study, the author used a set of table data collected from socio-economic surveys in Germany (1984-2004). Subjects of the study were foreigners working in Germany ranging in age from 20 to 60. The Heckman model was applied to control variables such as human capital, health, and social capital. Bonding and bridging social capitals were considered in the study by measuring relationship networks, volunteering activities, and migrant employment in relation to the original population. The results of the study confirm that various types of social interactions have a positive impact on the economic performance of immigrants. Migrant workers that are members of a group tend to achieve higher professional status than those that do not join a group. In particular, bridging social capital helps migrant workers develop opportunities to achieve a higher professional status, although this did not necessarily increase their income. In this study, the author did not measure social capital in terms of trust.

Mahuteau, et al. [10] uses data on migrant workers in Australia from household and labor income surveys in Australia during the 2002-2010 period. A social network index is constructed by means of principal component analysis (PCA). Social capital is measured by the following factors: (1) network participation, (2) interaction frequency, and (3) mutual contact and trust. A logit regression method was applied to answer the question, "Does social capital increase the probability of finding a job and increasing the income for immigrants?" Results of the study provide empirical evidence that the socio-economic environment has a positive impact on the probability of finding a job and increasing income for migrants, especially for women. In addition, it also shows that there is a significant difference in socio-economic status of migrant workers from different communities. For example, when the social capital index is increased by one unit while other factors remain unchanged, the odds ratio for employment was 32% for workers in general, 28% for migrant workers from English-speaking communities, and 17% for migrant workers from non-English speaking communities. Further classifying persons into white-collar workers (skilled workers) and blue-collar workers, shows that social capital has a statistically significant impact on increasing the probability of finding employment for white-collar migrants [19].

III. Research Methodology and Results

In the early 1970s and in various publications in 1980, Thomas L. Saaty, an American mathematician, introduced a multi-standard decision-making method referred to as the analytic hierarchy process (AHP) [20-22]. This is a theoretically quantitative method that assists individuals or groups in evaluating, analyzing, and making decisions on complex multi-standard issues. The goal of this method is to quantify the relationships among various priorities in a given set of options by using a rating scale based on opinions and comments [23-26]. In particular, it emphasizes the importance of the evaluator's intuitive judgments as well as consistency in comparing the factors. AHP allows an expert to apply their knowledge and combine this with objective and subjective data in a logical hierarchy [27-30]. In addition, AHP combines both components of human thinking: qualitative and quantitative. The qualitative component is expressed by hierarchical arrangement while the quantitative component is expressed by description and evaluation. Preference is expressed by figures that can be used to describe perceptions of all intangible and tangible matters. It can be used to describe a person's emotions, intuition, and evaluation.

In recent years, many socio-economic scientists have utilized approaches based on set theory, especially fuzzy set theory. When Zadeh (1965) introduced the concept of fuzzy set, it was widely applied in a variety of disciplines, especially in the fields of engineering, and for computers and medicine. However, the application of fuzzy set theory in social and economic science is very limited.

Geometric Mean Method (GMM) is constructed solely based on fuzzy set theory and was introduced by Professor Zadeh (1965). He introduced fuzzy set theory to deal with uncertainty due to inaccuracy, unclearness, and ambiguity. For example, we cannot use any mathematical formula or explanation to express a saying, such as "He is a tall man." For example, the "high" characteristic here in Vietnam may be 1.7m for many people, but others may have another number in mind [24, 31]. Science has proven that fuzzy set theory is quite effective in dealing with problems without sharp boundaries and exact numbers. Moreover, fuzzy numbers are not the same as rigid words and mathematical equations, and instead it is very close to the natural language of humans [32-34].

The following section manifests the computational process of the weights of the social capital indicators [34-39]:

Step 1. According to the experts about the relative importance of social capital indicators, the pairwise comparison matrices can be obtained. We use the fuzzy numbers determined in Table 1.

Step 2. We calculated the elements of the synthetic pairwise comparison matrix by using the geometric mean method suggested by Buckley [23]:

TABLE I. Linguistic Scales for the Relative Importance

Fuzzy number	Linguistic	The scale of the fuzzy number
1	Equal	(1,1,1)
2	Weak Advantage	(1,2,3)
3	Not bad	(2,3,4)
4	Preferable	(3,4,5)
5	Good	(4,5,6)
6	Fairly good	(5,6,7)
7	Very good	(6,7,8)
8	Absolute	(7,8,9)
9	Perfect	(8,9,10)

$$\tilde{a}_{ij} = (\tilde{a}_{ij}^1 \otimes \tilde{a}_{ij}^2 \otimes \tilde{a}_{ij}^3 \otimes ... \otimes \tilde{a}_{ij}^n) \tag{1}$$

where $\tilde{a}_{ij}$ is the fuzzy comparison value of criterion *i* to criterion *j*.

Step 3. To calculate the fuzzy weights of social capital indicators, we need to calculate [40, 41]:

$$\tilde{r}_i = (\tilde{a}_{i1} \otimes \tilde{a}_{i2} \otimes \tilde{a}_{i3} \otimes ... \otimes \tilde{a}_{in})^{1/n} \tag{2}$$

Moreover, for the weight of each criterion:

$$\tilde{w}_i = \tilde{r}_i \otimes (\tilde{r}_1 \oplus \tilde{r}_2 \oplus \tilde{r}_3 ... \oplus \tilde{r}_n)^{-1} \tag{3}$$

where $\tilde{w}_i$ is the geometric mean of the fuzzy comparison of the i^{th} criterion, which is indicated by a triangle fuzzy number $\tilde{w}_i = (Lw_i, Mw_i, Uw_i)$.

Step 4. The The fuzzy weights are defuzzified by any defuzzification method. In this paper, we applied the following CoA method [42]:

$$BNP_{w_i} = [(U_{w_i} - L_{w_i}) + (M_{w_i} - L_{w_i})]/3 + L_{w_i} \tag{4}$$

where BNP_{w_i} is the Best Nonfuzzy Performance (BNP) value of the fuzzy weights of the i^{th} criterion.

IV. Results and Discussion

Ho Chi Minh City Open University was incorporated in 1990 and is the first open-model university in Vietnam. Over the last 26 years since establishment and subsequent development, the university has built up a system of facilities and affiliates throughout southern Vietnam, from Quang Ngai to Ca Mau, and in the Central Highland provinces and coastal regions and islands such as Con Dao, Phu Quoc. With a policy of socializing education to create learning opportunities for all people with many levels of education, the university has developed many forms of educational curriculums with flexible and convenient methods for learners, with subjects ranging from business administration to science and technology and for social sciences and the humanities.

To help learners find the right jobs and be able to work after graduation, the university has developed a quality training curriculum. Additionally, the university has organized a variety of activities for students to participate in, such as activity groups, clubs, cultural events, entertainment, professional development activities, academic clubs, skills clubs, sports events, and extra-curricular activities such as Green Sunday, Saturday for Volunteer Work, student assistants, blood donation, green summer, volunteer spring, etc. Additionally, there are activities within various organizations, groups, and communities, such as the Communist Party of Vietnam, Youth Union, Student Union, etc. The annual rate of full-time students graduating from Open University is increasing in comparison with the number of enrolled students, from 46.68% in 2012 up to 70.75% in 2014 and 91% in 2015. In addition, about 90% of students gained employment after one year of graduation and about 67% of graduates have obtained work that suits their majors, and about 66% of students were employed immediately after graduation during the 5-year period of 2011-2015.

In the social capital index, the results of our study show that two components, network and trust, of graduates at Ho Chi Minh City Open University, have almost equal weights of 0.4095 and 0.5905, respectively. This shows that both components, network and trust, play a very important role in the allocation of social capital. This result is consistent with the theoretical foundation of social capital [43].

Considering the network component only, a linking network is the most important type of network with a weight of 0.495 compared to 0.346 for a bonding network and 0.172 for a bridging network. Indeed, in traditional Asian cultures such as of the Vietnamese, the bonds among family members, relatives, and friends are very tight. People are willing to share and help each other, such as by providing food, accommodation, and economic support for children when they attend colleges. In addition, the role of the network is reflected in organizations and community groups such as the Party, Youth Union, and Student Union at Ho Chi Minh City Open University. In fact, Ho Chi Minh City Open University frequently organizes many extracurricular activities, social work, field trips, etc., to create an active environment for students to interact with others, to learn, and improve their life skills as well as soft skills before graduation. By participating in such activities, especially Delegation activities, students can gain countless benefits, and establish a strong foundation for their development in the future. It is important to meet and have exchanges with alumni students to have opportunities to improve their understanding and life experience. Additionally, network opportunities provide experiences to develop skills in communication, social interaction, organization of personal work, and in organizing events for the university, unions, associations, etc. Finally, regarding the trust component, specific trust is more important than general trust. This means that individuals in the network are more likely to engage in mutual trust because of their willingness to help and support each other.

V. Conclusion

This research developed a social capital index for graduates of Ho Chi Minh City Open University that consists of two components: network and trust. In particular, the network component includes bonding, bridging, and linking networks. The trust component consists of general trust and specific trust. The application of Fuzzy Analytical Hierarchy Process based on Geometric Mean Method shows the important roles of networks, where a linking network is the most important, followed by a bonding network, and bridging network. Additionally, trust plays a more important role than networks; and specific belief is more important than general belief. Importantly, social capital is a multidimensional, multidisciplinary concept and, subject to the purpose, approach and investigational method, where researchers determine the inherent contents of social capital differently. Future researchers can apply the findings from this study to explore the role of social capital in various areas such as the role of social capital in graduate job searches, how to improve student income and satisfaction level at the workplace after graduation, etc. In addition, further research may aim to provide solutions to enhance the social and behavioral satisfaction of students after graduation.

Acknowledgment

The authors gratefully acknowledge Ho Chi Minh City Open University (HCMCOU), 97 Vo Van Tan Street, Ward 6, District 3, Ho Chi Minh City, 72407, Vietnam for their supporting this research.

References

[1] L. J. Hanifan, "The rural school community center," The Annals of the American Academy of Political and Social Science, vol. 67, no. 1, pp. 130-138, 1916.

[2] P. Bourdieu, "The forms of capital," Cultural theory: An anthology, pp. 81-93, 2011.

[3] R. Leonard and J. Onyx, "Networking through loose and strong ties: An Australian qualitative study," Voluntas: International Journal of Voluntary and Nonprofit Organizations, vol. 14, no. 2, pp. 189-203, 2003.

[4] M. Paldam, "Social capital: one or many? Definition and measurement," Journal of Economic Surveys, vol. 14, no. 5, pp. 629-653, 2000.

[5] F. Fukuyama, Trust: The social virtues and the creation of prosperity (no. D10 301 c. 1/c. 2). JSTOR, 1995.

[6] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," Academy of management review, vol. 20, no. 3, pp. 709-734, 1995.

[7] J. F. Helliwell and R. D. Putnam, "Economic Growth and Social Capital in Italy," Eastern economic journal, vol. 21, no. 3, pp. 295-307, 1995.

[8] R. D. Putnam, "Bowling alone: America’s declining social capital," in Culture and politics: Springer, 2000, pp. 223-234.

[9] C. L. Casey, "Linking Social Capital and Indirect Policy Tools: Fostering Equitable Community Reinvestment Responses?," Journal of Planning Education and Research, vol. 28, no. 4, pp. 413-425, 2009.

[10] S. Mahuteau, M. Piracha, M. Tani, and M. V. Lucero, "Immigration policy and entrepreneurship," International Migration, vol. 52, no. 2, pp. 53-65, 2014.

[11] B. Lancee, Immigrant performance in the labour market: Bonding and bridging social capital. Amsterdam University Press, 2012.

[12] M. S. Granovetter, "The Strength of Weak Ties," American journal of sociology, pp. 1360-1380, 1973.

[13] W. Poortinga, "Social relations or social capital? Individual and community health effects of bonding social capital," Social Science & Medicine, vol. 63, no. 1, pp. 255-270, 2006.

[14] S. Szreter and M. Woolcock, "Health by association? Social capital, social theory, and the political economy of public health," International Journal of Epidemiology, vol. 33, no. 4, pp. 650-667, 2004.

[15] P. Dasgupta, "Trust and cooperation among economic agents," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, no. 1533, pp. 3301-3309, 2009.

[16] T. Harpham, "The measurement of community social capital through surveys," in Social capital and health: Springer, 2008, pp. 51-62.

[17] L. Xue, "Social capital and employment entry of recent immigrants to Canada," Research and evaluation paper). Ottawa: Citizen and Immigration Canada, 2008.

[18] A. M. Kanas, The economic performance of immigrants: the role of human and social capital. Utrecht University, 2011.

[19] P. B. Sørenson, "Public finance solutions to the European unemployment problem?," Economic Policy, vol. 12, no. 25, pp. 222-264, 1997.

[20] A. Daghouri, K. Mansouri, and M. Qbadou, "Information system evaluation based on multi-criteria decision making: A comparison of two sectors," International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, pp. 291-297, 2018.

[21] M. B. Alqaderi, W. Emar, and O. A. Saraereh, "Concentrated solar power site suitability using GIS-MCDM technique taken UAE as a case study," International Journal of Advanced Computer Science and Applications, vol. 9, no. 4, pp. 261-268, 2018.

[22] S. Ahriz, A. El Yamami, K. Mansouri, and M. Qbadou, "Cobit 5-based approach for IT project portfolio management: Application to a Moroccan university," International Journal of Advanced Computer Science and Applications, vol. 9, no. 4, pp. 88-95, 2018.

[23] M. A. Marhraoui and A. El Manouar, "An AHP Model towards an Agile Enterprise," International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, pp. 151-156, Nov 2017.

[24] S. K. Sehra, Y. S. Brar, and N. Kaur, "Applications of Multi-criteria Decision Making in Software Engineering," International Journal of Advanced Computer Science and Applications, vol. 7, no. 7, pp. 472-477, Jul 2016.

[25] A. El Yamami, S. Ahriz, K. Mansouri, M. Qbadou, and E. Illoussamen, "Developing an Assessment Tool of ITIL Implementation in Small Scale Environments," International Journal of Advanced Computer Science and Applications, vol. 8, no. 9, pp. 183-190, Sep 2017.

[26] A. A. Malik, T. M. R. Khan, and A. Mahboob, "Evaluation of OLSR Protocol Implementations using Analytical Hierarchical Process (AHP)," International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, pp. 338-344, Nov 2016.

[27] H. Setiawan, J. E. Istiyanto, R. Wardoyo, and P. Santoso, "The Group Decision Support System to Evaluate the ICT Project Performance Using the Hybrid Method of AHP, TOPSIS and Copeland Score," International Journal of Advanced Computer Science and Applications, vol. 7, no. 4, pp. 334-341, Apr 2016.

[28] A. R. Asghar, A. Tabassum, S. N. Bhatti, and S. A. A. Shah, "The Impact of Analytical Assessment of Requirements Prioritization Models: An Empirical Study," International Journal of Advanced Computer Science and Applications, vol. 8, no. 2, pp. 303-313, Feb 2017.

[29] A. I. El-Dsouky, H. A. Ali, and R. S. Rashed, "Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process," International Journal of Advanced Computer Science and Applications, vol. 7, no. 2, pp. 164-173, Feb 2016.

[30] Juhartini and M. Suyanto, "The Use of Programming Languages on the Final Project Report by Using Analytical Hierarchy Process (AHP) (Case Study on the Student of Information Management Amikom Mataram)," International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 310-317, Sep 2015.

[31] D. U. Ozsahin et al., "Evaluating Cancer Treatment Alternatives using Fuzzy PROMETHEE Method," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 177-182, Oct 2017.

[32] D. U. Ozsahin, B. Uzun, M. S. Musa, and I. Ozsahin, "Evaluating X-Ray based Medical Imaging Devices with Fuzzy Preference Ranking Organization Method for Enrichment Evaluations," International Journal of Advanced Computer Science and Applications, vol. 9, no. 3, pp. 7-10, Mar 2018.

[33] S. Basaran and O. J. Aduradola, "A Multi-Criteria Decision Making to Rank Android based Mobile Applications for Mathematics," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 99-107, Jul 2018.

[34] K. Gulzar, J. Sang, A. A. Memon, M. Ramzan, X. F. Xia, and H. Xiang, "A Practical Approach for Evaluating and Prioritizing Situational Factors in Global Software Project Development," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 181-190, Jul 2018.

[35] P. T. Nguyen, T. A. Nguyen, Q. L. H. T. T. Nguyen, V. D. B. Huynh, and K. D. Vo, "Ranking project success criteria in power engineering companies using fuzzy decision-making method," International Journal of Advanced and Applied Sciences, vol. 5, no. 8, pp. 91-94, 2018.

[36] N. L. H. T. T. Quyen, P. T. Nguyen, and V. D. B. Huynh, "Prioritization of social capital indicators using extent analysis method," International Journal of Advanced and Applied Sciences, vol. 4, no. 10, pp. 54-57, 2017.

[37] N. T. Phong and N. L. H. T. T. Quyen, "Application fuzzy multi-attribute decision analysis method to prioritize project success criteria," AIP Conference Proceedings, vol. 1903, no. 1, p. 070011, 2017.

[38] K. Gulzar, J. Sang, A. A. Memon, M. Ramzan, X. Xia, and H. Xiang, "A practical approach for evaluating and prioritizing situational factors in global software project development," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 181-190, 2018.

[39] A. A. A. Gad-Elrab, T. A. A. Alzohairy, and A. S. Alsharkawy, "Cluster-Based Context-Aware Routing Protocol for Mobile Environments," International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, pp. 1-10, Jan 2015.

[40] J. J. Buckley, "Fuzzy hierarchical analysis," Fuzzy Sets and Systems, vol. 17, no. 3, pp. 233-247, 1985/12/01 1985.

[41] T.-Y. Hsieh, S.-T. Lu, and G.-H. Tzeng, "Fuzzy MCDM approach for planning and design tenders selection in public office buildings," International Journal of Project Management, vol. 22, no. 7, pp. 573-584, 10// 2004.

[42] K. Vahdat, N. J. Smith, and G. G. Amiri, "Fuzzy multicriteria for developing a risk management system in seismically prone areas," Socio-Economic Planning Sciences, vol. 48, no. 4, pp. 235-248, 2014.

[43] P.V. Nguyen, P. T. Nguyen, Q. L. H. T. T. Nguyen, and V. D. B. Huynh, "Calculating Weights of Social Capital Index Using Analytic Hierarchy Process," International Journal of Economics and Financial Issues, vol. 6, no. 3, 2016.

Position-based Selective Neighbors

Sofian Hamad[1], Taoufik Yeferny[2], Salem Belhaj[3]
Computer Science Department, Northern Border University
Arar Kingdom of Saudi Arabia

***Abstract*—In this paper, we propose a routing protocol, named Position-based Selective Neighbors (PSN), for controlling the Route Request (RREQ) propagation in Mobile Ad-hoc Networks (MANETs). PSN relies on the Residual Energy (RE) and the Link Lifetimes (LLT) factors to select the better end-to-end paths between mobile nodes. The key concept is to consider the RE and the LLT to find the best neighboring nodes to forward the received RREQs. A Simulation has been performed to compare PSN with other pioneering routing protocols. Experimental results show that PSN performs better than its competitors. Indeed, our protocol increases the network life time and reduces the network overhead. Furthermore, it reduces the overhead generated by the redundant RREQ, while maintaining good reachability among the mobile nodes.**

***Keywords*—Mobile Ad-hoc network (MANET); routing protocol; energy aware; link life time; AODV**

I. Introduction

Academics and industry have become increasingly interested in wireless research over the last decade. Wireless access was chosen because it allows free movement. A Mobile Ad-hoc Network (MANET) has proved very interesting in finding ways to improve its operation and performance. A MANET typically consists of interconnected mobile nodes using wireless links that have no access points or permanent infrastructure. Moreover, a lot of work has been performed across the layers of Open Systems Interconnection (OSI) while applying Medium Access Control (MAC). Particularly, many routing algorithms have been proposed to provide end-to-end routes. These can be reliable and robust against the mobility of the nodes.

For example, neighboring nodes in wireless networks share wireless media. In addition to this, the nodes must compete with others in order to gain access to these media (channels). A MAC layer will control such an operation. Basically, the MAC protocol governs the access of wireless devices to shared wireless media. This protocol imposes many time constraints in the process to properly regulate the shared resource and to avoid collisions. These can happen, as illustrated in Figure 1 In such a case, the node A does not know that the node B simultaneously receives data from the node C. As a result, it can start its own transmission, which will cause the collision with node B. The neighboring-node collision and interference, the hidden-node presence as well as the distances between senders and receivers have a significant effect on wireless network performances. MANETs face such a problem, particularly while having a lot of data, controlling packet traffic as well as mobile topology.

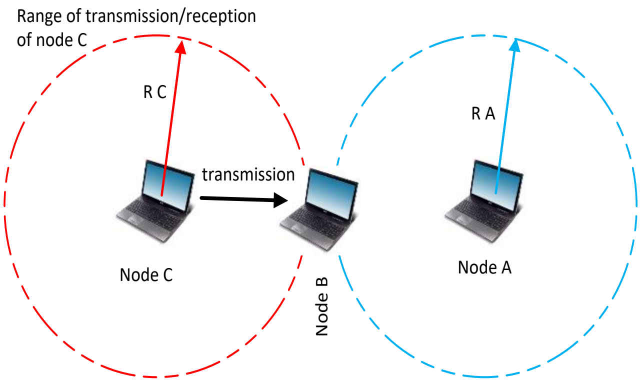


Fig. 1. Instance of Hidden Node Problem.

Because MANET’s topology is highly mobile and the nodes generating data and forwarding entities within networks, designing efficient and robust routing protocols requires a lot of effort. Several routing protocols have been recently put forward for MANETs, whose goal is to establish end-to-end paths in multi-hop scenarios between sink destination nodes and data-generating sources [1-6]. Nodes discover routes to a specific destination within conventional on-demand routing protocols [4-5], through the broadcasting of a Route Request (RREQ) packet. On the reception of a RREQ, the node will check if that packet was previously received. If it is the case, the node will drop it. If it is the contrary, a Route Reply (RREP) will be then sent back to the source node according to the availability of the route. In either case, the RREQ will be rebroadcasted by this node to its immediate neighbors until finding its destination. As a matter of fact, that route-discovery method is called blind flooding. The rebroadcasting of a copy of the received RREQ by each mobile node results, within the global network, in a maximal N – 2 number of rebroadcasts. In such a situation, N in this network is the number of nodes. Thus, there is possibly excessive redundant retransmission, hence high channel contention. This may lead to excessive packet collisions within dense networks. This can be called a broadcast storm problem [7], so it greatly raises end-to-end delay and network communication overhead, whilst rising bandwidth loss [7,9].

A lot of existing approaches have attempted to resolve the flooding problem through the reduction in the number of redundant messages. On the other hand, this results in low coverage and connectivity degree. In fact, the interdependence between both phenomena is problematic for balancing message overhead (in other words, that redundancy level) and coverage [8].

Therefore, the decrease in collisions within the network is able to ameliorate network performances, mainly for MANETs, where nodes may collaborate together so as to connect to those nodes not actually being within their transmission range. In addition to that, broadcasting RREQ messages generate duplicate messages across the full network, looking at the same time for an end-to-end path, hence a big chance of potential collisions. The elimination of the unnecessary RREQ packets is able to decrease the number of packet collisions, which will ameliorate the network performance.

This paper puts forward one novel algorithm which minimizes within the global network the RREQ propagation and simultaneously keeps the network connectivity. The (x, y) coordinates of all nodes and their neighbors in the suggested algorithm are known. According to those positions, the best neighbors are selected by one node to further rebroadcast RREQs. We divide the source node's transmission range to 4 equal zones. The latter are as follows: (Zone1, Zone2, Zone3 and Zone4) in a set M= {M1, M2, M3, M4}. Furthermore, we select 4 neighbor nodes from these zones based on the quality of their links to the source node as well as on their residual energy levels.

Our workout lines one effective routing protocol, which can tackle such a flooding problem and minimize RREQs propagation when at the same time maintaining among nodes within a global network comparable reachability.

II. Related Work

Several approaches have recently proposed manners of decreasing the broadcast-storm effect due to simple flooding [7-10]. Indeed, we can classify these approaches to five categories [8]: the neighbor knowledge methods, the probability-based approach, the position-based methods, simple flooding and different other approaches utilizing various techniques. The simple flooding has been discussed earlier in the introduction

A. *Neighbor Knowledge Methods*

From a first point of view, a main concept for this method is expanding data concerning the node neighbors. According to that, every node sends a neighbor node address or two ones to their neighbors. After that, this node uses existing "hello" messages in the purpose of sending this information periodically. As a result, every node can implicitly know what is in common with others. In the same vein, the writers of [11] suggested the two-hop backward-neighbor information concept. This latter was utilized to minimize the number of forwarding nodes. It also reduced the collisions in the network. Generally, such a suggested mechanism required exchanging one-hop hello messages. A novel joint one-hop neighbor information-based flooding scheme was put forward in [12], consisting of two sub-algorithms: receiver-phase and sender-phase. The sender-phase algorithm would facilitate for the node the selection of one subset of its one-hop neighbors in order to forward flooding messages. It would also select forwarding nodes that could greatly contribute to flooding message dissemination. The writers in [13] put forward an efficient flooding scheme. Indeed, this latter was based on one-hop Information within MANETs. Basically, every node would use its one-hop neighbor data. Looking for one new route, every node would determine a subset of its neighbors as candidate ones for rebroadcasting that message when they received it. Accordingly, the addresses of these nodes were attached to the RREQ message. Once a RREQ was received, the node would search for its address. When the latter was found by the receiving node, the sending one would provide the candidates from a novel subset of its neighbors and would rebroadcast the RREQ. If it was not the case, the node would drop the RREQ.

From a second point of view, neighbor-knowledge methods succeed in the reduction of extra RREQs in the network. On the other hand, the addresses of all the neighboring nodes are carried by periodic hello messages, hence the use of the available bandwidth, which might rise up the overhead. In addition, because of the nodes mobility, the gathered two-hop or one-hop data are not all the time exact.

B. *Position-based Methods*

It is worth mentioning that area-based methods comprise location-based and distance-based schemes. These methods show the area that can be offered by one node when rebroadcasting the same received message. In fact, in the transmission range of that node, a great distance from a previous broadcasting node will result in an additional coverage to be acquired, hence a big chance to reach more nodes. Actually, the writers put forward in [14] an approach which is known as Flooding based on One-hop Neighbor Information and Adaptive Holding (FONIAH).

The authors could assume that nodes knew their geographical location. Added to that, the sharing positions among the nodes would require that every node sent hello messages continuously having location information. One main idea of FONIAH is the node's ability to select those furthest nodes within its transmission range. Afterwards, it could calculate the distance (Maximum Distance (Dmax)) between these furthest nodes and itself .Such a distance was utilized for calculating waiting time at that receiver node. Abolhasan and Wysocki suggested in [15] Position-based Selective Flooding (PSF), where one novel scheme was applied to select forwarding nodes. Mainly, a received RREQ Would be rebroadcast by the receiving nodes just as it enters the Forwarding Region (FR), as it is illustrated in Fig. 2. That was a good position from which neighbors could rebroadcast RREQs, and therefore there would be a strong signal will probably be and a great coverage area. However, such a technique might not find the requested destination for the reason that the destination node was opposing the forwarder.

Hamad et al. suggested in [16, 27] a new algorithm for the reduction of overhead generated by redundant RREQ messages. Part of their work, candidate Neighbors rebroadcasting a RREQ (CNRR) would divide the transmission ranges. The latter were done for nodes sending or rebroadcasting those RREQs into four equal zones (Zone1, Zone2, Zone3 and Zone4). So after that, a node per zone would be selected in the aim of rebroadcasting the RREQ. This selection was on the basis of distance between a node and its neighbors.

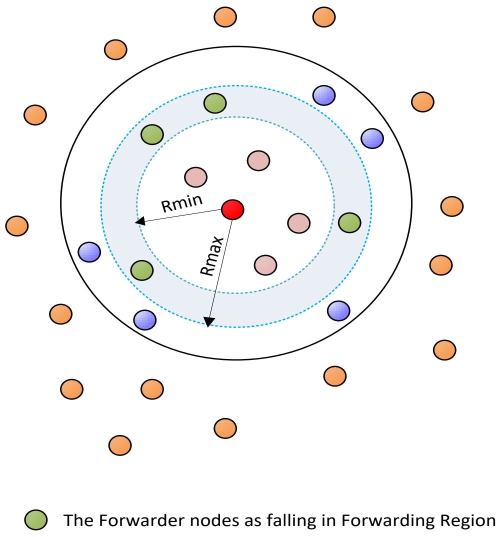


Fig. 2. Illustration of FR in PSF Method.

C. Probability-based Approaches

They are contingent with assigning various node-participation probabilities within a network. These probabilities are signs to nodes for discarding or rebroadcasting a received RREQ. Their values are able to differ for multiple algorithms and node conditions. Yassein et al. suggested in [17] a new probabilistic flooding algorithm that can build up the threshold value for a node having many neighbors. As a consequence, this node cannot rebroadcast the received RREQ. On the other hand, this node may rebroadcast the received RREQ in condition of having a low number of neighbor nodes. Nourazar et al. proposed in [18] a Dynamic Adjusted Probabilistic Flooding (DAPF) Algorithm. Its main goal was to rebroadcast the probability function one message dynamically adjustable with local observations and passing time. We can cite for instance the number of received duplicate messages and network density. Kim et al. suggested in [19] one dynamic probabilistic broadcasting approach. This latter was composed of two (probabilistic and position-based) methods. The probability here was assigned to nodes upon the basis of their distances from a RREQ sender. As a result, in case the receiver node was near that sender node, it might be difficult to rebroadcast the RREQ. Otherwise, it would be more probable to rebroadcast the RREQ and to achieve a wider coverage area.

D. Other Approaches

As it seems to be, various other approaches have been also considered by the research community in the objective of tackling the broadcast storm problem. For example, both studies of [20-21] considered node speed necessary to rebroadcast the RREQs.

Khamayseh et al. suggested in [20] two approaches for the enhancement of the route discovery phase and for the increase in overall routing performance. In addition, the authors considered node speed necessary to participate at the route discovery phase. Both approaches were Aggregate-AODV (Agg-AODV) and Per-Hop Mobility Aware (PH-MA-AODV), where the node would keep track of its speed.

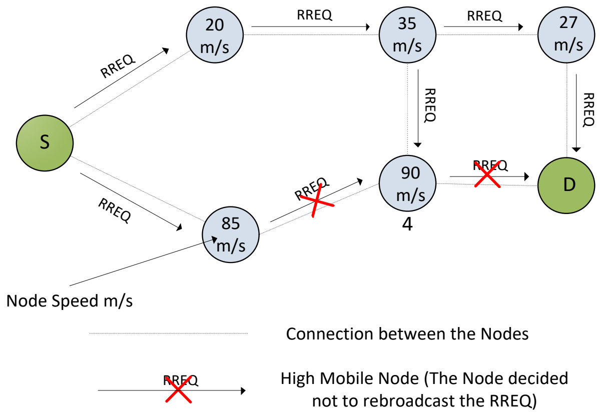

RREQ
RREQ
RREQ
RREQ
RREQ
RREQ
RREQ
RREQ
20 m/s
35 m/s
27 m/s
S
D
90 m/s
4
85 m/s
Node Speed m/s
Connection between the Nodes
RREQ
High Mobile Node (The Node decided not to rebroadcast the RREQ)

Fig. 3. RREQ Propagation in PH-MA-AODV Method.

Firstly, in case a RREQ is received, the node decides if it will forward that RREQ based on its speed. So if the latter is high, the received RREQ will be discarded. In case it is low, the node will take the decision of participating in the route and forwarding, in a way or another, the received RREQ. These nodes are illustrated in Fig. 3. Their speed is greater than 80 m/s and as a result, those received RREQs will be discarded. Secondly, the node attaches its speed. Afterwards, it will forward the received RREQ. Actually, the selection of the best route into the source node will be done by the destination node on the basis of the nodes' low aggregate speed.

III. Proposed PSN Protocol

We will discuss in this part the PSN routing protocol. In view of fact, AODV as an on-demand routing protocol follows blind flooding in the purpose of disseminating route discovery packets in global networks. This blind flooding can work well when reachability very significant. Nevertheless, since end-to-end route selection is carried out by these protocols utilizing hop counts, non-stable paths are able to return because of the MANET extremely mobile environment. To deal with such a problem, two solutions were proposed in [27]. The authors outlined a mechanism of placing into different zones sending/forwarding-node neighbors. The writers in [28] considered link stability and looked in an explicit manner into neighboring nodes' residual battery energy and quality of links. Such suggested protocols would decrease to a minimum the network-wide RREQ dissemination and at the same time preserve the desired connectivity. On the other hand, the previously proposed mechanisms had isolation problems. Let us take as an example the CNRR protocol. This latter just considered the locations of the neighboring nodes. In that way, the RREQ forwarding decisions were solely based upon distance. Despite the fact that RREQ dissemination considerably was reduced by that method, the energies of the remaining nodes and link quality were ignored. Consequently, these returned routes might not be stable for long. In the same context, during the route discovery phase, Link Stability and Energy Aware LSEA [28] protocols take into account nodes' residual energies and link quality. In fact, such a method returns stable paths, thus leading to a high throughput and fewer delays. Yet, because such a method does not give careful consideration to the positions of nodes when disseminating RREQs, we can compromise connectivity.

A. PSN Route Discovery Mechanism

In the following, there are three phases of the route discovery process of the proposed PSN protocol:

First of all, the neighbors of the ‘S’ node are divided ito 4 zones, so as to send an RREQ, which is in a precise manner the same mechanism in [27]. In addition, every neighbor (x,y) coordinates are made known for nodes by using specialized positioning devices like GPS [22].

Secondly, every one of the ‘S’ nodes will compare in a specific zone its average Link Lifetime (LLT) with the averaged LLT (LLTavg). This latter is got from all nodes’ specific times. These nodes share the links with the current ‘S’ node. In a similar way, the ‘S’ node will compare in the specific zone all its neighbors’ residual energies with (REavg).

Thirdly, a Candidate Node (CN) will be selected by the ‘S’ node among neighbors, based on specific conditions, in a specific zone in order to forward a current RREQ. This selection will be based on two conditions. Firstly, in condition that the neighboring nodes’ LLTs and REs are higher than LLTavg and REavg, in that case this current node will be selected as a Potential Candidate Neighbor (PCN) and added to Potential Candidate List (PCL)of the ‘S’ node. This is similar for PCL. Secondly, the ‘S’ node will select CNs from an already existing PCL on the basis of their LLTs and REs. As a matter of fact, a node will be selected from a PCL set when having a great number of LLTs and Res, compared to other PCL nodes. On the other hand, if in the specific zone there is one ‘S’ node neighboring node having the ability of meeting LLTavg and REavg conditions for PCLs, then such a node having in the specific zone most LLTs and REs will be selected as a CN. This will be similar for all the zones.

As an example, let consider the MANET topology, as it is depicted in Fig. 4, where node X has the intention of sending a RREQ to its neighbors. Firstly, that node will divide its transmission range to four zones. It can be assumed that node X compares all its neighbors’ LLTs and REs with LLTavg and REavg in Zone1. On the basis of the checks, nodes A, B and C will be in fact selected as Potential Candidate Neighbors, hence putting them within the PCL list. As a consequence, that node X PCL in Zone1 = {A, B and C}. After that, the same node will select the best node in order to forward the RREQ while comparing LLTs and REs. Added to that, consider (LLTC and REC) > (LLTB and REB) > (LLTA and REA). In that case, that node X will select node C, which will be considered as its CN in Zone1. In the same way, CNs will be selected within the other three zones. This will be done following the previously discussed procedure. Moreover, node X may have the capability of attaching all selected CNs addresses within RREQ packet respective zones. To make it clear, all nodes X zone neighbors will check whether their addresses are part of the address list upon receiving the RREQ packet. When they make sure their addresses are in the list, they can in fact forward the RREQ to their neighbors, This is done according to the PNS procedure. If it is not the case, the others will simply drop it.

In the objective of understanding the PSN route discovery mechanism concept, let us take into account Fig. 4 and Fig. 5, which, for simplicity, present just Zone1. In addition to that, The ‘S’ node interactions with all its neighbors can be seen within the specific zone. As a matter of fact, $L_{s-A}, L_{s-B}, L_{s-D}, L_{s-E}$ *and* L_{s-G} are respectively the neighboring (A, B, C, D, E and G) nodes links with the S’ node. It should be noted that every node LLT is shown above the link, On the other hand, REs are below each individual node. As a result, every node knows its neighbors LLTs and REs.

It was suggested by the authors of [28, 29] every node knew all its neighbor nodes LLTs and REs by exchanging ‘hello’ messages. In a similar way, the ‘hello’ message, in the proposed PSN protocol, was modified to convey to all the current node neighbors its (x, y) coordinates and REs. Indeed, such a frequent exchange of ‘hello’ messages would certainly help every node to get new data concerning its neighbor’s residual energy and link quality.

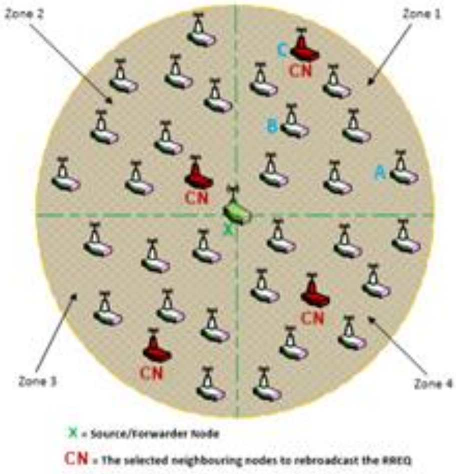

Zone 2
Zone 1
C
CN
B
A
CN
X
CN
CN
Zone 3
Zone 4
X = Source/Forwarder Node
CN = The selected neighbouring nodes to rebroadcast the RREQ

Fig. 4. Instance of Dividing Transmission Range into four Zones and Selecting CNs in Each Zone.

Fig. 5 shows the ‘S’ node has the intention of sending a RREQ packet to its neighbors. According to that, after computing (neighboring nodes) LLTavg as well as REavg, there will be a comparison of these values with the LLT and RE values of every by the ‘S’ node in the target of discovering which nodes have LLTs and REs higher than that of LLTavg and REavg. In particular, just the A, E and F nodes are included within the PCL.

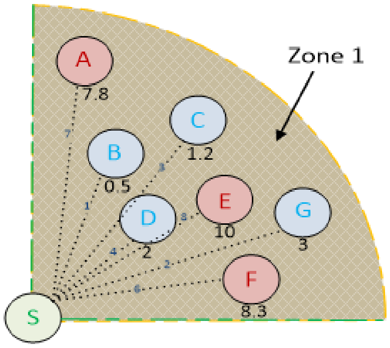


Fig. 5. Instance of Selecting best CN from PCL List in Specific Zone.

From another point of view, the B, C, D and G nodes in Zone1will be left off the PCL due to the fact that LLTavg and REavg are higher than their LLTs and / or REs. Added to that, Fig.5 indicates that node E, among the PCNs within the PCL, is the best candidate to be selected like a CN, which the case for node E, based on its good LLT and RE. That will be repeated by the 'S' node for all the zones in a way that one node is selected in every zone like a CN.

Finally, as a last phase, the 'S' node will include all CN addresses and broadcast them. A similar RREQ will be received by all the zones nodes. Every time that these addresses are seen within an address list, the current RREQ will be rebroadcasted according to the aforementioned method. The other neighboring nodes will only drop that RREQ.

Table 1 shows that Algorithm 1 selects four CNs in the aim of forwarding the RREQ as follows. First of all, one full area around the 'S' node will be split up into four separate zones. The latter are symbolized by the M = {M1, M2, M3, M4} set. As a matter of fact, a set of nodes inside every zone is represented by each member of set M. That is to say, M1 = $\{n_{1-\mathrm{M1}}, n_{2-\mathrm{M1}}, \ldots, n_{|\mathrm{M1}|-\mathrm{M1}}\}$, M2= $\{n_{1-\mathrm{M2}}, n_{2-\mathrm{M2}}, \ldots, n_{|\mathrm{M2}|-\mathrm{M2}}\}$, M3=$\{n_{1-\mathrm{M3}}, n_{2-\mathrm{M3}}, \ldots, n_{|\mathrm{M3}|-\mathrm{M3}}\}$, M4=$\{n_{1-\mathrm{M4}}, n_{2-\mathrm{M4}}, \ldots, n_{|\mathrm{M4}|-\mathrm{M4}}\}$. Next, it iterates through every node of the specific zone and selects the PCL set and therefore the CN in that zone. In the end, the 'S' node sends to chosen candidate nodes the RREQ packet.

TABLE I. PSN ALGORITHM

```
Algorithm 1
Input: Set of nodes N= {n1, n2, n3, ….,n|N|}
∈ the transmission range of sender S.
Output: Selection of four CNs to transmit the RREQ.
// Divides the nodes in the transmission range of 'S' into
four zones represented by M={M1, M2, M3, M4}, where each represents
a set of nodes in their respective zones.
for i= 1 to |N|
    If n [i]x ≥ Sx & n [i]y ≥ Sy
        n[i] ∈ M1
    else if n [i]x < Sx & n [i]y ≥ Sy
        n[i] ∈ M2
    else if n [i]x ≤ Sx & n [i]y> Sy
        n[i] ∈ M3
    else if n [i]x ≤ Sx & n [i]y< Sy
        n[i] ∈ M4
    end if
next i
// Selects the PCL and the four CNs from the PCL in each zone
$M_n$ ∈ M.
for j= 1 to 4
    for k= 1 to |Mj|
        //Node 'S' selects the PCL and CN in Zone Mj
        if ( LLTk ≥ $LLT_{Avg}$) and ( REk ≥ $RE_{Avg}$ )
            PCL_Mj = PCL_Mj U {nk-Mj}
        else
            next k
        end if
    Select CN in the PCL Mj based on the maximum RE and LLT in the set
    PCL_Mj .
next j
```

B. Percentage of RREQ Reception by Neighbour Nodes

According to what has been discussed in [7], which we can get a 61% higher coverage area across a full network offered by rebroadcasting RREQs [7]. PSN will offer more betterment and enhancement with an algorithm which will also help CNs check for optimized RREQ dissemination. As an example, if Algorithm 1 is chosen to be run by any sender/forwarder 'S' node, as represented and provided by Table 1, four CNs will be then selected among its neighbors. In addition, the 'S' node will attach the addresses of the selected CN to the RREQ packet and after that will broadcast it. Only the attached CNs will be permitted for further processing the received RREQ. This will happen if they find their addresses in a RREQ altered.

The verification of distances between every RREQ neighbor and the 'S' sender will result in checking how many of their neighbors got a similar one. If the transmission range of 'S' is more than the distance, the CN will assume that the neighbor obtained a similar RREQ as itself. Thus, any CN will be able to get the percentage of how many neighbors got a similar RREQ. Through extensive simulation, it is basically observed that the percentage which will improve a network performance is 75%.Hence, it is clear that when more than 75% of CN neighbors obtained a similar RREQ, the CN must not rebroadcast such an obtained one as most of its neighbors got it, so it will not be necessary to rebroadcast it. When lower than 75% of CN neighbors get a similar RREQ, it will be rebroadcasted by the CN. Fig. 6 illustrates the overhead / network link in case that CNs have a predefined percentage of what concerns the rebroadcasting of received RREQs.

As a matter of fact, the results presented in Fig. 6 demonstrate what follows: If the percentage is low, overhead will be as well low and vice versa. In other words, if fewer CN neighbors obtain a similar RREQ, the CN node will rebroadcast that obtained RREQ, thus the addition of more overhead to the network. Actually, in case this percentage is low, most CNs may keep such a RREQ. As a consequence, to find the intended destination will be improbable since few nodes will get the RREQ. From that reason, the balance between reachability and overhead added in the network is struck through the means of setting the percentage at 75%.

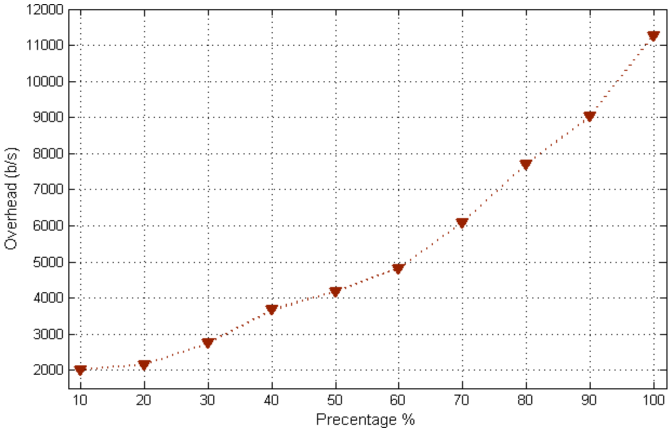


Fig. 6. CN Rebroadcasting Effect.

IV. PSN Performance Evaluation and Results Analysis

The PSN protocol was implemented in the NS2 modeler [23], version 2.34. NS2 is a discrete event network simulator tool used to a great extent when simulating real network scenarios. Added to that, it is freely available and was in the first place designed to simulate wired networks. On the other hand, it has been extended for the simulation of wireless networks including MANETs, wireless LANs and wireless sensor networks. Moreover, it can be organized as it is stated by the OSI reference model [24]. It was shown in [25] that 57% of all published papers based on simulation- utilized NS2 as their simulation tool. This confirms and demonstrates that NS2 is a network simulator which is powerful and trusted.

A. *Simulation Environment and Parameters*

The suggested PSN protocol is exhaustively analyzed through its comparison with our previous proposed A-LSEA and C-CNRR schemes, while depicting its performance. The following section will discuss in detail the results got after comparing between AODV, C-CNRR, A-LSEA and PSN through the use of the parameters given in Table 2. For the simulation of mobile nodes random way points are utilized, where every node will randomly move at a consistent [5 – 30 m/s] speed. At the same time that any node attains one definite random destination, it will take a pause of only two seconds. Afterwards, it will start moving again to a new random destination.

TABLE II. Simulation Parameters for Comparing AODV, C-CNRR, A-LSEA and PSN

Simulation area	600 x 600 M2
Nodes number	100
Data rate	2 Mbps
Transmission range	250 m
Mac protocol	802.11
Traffic type	CBR
Packet size	1000 bits
Traffic	5 packets/sec
Simulation time	600 sec
Speed	[5 m/s - 30 m/s]

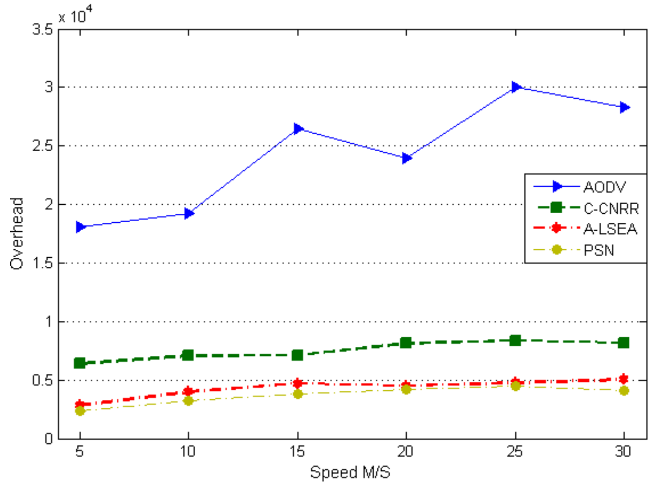


Fig. 7. Overhead Vs. Speed.

B. *Results and Discussion for the First Simulation*

This sub-section will analyze in detail the obtained results and it will present the comparative discussion.

1) *Total overhead:* Fig. 7 shows our comparison of all the overhead of proposed schemes to the AODV, C-CNRR and A-LSEA overheads. Fig. 7 demonstrates also that the overhead goes up significantly when mobility grows for AODV. By way of contrast, this rise is constant for these suggested PSN, A-LSEA and C-CNRR schemes. This due to the fact that the AODV protocol will flood any obtained RREQ with no constraints That is to say, without any energy level or link quality. Through the comparison of the other three schemes, it is clear that PSN outperforms A-LSEA and C-CNRR, as long as the PSN will consider LLTAVG and REAVG and will select as well a specific set of nodes (CNs) in the aim of rebroadcasting a RREQ. In addition to that, the PSN routing protocol will reduce to the least possible the overhead via the driving of the CNs to verify the exact number of their own neighbors obtaining a similar RREQ before sending it. On the other hand, C-CNRR considers only the distance,. In spite of that, A-LSEA considers both constraints. This is actually done without a zoning concept or even an extra verification of the number of neighbors receiving the same RREQ.

2) *Sent and received RREQs:* In the entire network, the number of sent and received RREQs is illustrated in Fig .8. Generally, a broadcast RREQ is sent by one node and afterwards all its neighbors receive it. As a matter of fact, there is a correlation between the number of sent RREQs and the number of received ones (high – high or low -low). The PSN outperforms all other protocols due to the fact that the suggested algorithm selects CNs on the basis of link quality and energy levels as well as the basis of how many node neighbors obtaining one RREQ. When a definite or specific number of ‘S’ node neighbors receive one RREQ, the latter will not be flooded within the network. Consequently, there will be across the entire network more control over RREQ dissemination. In the same way, A-LSEA performs better than C-CNRR due to the fact that A-LSEA path selection is more constant (in case that RE and LLT are considered) compared with C-CNRR (considering just the distances between nodes).

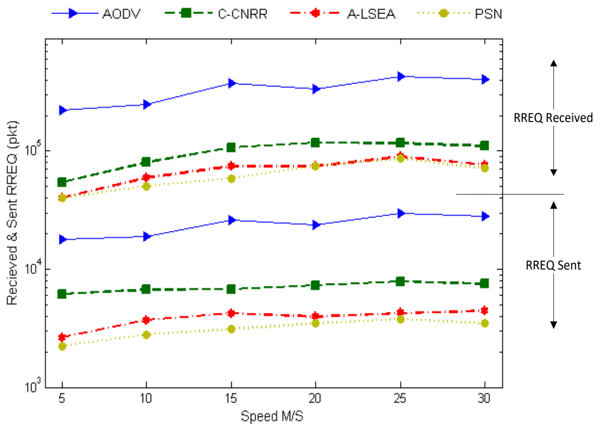


Fig. 8. Received and Sent RREQs Vs. Speed.

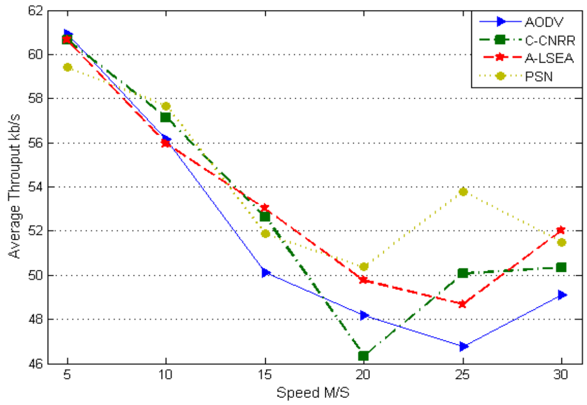


Fig. 9. Throughput Vs. Speed.

3) Average throughput: We demonstrate in Fig. 9 the average PSN routing protocol throughput while comparing it with other routing protocols (A-LSEA, C-CNRR and AODV).

In general, we can see that it decreases when nodes mobility increases for all analyzed protocols. In addition to that, the PSN has a good performance compared to other protocols as the PSN-selected paths holds out more time than the ones selected by other protocols. As a result, the PSN is better than the other protocols (A-LSEA, C-CNRR and AODV) since having the ability to send more data because of very good path lifetimes.

4) Data received: We show in Fig. 10 the data received for PSN as well as the other routing protocols (A-LSEA, C-CNRR and AODV)., for which it is demonstrated that the amount of received data will decrease in case mobility increases. This has an effect on the established routes and links. These latter require being re-established whenever breakages occur. By way of contrast, the amount of received data in the PSN routing protocol will decrease in case the speed rises from 5 m/s to 15 m/s. On the other hand, it stays approximately constant above 15 m/s for the reason that the PSN-algorithm links judge residual energy and link lifetimes. This makes easier to have an impact on high speeds through the involvement of just the nodes selected by the developed algorithm (the selection of a one best node in every zone). This is performed in an end-to-end path.

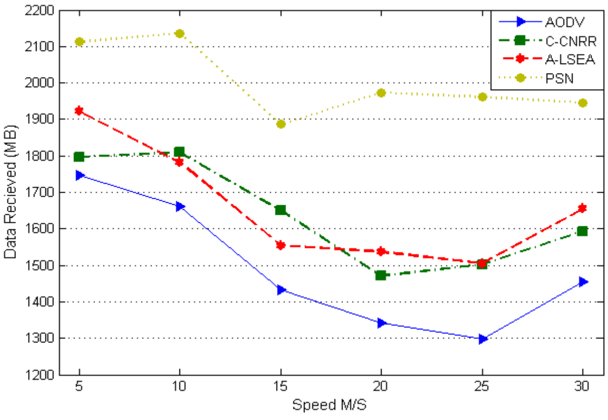


Fig. 10. Data Received vs. Speed.

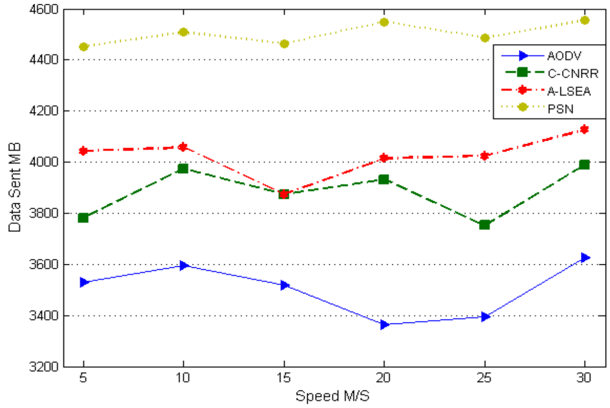


Fig. 11. Data Sent Vs. Speed.

5) Data sent: Fig. 11 depicts the amount of information that has been sent during simulation in a successful manner. The node power supply in MANETs is not permanent because it is naturally mobile. Therefore, any sent or obtained information to and by a node will lead to the reduction in energy levels. In Fig. 8, it can be noticed that AODV is the worst protocol as regards sent or obtained RREQs. A big number of sent or obtained RREQs that are not necessary will greatly decrease the battery life of nodes. Added to that, the PSN is a better protocol when compared to any other routing protocols since it sends a lower number of RREQs, even though it sends more data successfully.

6) Network lifetime: We illustrate in Fig. 12 the proposed-PSN, A-LSEA, C-CNRR and AODV network lifetime results. . We show as well that all other routing protocols are outperformed by the PSN, while giving better Network Lifetime results, due to the fact that the PSN routing protocol will select just 4 nodes for rebroadcasting received RREQs.

In addition to that, an advanced algorithm is run by the selected CN nodes in the goal of eliminating RREQ redundancy by verifying the exact number of their neighbors receiving the same one. According to this, the CN nodes will discard or rebroadcast the obtained RREQs. AS a matter of fact, saving energy will lead to node energy, hence the growth in network lifetime.

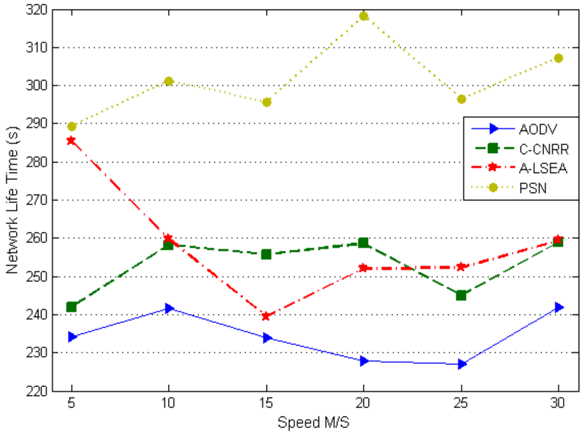


Fig. 12. Network Life Time vs. Speed.

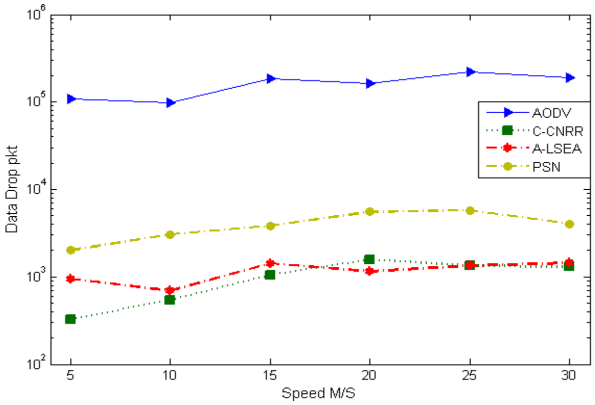


Fig. 13. Data Drop vs. Speed.

7) Data drop: We depict in Fig. 13 (in packets) the amount of data dropping. This is carried out during simulation of proposed-PSN, A-LSEA, C-CNRR and AODV protocols. It is clear that the PSN routing protocol is better than the other routing protocols for these former performance metrics.

On the other hand, Fig. 13 shows that C-CNRR is greatly better than all the other routing protocols in relation to the dropping of data. Despite the fact that the PSN selects better paths compared in fact to the other routing protocols, there will be no advantage of performing better basically as regards any dropped data packets within the network. This can be because the end-to-end-path C-CNRR selection is made on the basis that the distance between the route nodes is advantageous owing to the signal strength for sending and receiving data in distances smaller than those of the routing protocols.

C. Results and Discussion for Second Simulation

For further verification and validation, the authors in [20] implemented Mobility-Aware AODV in NS2. It was also compared with the proposed PSN approach. The latter is used in this section in the goal of achieving a better performance while taking into account our previously introduced (A-LSEA and C-CNRR) routing protocols as well as that standard AODV. For this reason, we consider this approach (PSN) in this paper the best proposed routing protocol. In fact, PSN is selected to be compared to AODV and the work suggested in [20] through the use of similar simulation parameters, as provided by Table 3.

TABLE III. SIMULATION PARAMETERS FOR COMPARING AODV, MA-AODV AND PSN

Simulation area	700 x 700 M2
Nodes number	100
Data rate	2 Mbps
Transmission range	250 m
Mac protocol	802.11
Traffic type	CBR
Trafic number	15 flows
Packet size	1000 bits
Traffic	5 packets/sec
Simulation time	700 sec
Pause time	[2 s – 12 s]

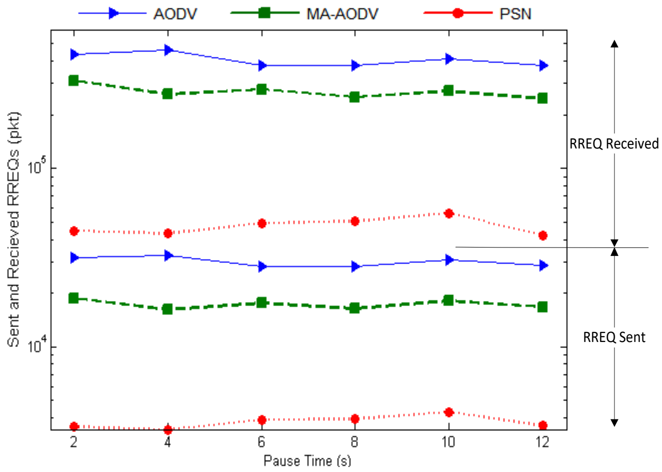


Fig. 14. Sent and Received RREQs vs. Pause Time.

In the next through Fig. 14 – Fig. 20, we will illustrate different metrics [30-32] for the comparisons between PSN, MA-AODV and AODV. It is clear in Fig. 14 that the PSN routing protocol was able to send and receive in the network fewer RREQ packets. This is due to the fact that end-to-end routes are selected by the PSN on the basis of LLT and RE factors

On the other hand routes are selected by MA-AODV only on the basis of node speeds. An edge is given to the PSN over MA-AODV by these two factors for the reason that the route selected by PSN endures for more time than that selected by MA-AODV.

Added to that, those routes selected by MA-AODV endure generally less. Afterwards, the nodes will establish a novel path through the initiation of a new RREQ discovery process. As a consequence, many RREQs will be sent and received in addition to the entire overhead, as depicted in Fig. 15.

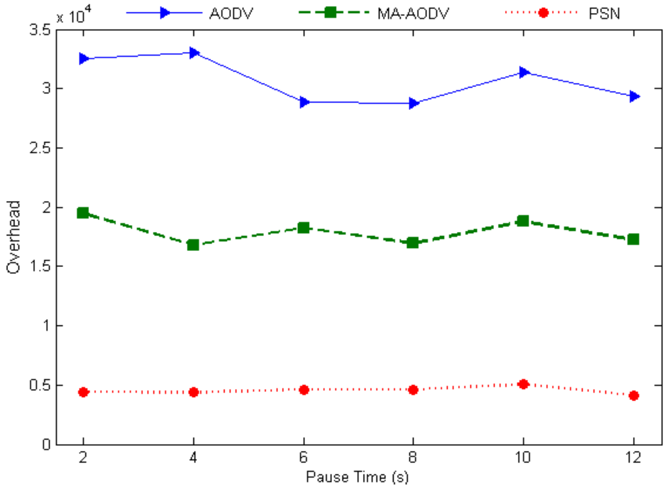


Fig. 15. Average Overhead vs. Pause Time.

The average delivery ratio is illustrated in Fig. 16 for the AODV, MA-AODV and PSN routing protocols. We can see that all routing protocol delivery ratios go up as soon as the pause time increases.

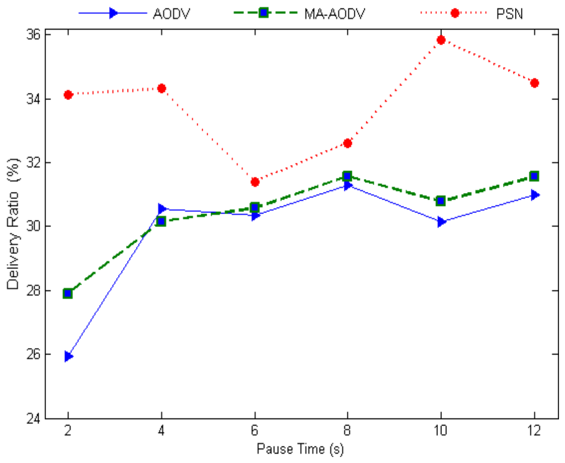


Fig. 16. Delivery Ratio vs. Pause Time.

This is because of the stillness in the mobile nodes. What is more, this PSN routing protocol performs in a good way compared to MA-AODV and AODV routing protocols due to the fact that PSN selects paths lasting longer compared to the ones selected by MA-AODV for the reason that selected PSN end-to-end paths are based on Link Lifeitmes (LLT) and Residual Energy (RE) of the nodes involved in the route. By way of contrast, MA-AODV forwards node-speed-based RREQs. As a result, this MA-AODV algorithm will lead to the end-to-end routes which are entirely having along the path low-speed nodes. On the other hand, this approach, does not ensure any very good paths as regards MANETs because, first of all, , there may be two slow oppositely moving nodes.

As a matter of fact, these two nodes link lifetime can terminate since they move apart. In addition to that, imagine two neighboring nodes that move quite quickly in the same direction. By way of contrast, these two nodes link lifetime is valid for a long time compared to two oppositely moving low-speed nodes. MA-AODV will just consider that node speed in order to forward a received RREQ. At the same time, the PSN will consider both direction and speed and will calculate in fact the any two neighbor nodes link lifetimes. In the second place, the PSN will provide a very strong packet delivery ratio. This is on account of how the nodes Residual Energies are considered by the PSN during the route selection decision.

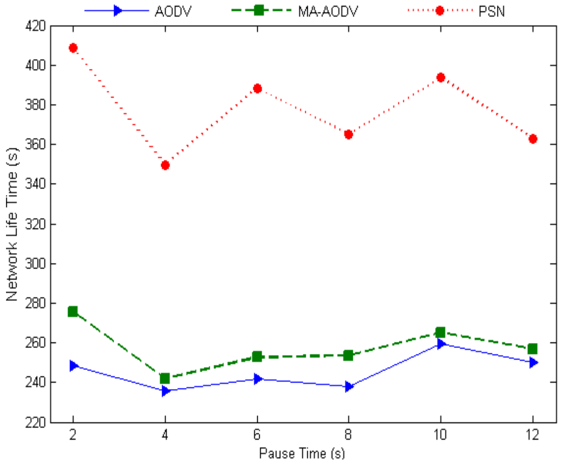


Fig. 17. Network Life Time vs. Pause Time.

Moreover, it is noticeable in Fig. 17 that this suggested PSN runs in a successful way the network for more time compared to the MA-AODV routing protocol. In the first place, the PSN considers the Link Lifetimes and the Residual Energies of the nodes that are in fact involved within end-to-end routes whose role is returning stable paths.

Secondly, energy is conserved by discontinuing sending/receiving RREQ packets which are not necessary and which consume a big amount of the node energy. In the same vein, MA-AODV considers just the nodes speed is the latter cannot be an accurate parameter for the selection of unchanging paths. Running a network much longer will enable nodes to send and receive data lot of information, as depicted in Fig. 18. As a result, the PSN outperforms MA-AODV as it sends/ receives more data packets.

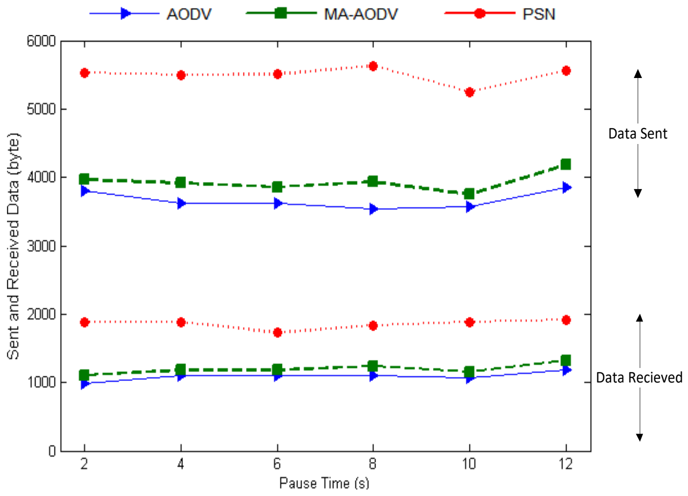


Fig. 18. Sent and Received RREQs vs. Pause Time.

A throughput comparison of AODV, MA-AODV and PSN routing protocols is clearly shown in Fig. 19, where the PSN mostly outperforms MA-AODV on account of that improved algorithm, which improves and stabilizes in a better way end-to-end paths.

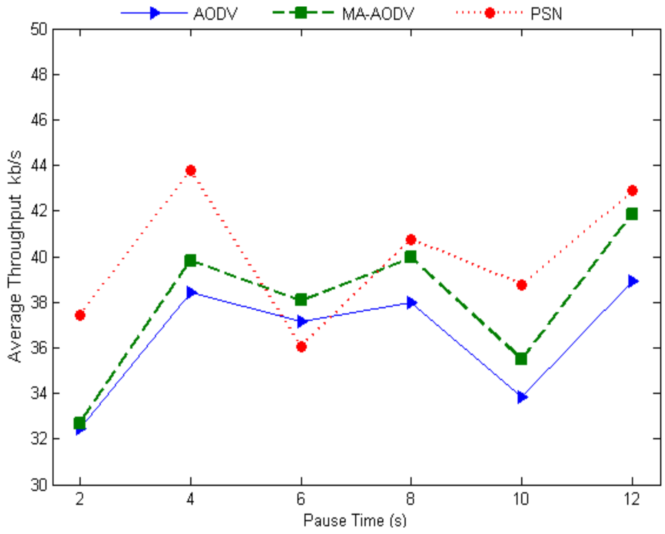


Fig. 19. Average Throughput vs. Pause Time.

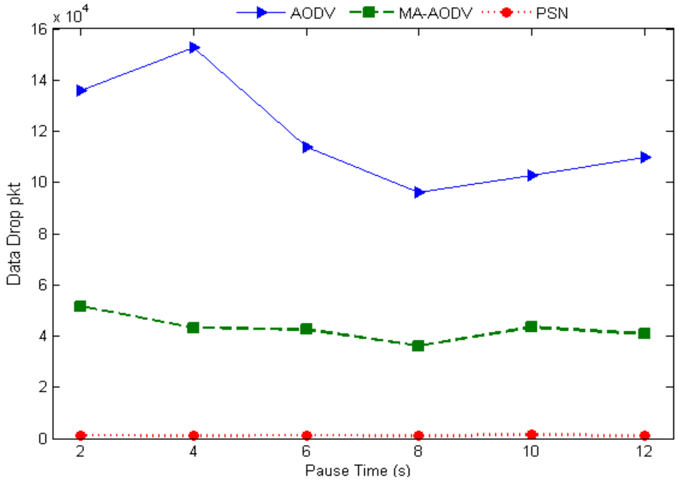


Fig. 20. Data Drop vs. Pause Time.

We can notice as well that the shape of the general throughput rate curve is incremental to the same degree as the pause times increases. This is in spite of some fluctuations caused by nodes mobility randomness. Theoretically speaking, the sex-second-pause-time scenario throughput rate should be greater than a scenario that has a four-second pause time, while having during simulation identical trajectories travelled by nodes. On the other hand, the positions towards which nodes move, under Random Waypoint mobility models, are chosen in a random way and vary from a scenario to another.

Finally, we note that the PSN outperforms the MA-AODV routing protocol as regards the information drop packets, as depicted in Fig. 20, That can be owing to the abovementioned reasons.

V. Conclusion

This paper presents the PSN protocol, a routing protocol for controlling RREQ propagation within networks, which allows end-to-end paths to be selected based on the Residual Energy (RE) and the Link Lifetimes (LLT). PSN benefits from the combination of these two important factors. Moreover, when the CNRR [27] and LSEA [28. 29] concepts are merged, the RREQs dissemination into the network will actually be reduced without causing reachability loss between the nodes. In addition, we introduced a threshold percentage based method, in which the nodes verify that their neighbors have received before rebroadcasting a RREQ. By preventing nodes from sending duplicate RREQs, this mechanism more intelligently controls network-wide flooding, based on a defined threshold relating to the percentage of its neighbors that have received the RREQ. We performed a simulation-based comparison between the proposed PSN and other routing protocols for different metrics and we have discussed the results.

This increases network lifetimes, improves throughput, and enables more data to be sent and received. The proposed scheme combines both the Residual Energy (RE) and Link Lifetime (LLT) factors in the routing management process, rather than using only a single factor, as in the case studies of [20, 26].

Acknowledgment

Financial support for this study was provided by deanship of Scientific Research, Northern Border University under grant no. (7330-SCI-2017-1-8-F).

References

[1] Perkins, C.E. and Bhagwat, P. "Highly Dynamic Destination-Sequenced distance-vector routing (DSDV) for mobile computers", ACMSIGCOMM Computer Communication Review, vol. 24, no. 4, pp. 234-244, 1994.

[2] Qayyum, A., Jacquet, P. and Muhlethaler, P. "Optimized Link State Routing Protocol (OLSR)", 2003.

[3] Ogier, R., Templin, F. and Lewis, M. (2004) Topology dissemination based on reverse-path forwarding (TBRPF), 2004.

[4] Das, S., Perkins, C. and Royer, E. "Ad hoc on demand distance vector (AODV) routing", Mobile Ad-hoc Network (MANET) Working Group, IETF, 2002 .

[5] Johnson, D.B., Maltz, D.A. and Broch, J. "DSR: The dynamic source routing protocol for multi-hop wireless ad hoc networks", Ad hoc networking, vol. 5, pp. 139-172, 2001.

[6] Z. J. Haas, M. R. Pearlman, and P. Samar, "The Zone Routing Protocol (ZRP) for ad hoc networks," IETF Internet Draft, draft-ietf-manet-zonezrp-04.txt., July 2002.

[7] Tseng, Y.C., Ni, S.Y., Chen, Y.S. and Sheu, J.P. "The broadcast storm problem in a mobile ad hoc network", Wireless networks, vol. 8, no. 2, pp. 153-167, 2002.

[8] Abdulai, J., Ould-Khaoua, M. and Mackenzie, L. "Improving probabilistic route discovery in mobile ad hoc networks", Local Computer Networks, 2007.LCN 2007. 32nd IEEE Conference, pp. 739, 2007.

[9] Y.-C. Tseng, S.-Y.Ni, and E.-Y. Shih, "Adaptive approaches to relieving broadcast storms in a wireless multihop mobile ad hoc networks," Proceedings of IEEE Transactions on Computers, vol. 52, pp. 545--557, May 2003.

[10] B. Williams and T. Camp, "Comparison of broadcasting techniques for mobile ad hoc networks," Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing, MOBIHOC, pp. 194 - 205, June 2002.

[11] Le, T.D. and Choo, H. "Efficient flooding scheme based on 2-hop backward information in ad hoc networks", IEEE ICC'08 International Conference on Communications, pp. 2443, 2008.

[12] Yang, S.R., Chiu, C.W. and Yen, W.T. "A novel convex hull-based flooding scheme using 1-hop neighbour information for mobile ad hoc networks", Wireless Networks, vol. 17, no. 7, pp. 1715-1729, 2011.

[13] Liu, H., Wan, P., Jia, X., Liu, X. and Yao, F. "Efficient flooding scheme based on 1-hop information in mobile ad hoc networks", Proc. IEEE INFOCOM, 2006.

[14] Lee, S.H. and Ko, Y.B. "An Efficient Neighbor Knowledge Based Broadcasting for Mobile Ad Hoc Networks", Computational Science–ICCS, pp. 1097-1100, 2006.

[15] Abolhasan, M. and Wysocki, T. "GPS-based route discovery algorithms for on-demand routing protocols in MANETs", Wireless On-Demand Network Systems, pp. 144-157, 2004.

[16] Hamad, S., Noureddine, H., Radhi, N., Shah, I. and Al-Raweshidy, H. "Efficient flooding based on node position for mobile ad hoc network", IEEE Innovations in Information Technology (IIT), pp. 162, 2011.

[17] Yassein, M.B., Khaoua, M.O., Mackenzie, L. and Papanastasiou, S. "Improving the performance of probabilistic flooding in manets", In Proc. of International Workshop on Wireless Ad-hoc Networks (IWWAN), 2005.

[18] Nourazar, F. and Sabaei, M. "DAPF: An Efficient Flooding Algorithm for Mobile Ad-hoc Networks", 2009 International Conference on Signal Processing Systems IEEE, pp. 594, 2009.

[19] Kim J, Zhang Q, Agrawal DP. In: Probabilistic broadcasting based on coverage area and neighbor confirmation in mobile ad hoc networks. Global telecommunications conference workshops, GlobeCom workshops, p. 96-101, 2004.

[20] Khamayseh Y, Darwish OM, Wedian SA. “Ma-aodv: Mobility aware routing protocols for mobile ad hoc networks”, IEEE ICSNC'09 fourth international conference on Systems and networks communications, p. 25-29, 2009.

[21] Hamad S, Noureddine H, Al-Raweshidy H. In: Link stability and energy aware for reactive routing protocol in mobile ad hoc network. Proceedings of the 9th ACM international symposium on mobility management and wireless access; ACM; p. 195-8, 2011.

[22] Kaplan, E.D. and Hegarty, C.J. Understanding GPS: principles and applications, Artech House PublishersK, ISBN 1-58053-894-0, 2006.

[23] "The Network Simulator Ns-2-the VINT pro,"http://www.isi.edu/nsnam/ns/ns-build.html; Retrieved in July 2012.

[24] Lewan, D. and Long, H. "The OSI file service", Proceedings of the IEEE, vol. 71, no. 12, pp. 1414-1419, 1983.

[25] Gast, M. 802.11 wireless networks: the definitive guide, O'Reilly Media, ISBN 0-596-10052-3, 2005.

[26] Xie, F., Du, L., Bai, Y. and Chen, L. "Energy aware reliable routing protocol for mobile ad hoc networks", IEEE Wireless Communications and Networking Conference WCNC, pp. 4313, 2007.

[27] Hamad, Sofian, Salem Belhaj, and Muhana M. Muslam. "Smart Selection of Candidate Neighbors for Efficient Route Discovery in MANETs." Journal of Applied Sciences 17.3 (2017): 126-134.

[28] Hamad, Sofian, Hadi Noureddine, and Hamed Al-Raweshidy. "LSEA: Link stability and energy aware for efficient routing in mobile ad hoc network." Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on. IEEE, 2011.

[29] Hamad, Sofian, Salem Belhaj, and Muhana M. Muslam. "Average Link Stability with Energy-Aware Routing Protocol for MANETs." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 9, no. 1 (2018): 554-562.

[30] T. Yeferny, K. Arour, A. Bouzeghoub. "An efficient peer-to-peer semantic overlay network for learning query routing". 27th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 1025-1032 (2013).

[31] T. Yeferny, K. Arour. "Efficient routing method in p2p systems based upon training knowledge". 26th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 300-305 (2012).

[32] Taoufik Yeferny, Sofian Hamad and Salem Belhaj. “CDP: a Content Discovery Protocol for Mobile P2P Systems". International Journal of Computer Science and Network Security, VOL.18 No.5, May 2018

Comparative Study of Data Sending Methods for XML and JSON Models

Anca-Raluca Breje[1], Robert Győrödi[2], Cornelia Győrödi[3], Doina Zmaranda[4], George Pecherle[5]
Department of Computer Science and Information Technology
University of Oradea Oradea, Romania

Abstract—Data exchange between different devices and applications has become a necessity nowadays. Data is no longer stored locally on the device, but in the cloud. In order to communicate with the cloud and exchange data, web services are being used. To keep the communication consistent across different devices and platforms, the data needs to be formatted using a standard data format, such as JSON or XML. This paper compares both standards and provides an in depth analysis of their performance. In order to perform the analysis a web API was built in the PHP framework Laravel, which was then tested with the help of the API development environment called Postman for different number of transferred items.

Keywords–XML; JSON; data model; data transfer; application programming interface

I. Introduction

Mobile applications have taken a large scale lately and they often require to be connected to a database hosted on a server. Common approaches imply that access to these data to be accomplished via web APIs [1], and consequently, these web APIs must have a very good response time to ensure that there are no delays in displaying the required data to the end-user.

In order to have the data that reaches the applications usable, it needs to be formatted according to a standard that can be parsed, read and used by both the API and the application that will access it. Two of the most popular formatting standards for applications that use web APIs are JSON and XML. Both JSON and XML formats have strengths and drawbacks that qualify them for specific purposes and each of them can be used according to the need of the system [2]. It is well known nowadays that JSON power stays into its simple structure that makes it suitable for simple data transmission [3]. On the other hand, one of the main advantages of XML is represented by its flexibility, given by the possibility of storing (theoretically) all possible data types, unlike JSON where storing is limited to common data types [4]. This flexibility comes with a cost, XML format being much more difficult to parse and to convert to objects, due to its strict structure definition (tree-like), than the much more simplistic JSON format [5].

Another well-known advantage of XML format is represented by the large availability of technologies that could be used for validating XML documents, such as XML namespaces or XMLSchema [6]. Even if for JSON format, in the last years, similar technologies emerged, such as JSON Schema, their range of functions are still not comparable with XML technologies ones [7].

Starting from these two well-known technologies, a performance comparison between JSON and XML data formats, from both runtime and memory usage point of view, is presented in this paper. The paper starts by reviewing the related work, as presented in Section II. A specific testing architecture was specifically developed for running the tests: the algorithm and data structure that are involved are described in Section III. Section IV presents the API developed for accessing the database while Section V illustrates issues related to validation. In Section VI several tests were run, and the obtained results were analysed from several perspectives. Furthermore, Section VII resumes the conclusions of the study.

II. Related Work

Several comparisons based on different scenarios were done between the two formats in [8]. Also, a comprehensive analysis of XML and JSON for web technology is described in [2]. In [9], the process data exchange between a mobile application and remote servers using JSON format is described. A performance comparison between the two interchange formats for simple structures is described in [10]. As outlined in [11], switching between a format to another is possible by using converters, meanwhile preserving data content. Common conclusion that results is that generally, JSON with its simple format behaves faster and uses fewer resources than XML. However, with is much more complex structure and validation techniques, XML remains actual for applications that are manipulating various types of data.

In this idea, this paper presents a benchmark performance comparison between JSON and XML data formats, from both runtime and memory usage point of view, when different types of data were involved. Thus, the performance tests carried on in this paper explore the execution time and memory footprint when using XML and JSON for data sending methods for various applications. Besides other comparative studies existing in the literature, we tried to consider into comparison, for both formats, two additional issues: data validation and data compression. The main goal is to assess the performance impact on the two methods when including data validation and also when compression on the server is enabled.

III. TEST PERFORMANCE ALGORITHM AND DATA

No matter what format is chosen, conversion of data needs using some specific server-side language, such as PHP. Consequently, to be able to analyse the two data sending formats, a web API using the PHP framework Laravel was built [12]. The developed API supports both formats for the four basic operation that are related to data storage, namely data request, sending and inserting data in the database, sending and updating data in the database, deleting data.

Each of the four operations was tested through the developed web API, which for this paper was called using an API development environment called Postman.

These operations were then analysed based on two criteria: data receiving speed and the size of the data received.

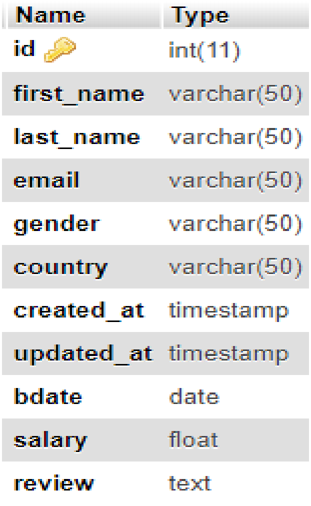
To test the data parsing performance of the two formats, JSON and XML, we used the following generic algorithm

```
var XML / var JSON;
var TIME BEFORE = get current timestamp
validate XML/JSON;
decode XML/JSON;
var TIME AFTER = get current timestamp
display TIME AFTTER - TIME BEFORE
```

This performance test uses the PHP language to execute the parsing of an array with three elements that is encoded in XML and JSON format. The functions used for the test were *json_decode* for the JSON data and *simplexml_load_string* for the XML data.

Fig. 1 illustrates the structure of the database table that is used in the application, containing sample data of some people, and it has the following columns: id, first name, last name, email, gender, country. In order to illustrate the impact of using different types of data for both XML and JSON formats, columns of several other types were included into the table: timestamp, date, float and text.

The table data is handled using SQL queries written in PHP using the Eloquent ORM, which is part of the Laravel framework [12].



Name	Type
id 🔑	int(11)
first_name	varchar(50)
last_name	varchar(50)
email	varchar(50)
gender	varchar(50)
country	varchar(50)
created_at	timestamp
updated_at	timestamp
bdate	date
salary	float
review	text

Fig. 1. Users Table with Different Column Types.

IV. API URLS AND HTTP REQUEST METHODS

The API built for the testing of the two models, is divided in two big parts, the part that expects data and send back data formatted as JSON, and the other part that expects data and sends it back formatted as XML, the URL bases for the two parts are the following [1]:

- http://localhost/api/json
- http://localhost/api/xml

When calling one of the API URLs, the method with which the API is called must be also specified, the available methods are the following [13]

- GET – for requesting data
- POST – for adding new data
- PUT – for updating existing data
- DELETE – for deleting data

Fig. 2 explains the interaction between the request of the API and the data stored in the database.

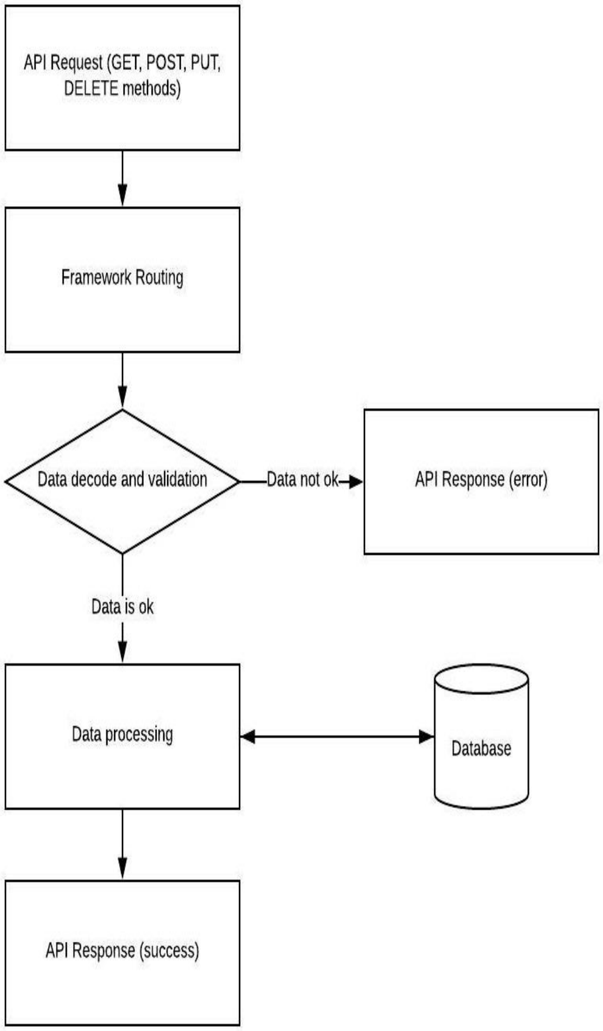


Fig. 2. API Request Diagram.

V. Data Validation

To make sure the data received by the API is correct and it contains all the needed elements, validation schemas were used.

For the JSON format data, we used JSON Schema which is a vocabulary to annotate and validate JSON documents [14]. A library that implements JSON Schema that is compatible with Laravel is *JSON Schema for PHP* [15]. An example of how the name field is defined in the JSON schema, to be validated in the input, is the following

```
[
   "type"=>"array",
       "properties"=>(object)[
          "first_name"=>(object)[
             "type"=>"string"
          ],
          "last_name"=>(object)[
             "type"=>"string"
          ],
          …
        ]
     ]
```

For the XML format data, we used a XSD schema which defines the elements of the correct XML document. A library that implements XSD Schema and is compatible with Laravel is *PHP Xml validator* [16]. An example of how the name field is defined in the XML schema, to be validated in the input, is the following

```
<xs:element     name="record"     minOccurs="0"
maxOccurs="10000">
  <xs:complexType>
    <xs:sequence>
      <xs:element     ref="id"     minOccurs="0"
maxOccurs="1"/>
      <xs:element                 ref="first_name"
minOccurs="0" maxOccurs="1"/>
      <xs:element ref="last_name" minOccurs="0"
maxOccurs="1"/>
…
    </xs:sequence>
  </xs:complexType>
</xs:element>
…
<xs:element                   name="first_name"
type="xs:string"/>
<xs:element name="last_name" type="xs:string"/>
```

VI. Comparison between Transmission of Data for XXL and JSON Models

A. Data Request (GET)

For the GET method, the data is obtained by calling one of the following URLs (the example is using the *cURL* function), where the limit parameter represents how many elements we want to get from the database with the API.

To request data in JSON format

```
curl -X GET
http://localhost/api/json/people?limit=1000
```

To request data in XML format:

```
curl -X GET
http://localhost/api/xml/people?limit=1000
```

In order to obtain the data, the called API web function will execute a database query, which returns a number of rows less than or equal to the value in the limit parameter.

The data returned by the SQL query is then converted using *simple_load_string()* for XML and *json_decode()* PHP functions to the required format and returned as a response to the API call.

Formatting data in the XML format can consume more memory than in JSON format. Moreover, adding attributes to the XML tags contributes to the final data size of the XML result text, to lowering the performance of the applications that use it.

In contrast to the XML format, JSON is more simplistic and easier to use in applications, with lower impact on application performance, especially because JSON is a format based on JavaScript objects, which most programming languages (PHP, JS, C #, etc.), can use without the need of including any external libraries [17].

As can be seen in Fig. 3 and Table 1, the XML format for the same data and for the same number of items takes up more storage space, but not more than 20% more than the JSON format. This also affects the data response time, the impact on this being noticeably higher, being more than three times higher (in the case of 10,000 items).

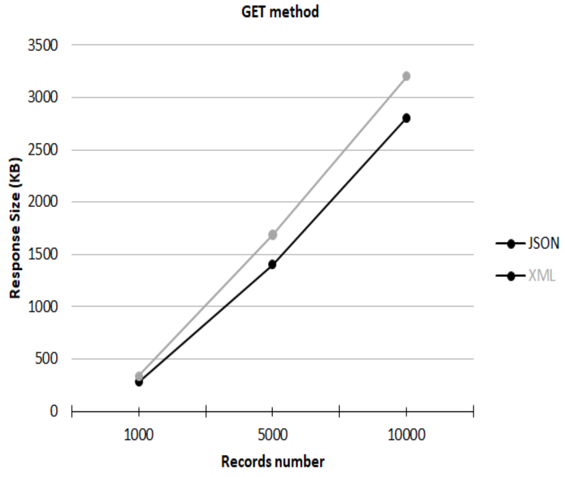


Fig. 3. XML and JSON Graph for Response Size in KB – GET Method without GZIP Compression.

TABLE I. XML and JSON Response Size in KB – GET Method – without GZIP Compression

Records no.	JSON	XML	% XML over JSON
1000	281 KB	337 KB	119
5000	1402 KB	1689 KB	120
10000	2805 KB	3205KB	114

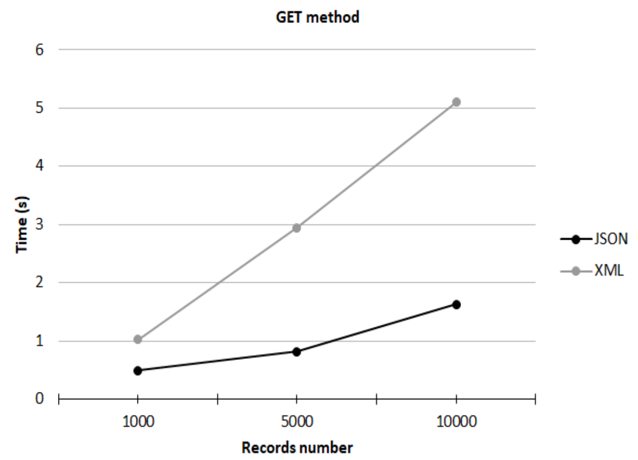


Fig. 4. XML and JSON Graph for Response Time in seconds GET Method without GZIP Compression.

TABLE II. XML AND JSON RESPONSE TIME IN SECONDS GET METHOD WITHOUT GZIP COMPRESSION

Records no.	JSON	XML	% XML over JSON
1000	0.49 s	1.03 s	210
5000	0.82 s	2.95 s	359
10000	1.63 s	5.1 s	312

As presented in Fig. 4 and Table 2, for JSON format, the call response time for 1000 items is 0.49 s, the response size is 281 KB, and for XML format, for the same number of items, the response time is 1.03 s, the response size is 337 KB.

For the GET method, we also tried requesting the data from a server that has the gzip compression activated to see if any if there are benefits in having a server-side compression.

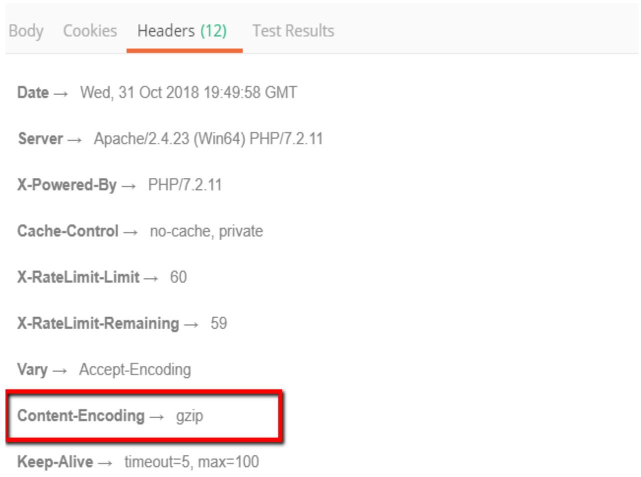
Before using the gzip compression, the deflate module needs to be enabled (on an Apache server) and also the data that should be compressed needs to be specified on the *.htaccess* file of the project or on the virtual host configuration.

In order to have the compression for the JSON and XML format we are using, the following lines were added in the *.htaccess* file of the project:

```
<IfModule mod_deflate.c>
    AddOutputFilterByType                DEFLATE
            application/json application/xml
</IfModule>
```

To see if the gzip compression was used on the data from the response, the header of the response can be checked (Fig. 5), if the Content-Encoding is set to gzip, it means the gzip compression was correctly enabled for the data type sent from the server in the response.

As it is shown in Table 3 and Table 4, the size of the response changed after the gzip was applied, and the response time is better if gzip is enabled, namely for JSON the response size is approximately 5 times smaller and the response time is 25% to 35% faster.



Date → Wed, 31 Oct 2018 19:49:58 GMT

Server → Apache/2.4.23 (Win64) PHP/7.2.11

X-Powered-By → PHP/7.2.11

Cache-Control → no-cache, private

X-RateLimit-Limit → 60

X-RateLimit-Remaining → 59

Vary → Accept-Encoding

Content-Encoding → gzip

Keep-Alive → timeout=5, max=100

Fig. 5. Response Headers after the GZIP Compression is Enabled.

TABLE III. XML AND JSON RESPONSE SIZE IN KB – GET METHOD – WITH GZIP COMPRESSION

Records no.	JSON	XML	% XML over JSON
1000	52.9 KB	73.4 KB	139
5000	263.1 KB	356.8 KB	135
10000	529.3 KB	731.3KB	138

TABLE IV. XML AND JSON RESPONSE TIME IN SECONDS–GET METHOD WITH GZIP ENABLED

Records no.	JSON	XML	% XML over JSON
1000	0.32 s	0.6 s	187
5000	0.61 s	2.21 s	362
10000	1.22 s	4.14 s	339

For XML, the results are similar, the response size is around 4.6 times smaller, the response time is 20% to 40% better. Based on the results, if the server allows the gzip compression, it would be recommended to use it for both JSON and XML data request.

B. Send and Insert Data (POST)

For the POST method, sending and inserting data is done by calling the following URL (*cURL* function), while putting in the body of the call the data that is being sent and is going to be inserted.

To send and insert data in JSON format:

```
curl -X POST http://localhost/api/json/people
```

To send and insert data in XML format:

```
curl -X POST http://localhost/api/xml/people
```

Data sent to the web API is processed using PHP functions and inserted into the database through SQL INSERT queries, and as a result, a success message is returned if everything went well. Because the web API must know that it will receive data in JSON or XML format, we will specify this fact in the header of the call by setting the content type property as it follows:

- Content-type: application/json-for JSON
- Content-type: application/xml-for XML

The data that is sent to the web API, formatted as JSON needs to have the following shape:

```
[   {
            "first_name":"Lucille",
            "last_name":"Baddoe",
           "email":"lbaddoe0@earthlink.net",
            "gender":"Female",
            "country":"China",
            "bdate":"1989-05-12",
            "salary": 7433.55,
            "review":"Lorem ipsum...",
    },
    ... ]
```

As for the XML, data the need to be formatted like the example below:

```
<?xml version="1.0”?>
<root>
  <Collection>
    <row_0>
            <first_name>Elise</first_name>
            <last_name>McGurn</last_name>
            <email>emcgurn0@a8.net</email>
            <gender>Female</gender>
            <country>Philippines</country>
            <bdate>1989-05-12</bdate>
            <salary>7433.55</salary>
            <review>Lorem ipsum...</review>
    </row_0>
            …
    </Collection>
</root>
```

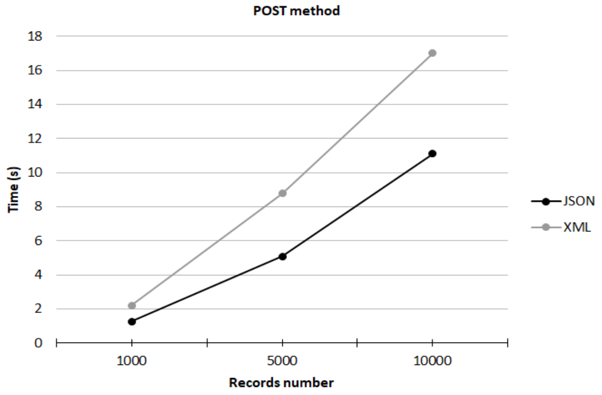


Fig. 6. XML and JSON Graph for Response Time in Seconds – POST Method.

TABLE V. XML AND JSON RESPONSE TIME IN SECONDS–POST METHOD

Records no.	JSON	XML	% XML over JSON
1000	1.25 s	2.21 s	176
5000	5.09 s	8.78 s	172
10000	11.11 s	17.02 s	153

Data sent in the XML format is bigger as size that the data send as JSON this will also affect the call response time, as the data needs more time to get to the server. This can be seen in Fig. 6 and Table 5 where for the same number of elements, the time is bigger for the call that send data as XML to the server.

The response time was bigger each try for the data sent as XML, for 10,000 records sent to be inserted the time for JSON was 11.11 seconds and for XML it was 17.02 seconds, this means that the call using JSON formatted data is 1.53 times faster for 10,000 records. Even if using XML format implies more than 50% more response time, it should be noticed that this percent does not increase much if number of records increases, remaining more or less at similar levels or even decreasing. Consequently, the (negative) impact of using XML format is decreasing as the number of implied records increase.

C. *Send and Update Data (PUT)*

For the PUT method, sending and inserting data is done by calling the following URL (*cURL* function) and putting into the body of the call the data that is being sent and is going to be inserted.

To send and update data in JSON format

```
curl -X PUT http://localhost/api/json/people
```

To send and update data in XML format

```
curl -X PUT http://localhost/api/xml/people
```

The data that is sent to the server is updated if is found in the database by the ID that is specified for each record, as a response message we get a success one if all the records were updated successfully or an error one if the operation could not be executed.

The data that is sent to the web API, formatted as JSON needs to have the following shape:

```
[
    {       "id":1,
            "first_name":"Lucille",
            "last_name":"Baddoe",
           "email":"lbaddoe0@earthlink.net",
            "gender":"Female",
            "country":"China",
            "bdate":"1989-05-12",
            "salary": 7433.55,
            "review":"Lorem ipsum...",
    },
    …
]
```

As for the XML, data the need to be formatted like the example below:

```
<?xml version="1.0"?>
<root>
  <Collection>
    <row_0>
            <id>1</id>
            <first_name>Elise</first_name>
            <last_name>McGurn</last_name>
```

```
                <email>emcgurn0@a8.net</email>
                <gender>Female</gender>
                <country>Philippines</country>
                <bdate>1989-05-12</bdate>
                <salary>7433.55</salary>
                <review>Lorem ipsum...</review>
        </row_0>
                …
        </Collection>
    </root>
```

As can be seen from Fig. 7 and Table 6, for the PUT method, the response time for sending data and updating it, for both the XML format and the JSON format, has a similar value for a small number of elements, the difference between them appears only for a larger number of items, where the JSON format has a better time.

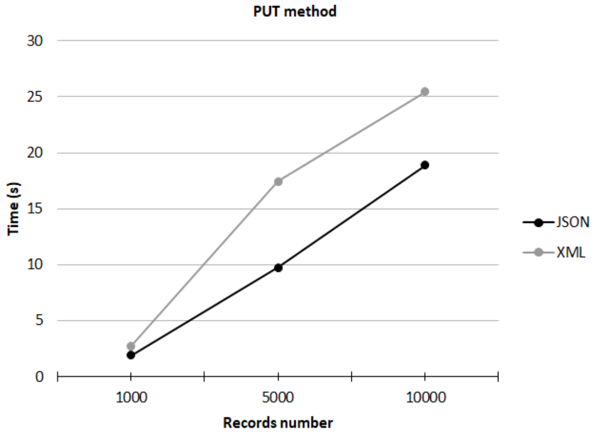


Fig. 7. XML and JSON Graph for Response Time in Seconds–PUT Method.

TABLE VI. XML AND JSON RESPONSE TIME IN SECOND –PUT METHOD

Records no.	JSON	XML	% XML over JSON
1000	1.9 s	2.73 s	143
5000	9.73 s	17.43 s	179
10000	18.9 s	25.44 s	134

D. Data Deletion (DELETE)

The deletion of the elements in the database is done by calling one of the following API URLs (*cURL* function), where the limit parameter represents how many elements will be deleted from the database through the API.

To delete data in JSON format

```
curl -X DELETE
http://localhost/api/json/people?limit=1000
```

To delete data in XML format

```
curl -X DELETE
http://localhost/api/xml/people?limit=1000
```

To delete the records, the called API function will execute a database query that deletes a number of rows less than or equal to the value sent in the limit parameter.

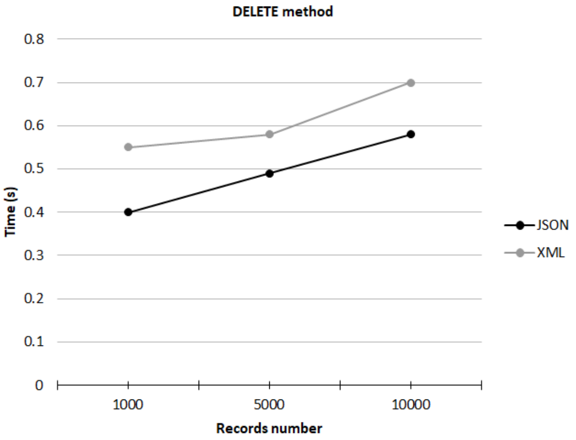


Fig. 8. XML and JSON Graph for Response Time in Seconds–Delete Method.

TABLE VII. XML AND JSON RESPONSE TIME IN SECOND–DELETE METHOD

Records no.	JSON	XML	% XML over JSON
1000	0.4 s	0.55 s	137
5000	0.49 s	0.58 s	118
10000	0.58 s	0.7 s	120

As can be seen in Fig. 8 and Table 7, for the DELETE method, the response time, both the XML format and the JSON format, has a similar value for all tested values for example when deleting 1000 or 5000 elements. For the deletion of 10000 elements, the response time is 0.58 s for JSON and 0.7 s for XML. However, the impact of using XML implies only around 20% more response time than for JSON.

Time response values are similar because the functions used to encode and to decode data in JSON or XML format are less used for this method, here the execution time of the SQL DELETE query is having a greater importance in the response time of the API.

VII. CONCLUSIONS

In conclusion, the case study from this paper presents a comparison of the data transfer methods for XML and JSON models.

The comparison of the two formats was achieved by building a web API in the PHP framework called Laravel. This API supports both formats for four operations that are related to data transfer: data request (GET method), send and insert data (POST method), send and update data (PUT method), deleting data (DELETE method).

Each of the four operations was tested through the API, which was called for this paper using an API development environment called Postman.

These operations were then analysed by two criteria, the response speed, in seconds, and the size of the data received, in KB. The data is also validated to prevent wrong data to be sent to the server. The tests are done with and without gzip server compression, because not all servers have this option enabled. The obtained results for each test were analysed and discussed in detail; as an overall evaluation, we noticed that for all operations, except the deletion one, the JSON format is more effective both in terms of data size and time of web API response time, which was in some cases 30 to 40% faster for JSON than the XML format, being generally lower as number of records implied is increasing. However, for some applications that require sending heterogenous complex structures, for which XML format offer better support, the above percentage impact over performance must be assumed. Consequently, when speaking about the necessity of XML utilization, an issue that could be investigated in the future implies the use of a converter for switching from XML to JSON. The possibility of using such converters and their impact on performance issues represent a future development that will be further investigated.

References

[1] B. Matthias, “Restful Api Design”, CreateSpace Independent Publishing Platform, ISBN-10: 1514735164, ISBN-13: 978-1514735169, 2016

[2] Z.U. Haq, G.F. Khan and T. Hussain, “A Comprehensive analysis of XML and JSON web technologies”, New Developments in Circuits, Systems, Signal Processing, Communications and Computers, pp. 102-109, 2013

[3] H. S. Padda1 and G. K. Gupta, “Analysing Impact of Delimiters on the Size of JSON Data Interchange Format”, International Research Journal of Engineering and Technology, Vol. 2, No. 8, -ISSN: 2395-0056, www.irjet.net, 2015

[4] A. Simec and M. Maglicic, “Comparison of JSONamd XML Data Formats”, Central European Conference on Information and Intelligent Systems; Varazdin, Croatia, pp. 272-275, 2014

[5] D. Peng, L.Cao, and W Xu, “Using JSON for Data Exchanging in Web Service Applications”, Journal of Computational Information Systems 7: 16, ISSN 5883-5890, 2011

[6] P. Bourhis, J. L. Reutter, F. Suárez and D.Vrgoč , “JSON: Data model, Query languages and Schema specification”, Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 123-135, 2017

[7] M. Šabanović, M. Saračević and E. Azizović, “Comparative analysis of AMF, JSON and XML technologies for data transfer between the server and the client”, Periodicals Of Engineering and Natural Sciences, Vol. 4, No.2, 2016

[8] S. Zunke and V. D’Souza, “JSON vs XML: A Comparative Performance Anaysis of Data Exchange Formats”, IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4, ISSN 2277-5420, www.IJCSN.org, 2014

[9] K. Ishwarjit, K. Sharanpreet and K. Gurinder, “Accessing Remote Database in IOS Application using JSON Parsing with Objective-C”, International Journal of Advanced Technology in Engineering and Science, Vol. no. 5, No. 1, www.ijates.com, 2017

[10] N. Nurseitov, M. Paulson, R. Reynolds and C. Izurieta, “Comparison of JSON and XML Data Interchange Formats”: A Case Study, ISCA 22nd International Conference on Computers and Their Applications in Industry and Engineering, pp. 157-162, 2009

[11] B. Šandrih, D. Tošić and V. Filipović, “Towards Efficient and Unified XML/JSON Conversion - A New Conversion”, IPSI BgD Transactions on Internet Research (TIR) vol. 13, no. 1, ISSN 1820-4503, 2017

[12] W. Natham, “Learning Laravel 5 - Building Practical Applications”, 5th edition, 2017

[13] G. David, T. Brian, S. Marjorie, A. Anshu and R. Sailu, “HTTP: The Definitive Guide”, O'Reilly Media, ISBN-13: 978-1565925090, ISBN-10: 9781565925090, 2002

[14] JSON Schema - http://json-schema.org/

[15] https://github.com/justinrainbow/json-schema

[16] https://packagist.org/packages/seromenho/xml-validator

[17] M. Tom, “JSON at Work”, O'Reilly Media, ISBN-13: 978-1449358327, ISBN-10: 1449358322, 2017

Abnormal Region Extraction from MR Brain Images using Hybrid Approach

Nikhil Gala[1]
Department of Electronics & Telecommunication
Mukesh Patel School of Technology Management &
Engineering, NMIMS University Mumbai, India

Dr. Kamalakar Desai[2]
Academic and Technical Advisor
Guru Gobind Singh College of Engineering & Research
Centre Nashik, India

Abstract—**Automatic brain abnormality segmentation from magnetic resonance images is a key task that is performed by computer aided algorithm or manual extraction by a medical expert. The regions are often partitioned based on the similarities of intensities that persist in a particular region. MR brain image segmentation is a critical step that helps to identify the abnormal region. Accurate identification of this abnormal region helps the radiologist and surgeons in surgical process and research. Through this paper we present a hybrid approach of algorithms based on clustering approach like region and edge based algorithm involved in segmenting abnormal region from MR brain images. The method is an integration of region based (pillar K-means) and edge based (level set) segmentation algorithm that aims to segment the abnormal region precisely. Experimental results show that the proposed approach could attain segmentation efficiency of 89.2%, mitigating the segmentation errors that were prevalent with region or edge based algorithms.**

Keywords—Clustering algorithm; hybrid approach; MR brain image segmentation; level set; pillar k-means; segmentation errors

I. Introduction

Brain is considered as one of the vital and important part of human body which is made up of nerve cell called neurons and the supporting cells called Glial Cells which are meant to send and receive the messages to different parts of the body and thus control the body parts. The abnormality in the brain is an extra tissue that has grown in any part of the brain. In some cases the brain cells are multiplied in an uncontrolled manner from these regions. In general these tumors are classified as benign and malignant based on its growth and orientation. The benign type of tumor grows slowly and does not spread to other region. Malignant tumors grow at a much faster rate unlike benign tumors thus causing pressure on surrounding tissues that may lead to interference in body parts functionality. There are multiple imaging methods by which the abnormality is detected but Magnetic resonance imaging (MRI) has proved to be best in detecting brain abnormality this is due to its high contrast projection and high resolution. The human body parts are thus safe from getting direct exposure to multiple deteriorating radiations. This MR imaging of brain is considered in this work as it can produce the projections in three different view perspectives like sagittal, axial and coronal planes [1].

Multiple planar projections help to precisely locate and detect the abnormality or lesion regions. Hence, it can be stated that from discussed points it is apparent that T1-weighted MR brain imaging is more adequate for detection of abnormal regions which are segmented along with other components like edema and necrosis [2][3][18].

The traditional approach of analyzing MR tumor images by medical expert is often tedious and consumes a lot of time. Therefore it is suggested to have an automatic segmentation approach that can provide an equivalent performance to manual observation. Effective segmentation approach may locate and can be employed to measure the density and volume of the abnormality which is crucial in deciding the stage of severity.

Multiple approaches have been proposed so far by researchers and most of them belong to region based or edge based approaches. Edge based approaches rely and focus on the information that is present at the corners of the regions while the pixel or intensity based approach focuses on the internal variations of intensities. Despite of numerous algorithms for segmentation and extraction of abnormal regions MR brain image segmentation is still a challenging problem for researchers and the medical practitioners, this is because of the presence of multiple variations of regions in intensity and shapes. This paper focuses on presenting a hybrid approach of integrating the level set which is an edge based approach with pillar K-means algorithm which is a region based approach. The paper is structured as follows: Section 1 emphasizes on the need and necessity of the research and a basic introduction about MR brain imaging. Section 2 covers the literature on the related work that has been conducted earlier. Section 3 presents the background concepts that were used in the proposed approach. Section 4 presents the proposed approach with its outcome and comparisons with earlier approaches ending with the conclusions and discussion in Section 5.

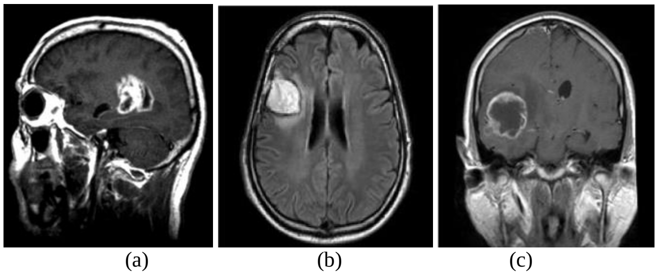


Fig. 1. MRI Brain Abnormality View in (a) Sagittal (b) Axial (c) Coronal.

II. Related Work

In [4] Maitra et al. has shown self-organizing maps based classification along with FCM (fuzzy C-means) clustering. The major contribution towards classification for MRI segmentation is done by Gibbs et al. [5], Zhu and Yan [6], Ho et al. [7]. Most extensively validated and appreciated system was presented by Clark et al. in [8] with two main components of this system being the Fuzzy C-Means.

A Multi class abnormal tissue i.e. brain tumor classification is proposed by S Dawood et al. in [9], using sparse coding and dictionary based learning. K-SVD algorithm is employed and the topological features are extracted to build the dictionary. The results of sparse are proven to be good than other methods. 3D image segmentation is proposed by Abbas and Farshad in [10] that aims to identify the image clusters and classify them. With this approach the processing time and memory utilization is reduced by 20 % however the classification is performed using Jacquard's coefficient.

A combination of clustering and region growing approach is proposed by Hooda and Verma in [11], in which they combined the region growing approach with the benchmark fuzzy C- means and K-means clustering approaches. Using this approach an accurate location and orientation of the abnormal tissue can be identified.

Segmentation with traditional Gradient Vector Flow (GVF) model was proposed by Tao wang et al. in [12] where they integrated the approach with traditional and BVF snake and magneto active contour model approaches. In this approach the GVF is applied for detecting the boundaries of the abnormal regions however this process is not suitable for 3D rendering hence interpolation is applied at the later stages.

A hybrid segmentation approach was proposed by K. Verma et al. in [13] that aims to find exact contour of the abnormal regions. This approach includes watershed segmentation along with some edge operators integrated with morphological operations. This approach has been able to detect and analyze the size of the brain tumor region in the acquired MR brain images.

A combination of co-clustering approach with morphological operation for the extraction of tumor region was proposed by Satheesh et al. in [14]. Firstly, mathematical morphological operations were employed on T1-weighted MR images that intend to remove the non-brain regions and tissues including skull, fat and muscles; this process increases the efficiency of the segmentation algorithm. The skull and fat regions are often interfering with brain region during the process of segmentation resulting in inefficient segmentation. After the process of skull removal the later obtained brain region is subjected to co-clustering algorithm for segmentation of brain tumor.

Reyes et al. in [19], proposed an ROI based abnormal segmentation by integrating active contour models, clustering approach and some morphological operations. With this approach they could attain an efficiency of 88. 2% than can be further increased.

Setyawan and others in [20] proposed a hybrid mechanism integrating clustering approach with morphological operations but could attain only 73.65% of accuracy.

Many of these methods have the limitations as in the case of edge and region based methods, so in order to overcome these deteriorations and accurately detect the abnormal region, a hybrid approach is proposed in this paper.

III. Background

A. Level-Set Approach using Non Re-Initialization

Level set approach is one of the effective ways to implement active contours highly recommended to partition multiple regions from background. Many researchers have proposed several algorithms to deal with the problem of segmentation in computer vision however active contour models employed with level sets are more effective.

Active contours are employed through zero level set methods. This method can be realized as a function ϕ which is time dependent that varies according to the equation mentioned below

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0 \tag{1}$$

Equation (1) is known as level set equation. In the above equation (1) the term "F" is termed as speed function that relies on image data structure and ϕ which is the level set function. It is mandatory to ensure that the evolving level set function is very close to signed distance function such that a stable curve may attain during the process of implementing the level set approach. To own this criteria re-initialization of the function is not recommended however this may result to higher computation and numerical errors.

Let $'I'$ be an image, and $'g'$ is an edge indicator function which is defined by below equation (2)

$$g = \frac{1}{1+|\nabla G_{\sigma} * I|^{2}} \tag{2}$$

Where G_{σ} is the Gaussian kernel with standard deviation σ.

The external energy of the function $\phi(x, y)$ is defined as

$$\mathcal{E}_{g,\lambda,v}(\phi) = \lambda \mathcal{L}_g(\phi) + v\mathcal{A}_g(\phi) \tag{3}$$

Where $\lambda > 0$ and v are constants.

The terms $\mathcal{L}_g(\phi)$ and $\mathcal{A}_g(\phi)$ are given by the below equations

$$\mathcal{L}_g(\phi) = \int_{\Omega} g\delta(\phi)|\nabla\phi| dx\, dy \tag{4}$$

$$\mathcal{A}_g(\phi) = \int_{\Omega} gH(-\phi) dx\, dy \tag{5}$$

Where δ is the univariate Dirac function, and H is the Heaviside function.

The total energy function is defined as

$$\mathcal{E}(\phi) = \mu\mathcal{P}(\phi) + \mathcal{E}_{g,\lambda,v}(\phi) \tag{6}$$

The term $\mathcal{P}(\phi)$ is called the internal energy function that penalizes the deviation of the function from the external energy and signed distance function. This inherits to drive the advancement of the zero level set towards region boundaries [15].

The evolution equation of this level set function is defined as

$$\frac{\partial \phi}{\partial t} = \mu\left[\Delta\phi - div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right] + \lambda\delta(\phi)div\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + vg\delta(\phi) \quad (7)$$

B. Segmentation using Pillar K-Means

Let us assume that there exists a data X={x_i |i=1,…,n} and 'k' being the clusters and C={c_i | i=1,…,k} where C represents the original position of centroids. Let there exists a subset SX ⊆ X which identifies the term X and selected in the subsequent process. Let the distance metric between the elements is termed as DM={xi |i=1,…,n} and the accumulated distance is termed as D={xi | i=1,…,n} that is calculated after each iteration. The mean of X is denoted as 'm'[16].

The steps of the proposed algorithm approach are described as below

- At first initialize C=Ø, SX=Ø, and DM=[ ]
- Calculate the mean (m) and also calculate the
- distance D from X and 'm'
- Initialize the number of neighbors nmin = α. n / k
- The maximum of distance is termed as dmax
- Mark the neighborhood boundary as nbdis = β . dmax
- Initialize the iteration i=0 and determine the ith initial centroid.
- Update the distance metrics DM = DM + D
- Select the maximum of distance metric and mark it as ж
- Update the subset with the marked ж as SX=SX U ж
- Now calculate the distance between X set to ж
- The number of points fulfilling D ≤ nbdis are termed as no
- Reset DM(ж)=0
- If no < nmin, go to step 8
- Assign D(SX)=0
- C = C U ж
- i = i + 1
- If i ≤ k, we need to update the distance matrix as per step 7
- Finish with C being the solution as per the optimized initial centroids.

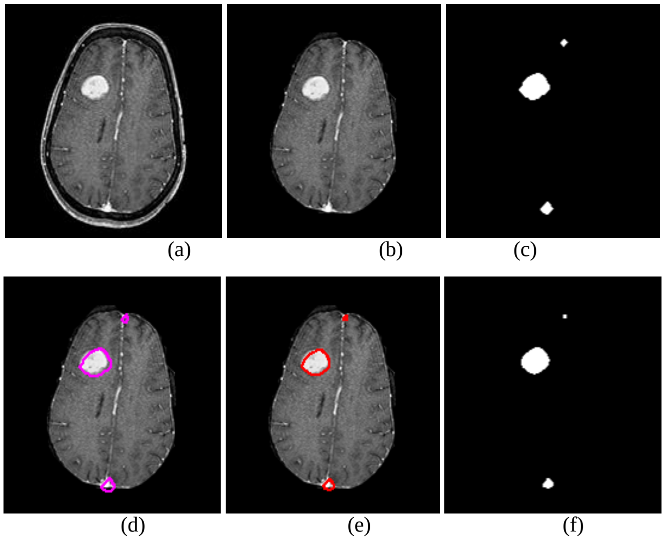


Fig. 2. Output at Various Stages (a) Original Image (b) Skull Removed (Pre-Processed Image) (c) Pillar K-Means Segmented Image (d) Initial Contour for Level Set (e) Final Contour Obtained after 500 Iterations (f) Final Segmented Image.

IV. Proposed Approach

The proposed approach is an integration of edge and region based algorithm. In this analysis T1-weighted MR image is subjected to pre-processing where the skull and fat regions are removed. Thus, obtained image is directed to pillar K-means clustering algorithm which is mentioned in the above section. The output of the clustering may contain segmentation blobs and it was observed that the over segmentation ratio is more. To mitigate this effect the boundary of the clustered abnormal region is given as an initial contour for the level set approach. In this present paper the skull removal approach mentioned by Satheesh et al. in [17] is utilized.

The step by step process is depicted in the below Figure 2. The original image is preprocessed which helps in accurate detection and extraction of abnormal region.

In order to evaluate the performance obtained for the proposed approach, the algorithm is tested with 20 patients data recorded with 1.5 Philips achieva device. As it is known that the abnormal regions are clearly visible in few of the slices, two or three slices are considered for each patient and few of the results are tabulated below.

Performance evaluation is carried out and tabulated for 8 MR images of different patients. The same MR images are used for all the 3 approaches i.e. Pillar K-means, proposed approach (Pillar + level set) and Pillar + GVF approach for comparative analysis.

Figure 3 shows the segmented output of the applied algorithms and proposed approach along with manual segmented image by expert radiologist. Table 1 shows the performance of Pillar K-means. Table 2 shows the proposed approach (Pillar + level set) output. Results are tabulated for Pillar + GVF in Table 3.

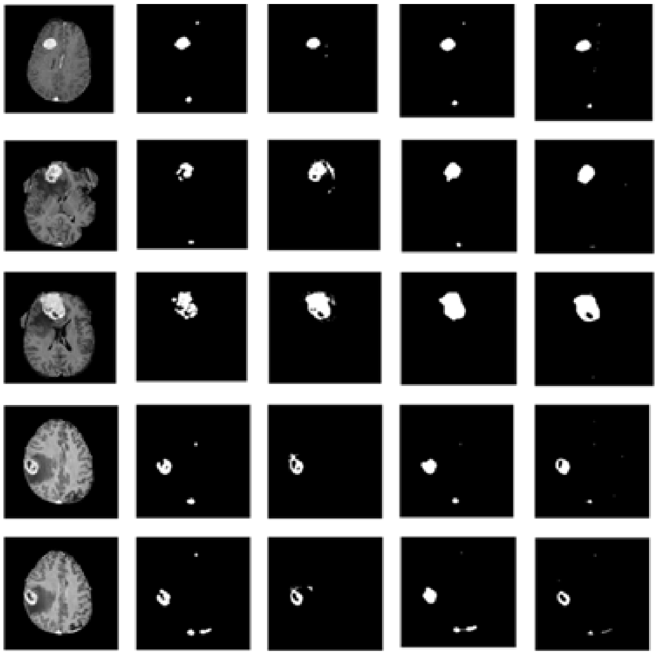


Fig. 3. (a) This Column Represents Original Pre-Processed Images (b) this Column Represents the Segmented Outputs with Pillar K-means Algorithm (c) this Column Represents the GVF+ Pillar Approach (d) this Column Represents the Segmented Output with Proposed Hybrid Approach (e) this Column Represents the Manual Segmented Images by Experts.

TABLE I. READINGS OBTAINED FOR METRIC ANALYSIS USING PILLAR K-MEANS

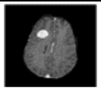
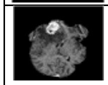
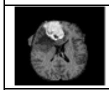
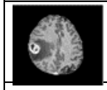
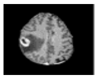
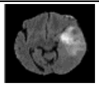
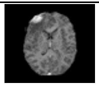
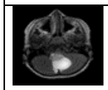
Image	**Pillar K-means**				
	SI	**CDR**	**OSE**	**USE**	**TSE**
	0.855	0.982	0.315	0.017	0.33
	0.725	0.649	0.139	0.305	0.409
	0.805	0.716	0.063	0.283	0.346
	0.8	0.877	0.313	0.122	0.436
	0.685	0.916	0.759	0.083	0.842
	0.613	0.476	0.52	0.07	0.6
	0.68	0.58	0.42	0.098	0.52
	0.59	0.31	0.52	0.06	0.58
AVG	0.719	0.688	0.381	0.129	0.507

TABLE II. READING OBTAINED FOR METRIC ANALYSIS USING PILLAR AND LEVEL-SET APPROACH

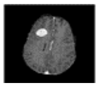
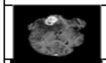
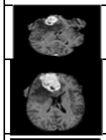
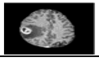
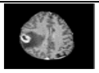
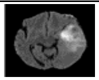
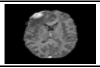
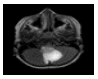
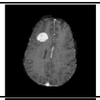
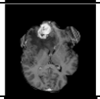
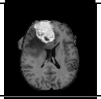
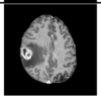
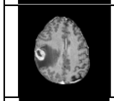
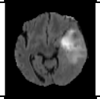
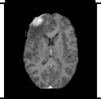
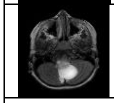
	Hybrid (Pillar + level-set)				
Image	**SI**	**CDR**	**OSE**	**USE**	**TSE**
	0.865	0.945	0.22	0.05	0.28
	0.84	0.832	0.141	0.165	0.317
	0.86	0.862	0.143	0.137	0.28
	0.812	0.887	0.299	0.112	0.413
	0.673	0.935	0.842	0.064	0.907
	0.645	0.52	0.48	0.092	0.57
	0.79	0.69	0.3	0.069	0.36
	0.68	0.65	0.44	0.088	0.48
AVG	**0.77**	**0.792**	0.358	**0.09**	**0.454**

TABLE III. READING OBTAINED FOR METRIC ANALYSIS USING PILLAR AND GVF APPROACH

Image	**Hybrid (Pillar + GVF)**				
	SI	**CDR**	**OSE**	**USE**	**TSE**
	0.862	0.833	0.07	0.16	0.241
	0.777	0.917	0.442	0.082	0.525
	0.877	0.927	0.138	0.072	0.211
	0.807	0.803	0.186	0.196	0.382
	0.717	0.778	0.278	0.221	0.5
	0.62	0.51	0.56	0.18	0.74
	0.72	0.68	0.39	0.16	0.55
	0.66	0.50	0.49	0.027	0.52
AVG	0.758	0.743	**0.312**	0.137	0.458

The proposed method is compared with Pillar K-means [16], Pillar + GVF approach and the metrical analysis is calculated with respected to manual segmented images by experts. Multiple metrics like segmentation efficiency/ Similarity Index (SE/SI), correct detection ratio (CDR), over segmentation error (OSE), under segmentation error (USE) and total segmentation error (TSE) that were mentioned in [17] were adopted for evaluating the performance of this segmentation approach. Table 4 below shows the comparative analysis of the proposed approach using these metrics. Bold values indicate better performance. Figure 4, Figure 5 and Figure 6 helps to understand that the proposed approach is more efficient in detection of the abnormal region and further reduces segmentation errors. As seen in Figure 7, the proposed approach has given better performance when SI, CDR, and TSE is compared.

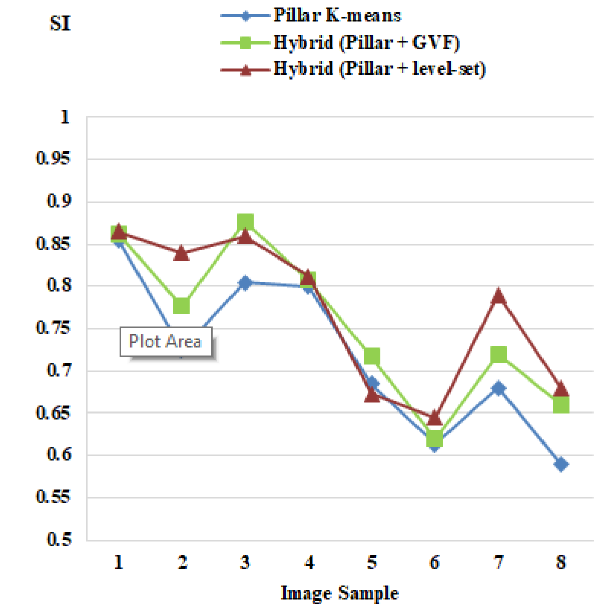


Fig. 4. Graph Representing Similarity Index Comparison.

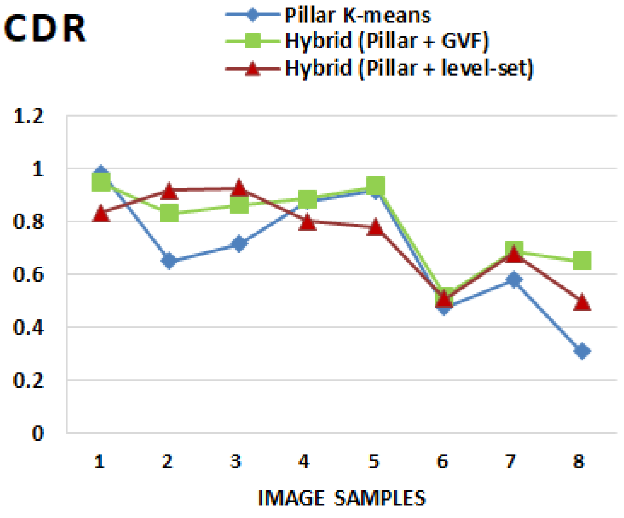


Fig. 5. CDR of Pillar K-Mean, Pillar+GVF and Proposed Approach.

TABLE IV. COMPARATIVE PERFORMANCE OF PROPOSED APPROACH

Approach	SI	CDR	OSE	USE	TSE
Pillar K-means	0.719	0.688	0.381	0.129	0.507
Pillar + level-set	**0.77**	**0.792**	0.358	**0.09**	**0.454**
Pillar + GVF	0.758	0.743	**0.312**	0.137	0.458

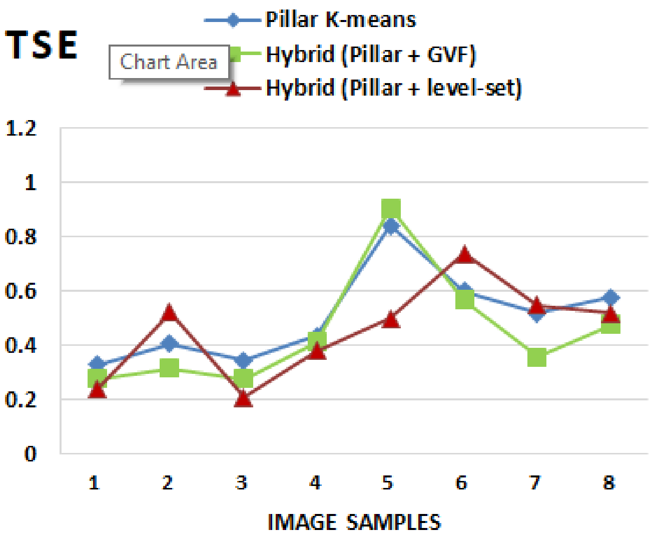


Fig. 6. Comparison of Total Segmentation Error.

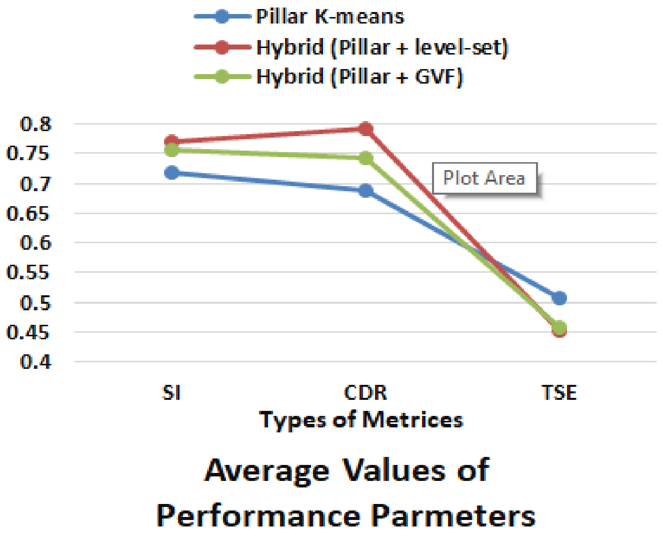


Fig. 7. Comparative Analysis of Proposed Approach.

V. CONCLUSION

This work presents a new low complex hybrid segmentation approach that incorporates pillar K-means and a level set method proposed. The method is able to efficiently segment the abnormal region from the input pre-processed images. The performance achieved from the proposed approach is compared with the pillar K-means and found that in all aspects it is yielding better results. In precise when compared with integrated GVF approach with pillar K-means the method has gained 2% more similarity and attained 5 % more correct detection. However, the method has over segmentation error larger than Pillar + GVF approach by 4 % that has to be mitigated. Finally, it can be concluded that the proposed approach has been able to reach the objectives of this work in minimizing the segmentation errors that were occurring with traditional clustering approaches.

References

[1] American Society of Neuroradiology. "ACR-ASNR Practice Guideline for the Performance and Interpretation of Magnetic Resonance Imaging (MRI) of the Brain" -2013

[2] Galloway, RL Jr, “Introduction and Historical Perspectives on Image-Guided Surgery ", In Golby, AJ, Image-Guided Neurosurgery, Amsterdam: Elsevier. pp. 3-4, 2015

[3] TSE, VCK; Kalani, MYS; Adler, JR, "Techniques of Stereotactic Localization". In Chin, LS; Regine, WF. Principles and Practice of Stereotactic Radio surgery. New York: Springer. pg.no 28,2015

[4] M. Maitra, A. Chatterjee,”Hybrid multi resolution Slantlet transform and fuzzy c-means clustering approach for normal-pathological brain MR image segregation”, Medical Engineering and Physics, Elsevier Publishers,2007.

[5] Gibbs, P., Buckley, D., Blanckb, S., and Horsman, A.,” Tumor volume determination from MR images by morphological segmentation”, Physics in Medicine and Biology, 41:2437–2446, 1996.

[6] Zhu, Y. and Yan, H. “Computerized tumor boundary detection using a Hopfield neural network” IEEE Transactions on Medical Imaging, 16:55–67, 1997.

[7] Ho, S., Bullitt, E., and Gerig, G. “Level set evolution with region competition: automatic 3D segmentation of brain tumors”, In 16th International Conference on Pattern Recognition, pages 532–535,2002.

[8] Clark, M., Hall, L., Goldgof, D., Velthuizen, R., Murtagh, F., and Silbiger, M., “Automatic tumor segmentation using knowledge- based techniques. IEEE Transactions on Medical Imaging, 17:238–251,1998

[9] Al-Shaikhli, S. D. S., Yang, M. Y., & Rosenhahn, B. “Brain tumor classification using sparse coding and dictionary learning”, In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 2774-2778). IEEE-2014

[10] Abbasi, S, Pour, F. T, “A hybrid approach for detection of brain tumor in MRI images”, In Biomedical Engineering (ICBME), 2014 21th Iranian Conference on (pp. 269-274). IEEE, 2014

[11] Hooda, H, Verma, O. P., & Singhal, T., “Brain tumor segmentation: A performance analysis using K-Means, Fuzzy C-Means and Region growing algorithm”. In Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on (pp. 1621-1626), 2014

[12] T. Wang, I. Cheng and A. Basu , "Fluid Vector Flow and Applications in Brain Tumor Segmentation," in IEEE Transactions on Biomedical Engineering, vol. 56, no. 3, pp. 781-789, March 2009.

[13] Verma, Kimmi, and Shabana Urooj. "Effective evaluation of tumor region in brain MR images using hybrid segmentation." 2014 International Conference on Computing for Sustainable Global Development (INDIACom), 2014

[14] S.Satheesh, Dr.K.V.S.V.R Prasad, Dr.K.Jitender Reddy, “Tumor Extraction And Volume Estimation For T1-Weighted Magnetic Resonance Brain Images”, Global Journal of Computer Science and Technology Neural & Artificial Intelligence, Volume 12 Issue 12 Version 1.0 Year 2012

[15] Chunming Li; Chenyang Xu; Changfeng Gui; Fox, M.D., "Level set evolution without re-initialization: a new variational formulation," Computer Vision and Pattern Recognition, 2005. CVPR 2005.IEEE Computer Society Conference on , vol.1, no., pp.430,436 vol. 1, 20-25 June 2005

[16] Ali Ridho Barakbah ; Yasushi Kiyoki, A pillar algorithm for K-means optimization by distance maximization for initial centroid designation, Computational Intelligence and Data Mining, 2009.

[17] S. Satheesh, R. T. Santosh Kumar, K.V.S.V.R Prasad and K. Jitender Reddy, "Skull removal of noisy magnetic resonance brain images using Contourlet transform and morphological operations," Proceedings of 2011 International Conference on Computer Science and Network Technology, Harbin, 2011, pp. 2627-2631.

[18] Nolen-Hoeksema, Susan. Abnormal Psychology (Sixth edn). New York, NY: McGraw-Hill Education. p. 67,-2014

[19] A. M. de los Reyes, M. Elena Buemi, M. N. Alemán and C. Suárez, "Development of a graphic interface for the three-dimensional semiautomatic glioblastoma segmentation based on magnetic resonance images," 2018 Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación (CACIDI), Ciudad Autónoma de Buenos Aires, Argentina, 2018, pp. 1-6.

[20] R. Setyawan, M. A. Almahfud, C. A. Sari, D. R. I. M. Setiadi and E. H. Rachmawanto, "MRI Image Segmentation using Morphological Enhancement and Noise Removal based on Fuzzy C-means," 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 2018, pp. 99-104.

Linear Intensity-Based Image Registration

Yasmin Mumtaz Ahmad[1], Shahnorbanun Sahran[2],
Afzan Adam[3]
Center for Artificial Intelligence Technology,
Faculty of Information Sciences and Technology,
Universiti Kebangsaan Malaysia,43600 Bangi, Malaysia

Syazarina Sharis Osman[4]
Department of Radiology
Hospital Canselor Tuanku Mukhriz
Cheras, Universiti Kebangsaan Malaysia,
56000 Kuala Lumpur, Malaysia

***Abstract*—The accurate detection and localization of lesion within the prostate could greatly benefit in the planning of surgery and radiation therapy. Although T2 Weighted Imaging (T2WI) Magnetic Resonance Imaging (MRI) provides an infinite amount of anatomical information, which ease and improve diagnosis and patient treatment, however, modality specific image artifacts, such as the occurrences of intensity inhomogeneity are still obvious and can adversely affect quantitative image analysis. Conventional high resolution T2WI has been restricted in this respect. On the contrary, Apparent Diffusion Coefficient (ADC) map has been seen as capable to tackle T2WI limitation when a functional assessment of the prostate capable to provide added value compared to T2WI alone. Likewise, it has been shown that diagnosis using ADC map combined with T2WI significantly outperforms T2WI alone. Therefore, to obtain high accuracy detection and localization, a combination of high-resolution anatomic and functional imaging is extremely important in clinical practice. This strategy relies on accurate intensity based image registration. However, registration of anatomical and functional MR imaging is really challenging due to missing correspondences and intensity inhomogeneity. To address this problem, this study researches the used of applying linear geometric transform to the corresponding point to accurately mapping the images for precise alignment and accurate detection. Transformation type is crucial for the success of image registration. The selection of transformation type is influenced by the type and severity of the geometric differences between corresponding images, the accuracy of the control point between images, its density and organization of the control points. A transformation type is selected to reflect geometric differences between two images in image registration. Often, the selection of the suitable transformation type for image registration is undeniably challenging. To make this selection as effective as possible, an experimental mechanism has to be carried out to determine its suitability. These transformations types are Affine, similarity, rigid and translation. Additionally, intensity based image registration is implemented to optimize the similarity metric mean square error through regular step gradient descent optimizer. Accuracies evaluation for each transformation type has been carried out through mean square error (MSE) and peak signal noise ratio (PSNR). The results have been presented in a chart form together with a comparison table.**

***Keywords*—Lineargeo metric transformation; image registration; point correspondence**

I. Introduction

The treatment of prostate cancer could greatly benefit from discovering imaging markers in MRI images that accurately predict existence lesion. Although MRI provide a massive amount of anatomical and functional information, which ease expert's daily task and improve diagnosis and patient treatment, However, modality specific image artefacts, such as the phenomena of intensity inhomogeneity in MRI, are still obvious and can adversely affect quantitative image analysis[1]. Image artefacts exist due to variability between patients during the MRI examinations even using the same scanner, protocol or sequence parameters [2]. The researcher in [3] and other researchers have emphasized the need to develop a system that can extract quantitative data in a more accurate and automated fashion. One possible approach for discovering these markers is to first align the anatomical and functional MR of a patient and then to analyse the imaging characteristics of suspicious lesion. This strategy relies on accurate registration. However, registration of anatomical and functional MR imaging is very challenging due to missing correspondences and intensity inhomogeneity between the MR imaging. To address this problem, this study researches the used of applying linear geometric transform to corresponding point to accurately maps the images for precise alignment and better observation. The process of mapping points from a reference image to the corresponding points in the target image is identified as an intensity based image registration[4]. Image registration comprises of four major components, which are a feature selection, a transformation model, a similarity metric and an optimization method. This paper examines the second step in image registration that is, transformation that is used to map corresponding points in the images. The type of transformation selected is vary depending on the type and density of the geometric differences between the images, the accuracy and organization of the matching control point [4].Although defining transformation type selection are essential, this paper will not cover the selection transformation criteria. Many authors have discussed method for transformation type selection. This paper presents the analysis for image transformation model, which is the second step in image registration. The linear geometric transformation will be investigated due to the fact that it is adequate to solve medical image registration acquired from same sensors [4]. This paper will examine the used of linear transformations for aligning two dimension (2D) Magnetic Resonance images (MRI) acquired from the same sources. The images are labelled as reference image for T2 Weighted Imaging (T2WI) and target for Apparent Diffusion Coefficient (ADC) image. These labelled will be consistently used in the entire paper. Both images acquired from axial plane as shown in Fig. 3a and 3b.

A. Prostate Cancer

The Prostate is divided into three glandular zones : the peripheral zone contains 70 percent of the glandular tissue and accounts about 70 percent of prostate cancer[5].The transition zone contains 5 percent of the gland tissues and accounts for 25 percent of prostate cancer[5] and the central zone contains 20 percent of glandular tissue and account most fewer percent of prostate cancer[5].Accurate early detection within peripheral zone and transitional zone is crucial as both zones are associated with favourable pathologic features and better recurrence free survival[5]. At present, clinical procedure for prostate cancer diagnosis is trans-rectal ultrasound and systematic biopsy.

B. MR Image

Conventional MRI of the prostate combines anatomic images T2WI and functional information obtained from ADC. This combination is identified as multi parametric-MRI(MP-MRI). This section clarifies the sequence used by radiologists in their daily diagnosis task. These sequences are T2WI and ADC.

1) T2-weighted imaging (T2WI): The first and most important step in an MP-MRI protocol used to perform diagnosis as it is well suited to render zonal anatomy of the prostate[6].Peripheral zone and central gland tissues are well observable in these image[7]. On this sequence, prostate cancer generally demonstrates low signal intensity (SI) contrast, to the high SI on the normal peripheral zone[8]. A sample of this image can be seen in Fig. 3a with delineated tumour. Since, T2WI is most important sequence to locate prostate cancer, therefore, it is the most frequently used sequence in computer aided detection system.

2) Apparent diffusion coefficient (ADC): Diffusion weighted imaging (DWI) measures the diffusion of water molecules within different tissues[9]. Normal prostate gland tissue has a higher water diffusion rate than cancer tissue[9].DWI is inherently T2WI but, unlike conventional T2WI, prostate cancer usually demonstrates increased SI on DWI effecting the lesion difficult to visualize within the normal high signal peripheral zone[9].To minimize the effect of this T2WI sine through effect that may result in the depiction of a false positive high intensity lesion[10], the ADC is calculated[8], which is calculated based on diffusion-weighted imaging(DWI) using various b values, is useful for differentiate existence lesion either benign or malignant [11].Furthermore, Prostate lesion visible as a high signal region on DWI but as low signal region on ADC map. Fig. 3b shows an ADC map with delineated lesion as low signal intensity (arrow).

C. Types of Linear Transformation

1) Affine: A common affine transformation from 2D to 2D as in below equation 1 derived from [12] which requires six parameters and computed from only three matching pairs of points [12]

$$[[xi ; yi]; [Xi; Yi]]i=1;3] \tag{1}$$

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \square \begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ 0 & 0 & 1 \end{bmatrix} \square \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

2) Similarity: The similarity transform represent global translation, rotation and scaling dissimilarities among reference and target images [13] and only need two point control. Equations 3 derived from [13] represent similarity transforms

$$X = xs\cos\theta - ys\sin\theta + h \tag{3}$$

$$Y = xs\sin\theta + ys\cos\theta + k \tag{4}$$

Where

s, θ, and (h, k) are scaling, rotational, and translation dissimilarities between the images. These parameters could be easily identified if the images correspondence control point coordinates are known. The rotational dissimilarity between the images is determined from the angle between the lines connecting the two points in the images. The scaling differences between the images are determined from the ratio of distances between the points in the images.

3) Rigid: A rigid transformation is described with translations and rotations. It has only three parameters as represented in equations 5.

$$T\ linear(x) = Rx + t \tag{5}$$

Where

R is a rotation matrix and *t* is a translation vector.

4) Translation: Two translation parameters required for 2-dimension transformation as represents in equations 6. Translated by the vector [Tx, Ty] is

$$f_T\ (x, y) = f\ (x - Tx, y - Ty) \tag{6}$$

Where

T_x is the horizontal translation and T_y is the vertical translation.

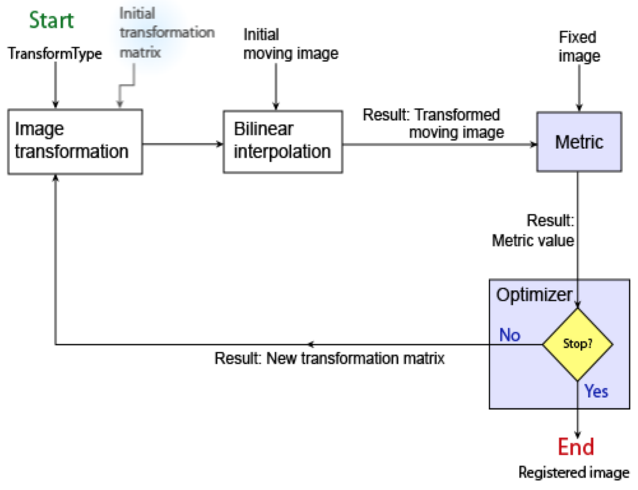


Fig. 1. Iterative Process of Intensity based Image Registration. Source from [16].

II. The Material and Method

Matlab has been adopted to evaluate the transformed images accuracy as it represents images in a matrices form [14]. An image is a M x N array of elements. Each element in the array is known as an intensity value. Intensity based image registration is an iterative process. To summarize the registration process, Fig.1 illustrates the iterative process of image registration. The ultimate goal of image registration is to search iteratively for a geometrical transformation that will minimizes the target image similarity metric until termination criteria is met [15]. For this iterative processes success, it requires a pair of images, a similarity metric, an optimizer and a transformation models. As mentioned in previous section, since this research will be registering images from the same sensor i.e. MRI, linear geometric transformations have been chosen, as this transformation model are seemingly adequate with the problem needs to be solved. The similarity metric calculates the image correspondences for weighing the accuracy of image registration. The optimizer describes the method for minimizing the similarity metric and the transformation model will transform the target image as a reference image. MR Images Acquisition.

The MR dataset obtained by the 3 Tesla(T) MRI Siemens Verio contained T1-weighted imaging, T2-weighted imaging, DWI with Apparent Diffusion Coefficient(ADC) mapping and dynamic contrast enhanced imaging(DCE)[17].Each patient was scanned with the clinical standard MRI protocol. Multiple acquisition in three orthogonal planes (axial, coronal, sagittal) were performed using high resolution T2-weighted. ADC maps were obtained from the DWI sequences[8].The size of each sequences is differ according to resolution. The size of T2 weighted imaging is a 512 x 512. The radiologist manually segmented the whole prostate. In addition, the ethical protocol was obtained and approved by the internal ethics committee and informed consent was obtained from the patients.

A. MR Image Analysis

All acquired sequences were reviewed for each patient. Initially, The T2WI has to be reviewed for the area of hypo intensity in contrast to the high signal intensity of the normal peripheral zone as shows in Fig. 3a. While on the ADC map the criteria for lesion existence were low signal intensity as depicted in Fig.3b.

B. Linear Geometric Transformation

Transformation model is one of the main characteristics of the image registration in order to precisely align reference and target images. In this section the linear transformation type will be explained further. Linear transformation changes only the spatial property of images (horizontal, vertical up,down)and does not change the property of pixels i.e intensity, color, shape [18]. This research study particularly about points transformation between two dimensions' spaces.. A 2D point has two coordinates and frequently denoted as either a row vector U=[x,y] or column vector U = [x,y]t [12]. Therefore, the used of homogeneous coordinates for points is more convenient as the selected transformation model is linear geometric transformation.

Linear geometric transformations is defined as a mapping of any straight line to a straight line [19] . It is classified into four categories: Affine, Similarity, Rigid and Translation. The simplest linear transformation model is a horizontal or vertical shift. This shifting transformation is known as translation.[20]. Affine is the most complex linear transformation includes of Translation, rotation, scaling and shear. Similarity is a composition of translation, rotation and scaling. Rigid is a combination of translation and rotation. In the next section, the geometric transformations are further discussed in mathematical representation form. In section 3 a comparison result between these models were represents. Detail explanation of linear geometric transformation as in equation 7,8 and 9 adapted from [19] represents coordinates of M pairs control points in two images of the same source.

$$[[xj, yj]; [Xj, Yj] : j = 1:M] \tag{7}$$

where a transformation function f(x,y) with elements $f_x(x,y)$ and $f_y(x,y)$ that fulfill

$$[Xj = fx(xj, yj)] \tag{8}$$

$$[Yj = fy(xj, yj), j = 1,...,M] \tag{9}$$

Once the coordinates of a point (x,y) in f(x,y) is identified, then the coordinates of the matching points in the target image can be identified as well. Here, the (x,y) is referred as reference coordinates and (X,Y) as target image coordinates. The reference image is reserved basic while the target image transform according to reference image.

C. Similarity Measure

Similarity measure computes the quantity of similarity between intensity in two images [21] [22]. The choice of an image similarity metric depends on the modality of the images needs to be registered [22]. The similarity metric used in this research is Mean Square Error (MSE).This metric is commonly used with mono-modal intensity based problem, which is the reason adopted for this research as this research is also solving mono-modal intensity based problem. The MSE is a cumulative squared error between transformed and reference image [22]. The mathematical representation for measuring MSE as in equation 10 derived from [22].

$$MSE = 1/MN * \sum^{M}_{y=1} \sum^{N}_{x=1} [J(x, y) - J'(X, Y)]^2 \tag{10}$$

Where J (x, y) is a reference image, J' (X, Y) is the target image and M, N are the dimensions of the images. A lower value signifies lower similarity error and higher similarity between the images [11]

D. Optimizer

Regular Step Gradient Descent (RSGD) is applied as optimization method. This optimizer suits most for minimization problem in image registration [23].Hence, it is chosen to minimize the similarity measure between two images. It initially moved from initial trial point X^1 moved along the steepest descent direction until the optimal point is found [23] as shown in Fig. 2. This optimizer will not terminate the iteration unless it reaches an optimal point. RSGD adopting hill climbing method begins with an opening estimate X^k of the MSE.

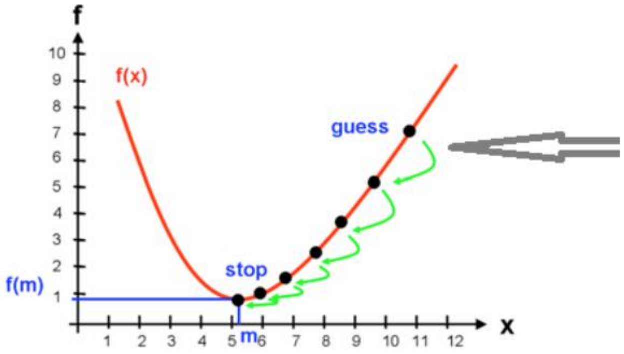


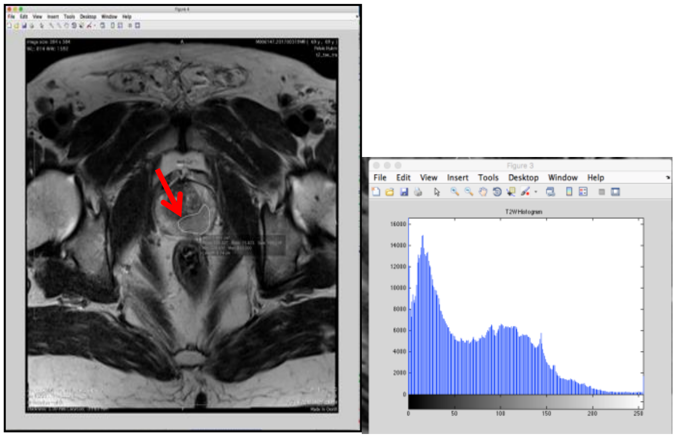
Fig. 2. Regular Step Gradient Descent Optimizer.Source from [12].

III. Results and Discussion

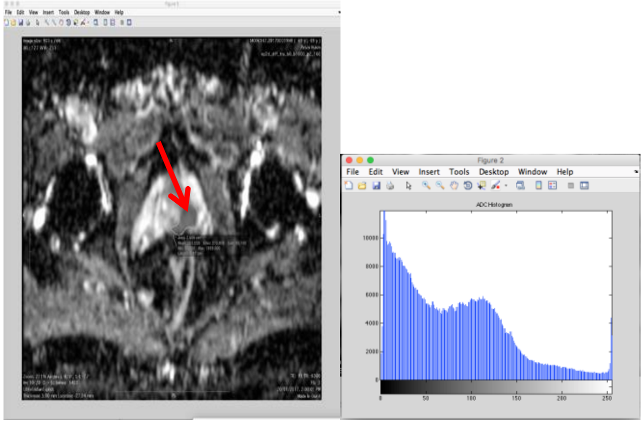
In this section, the registrations of linear geometric transform between 2D images are presented. To register a target image *J'* to a reference image *J*, a six parameter transformation (parameterized by q1, q2, q3, q4, q5q, q6) has been used. A dot in the reference image that is denoted by Xj is compared with dot in the target image. As stated in previous section, the parameters *q1-q6* are optimized by minimizing the mean square error between the images using Regular step gradient descent (RSGD).

A. Medical Image Registration using Linear Geometrical Transformation

The iteration started with the chosen 2D transformation, which is in this research a linear geometric transformation. The purpose of this transformation is to align precisely the target image ADC with the reference image T2WI. Both images were 8 bit but the width and height of both images are different; T2WI is 1080 x 1080 whereas ADC is 910 x 1080 as shown in Fig. 3a and 3b respectively and acquired by same sources. Thus, it is necessary to resize the ADC equally with T2W before the alignment process begins. The ADC image were resizing with Bi-cubic interpolation. Fig.3a (reference Image) and 3b (target Image) are MR images of a patient has prostate tumour. Tumours are marked with red arrows in respective images. Intensity histograms of both MR images are attached besides respective MR image. These MR images are taken at the same time using the same source.



(a)



(b)

Fig. 3. (a) Axial Plane T2WI Shows an Ill Defined Hypointense Lesion (Arrow) at Peripheral Zone.also this Sequence Labelled as Reference Image (T2W). Attached besides, is a Histogram Intensity for T2WI., (b) Low SI (Arrow) on ADC Map.Attached besides is a Histogram Intensity for ADC Map.

Next, the metric which is Mean squares error has been applied to compares the value of transformed ADC to T2W.Mean square error (MSE), Sum square differences (SSD) and cross correlation (CC) are commonly used for image registration acquired from the same modality (Panda, et all, 2017). Thus, MSE is applied as due to less computation time and complexity. Finally, the optimizer checks for a stopping condition. The optimizer search the minimum value of the similarity metric before it is terminated when the similarity metric reach the stopping condition [23].In this paper, regular step gradient decent has been used as optimization method. This optimization method can be expressed as in equation 11 based on [23]

$$\text{minimize}_{x \in R^n} f(x) \tag{11}$$

Where, the objective function f(x) is to locate the local minima of a transform using RSGD.

Fig. 4 shows the unregistered overlapping result of two MR images. Misregistration between two images is clearly visible as marked with red arrow. Expert has marked abnormality in both images manually. The purple is a reference image and the green is a target image.

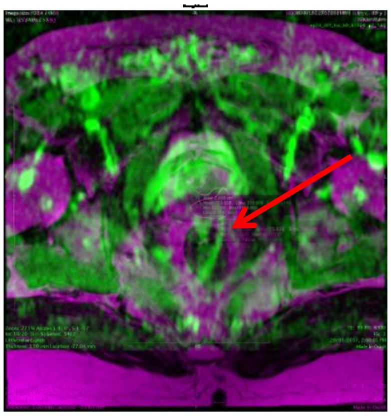
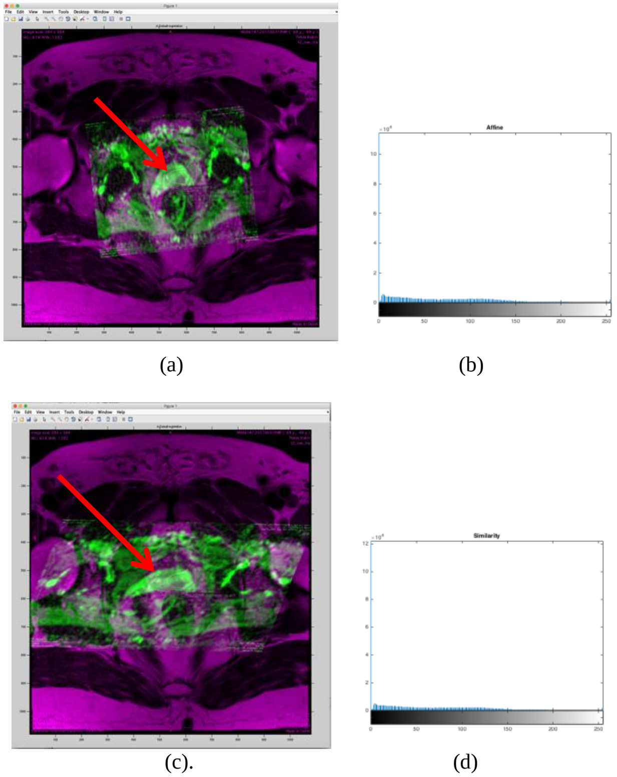


Fig. 4. Unregistered Overlapping of Reference and Target Images.

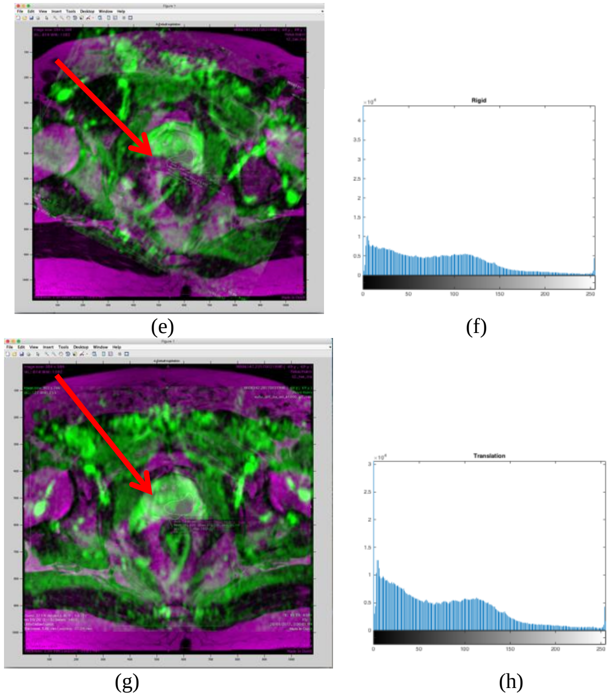
Linear Transformation models: Affine, Similarity, Rigid and translation has been implemented and evaluated for mono-modal prostate MR images. Fig. 5a-5h show the experimental results and represents histogram intensity of each linear geometrical transformation. The x and y axes in histogram represent intensity of the images. Fig. 5a shows an experimental registration result for affine transformation and generated histogram intensity. As seen here, rather than agreeing a homogenously spaced grid to select image blocks globally, affine transformation only adopt the locally spaced to select the image block. This is because Affine transformation originate from a combination of scaling, rotation, shearing and translation transformation [24].Hence, applying Affine transformation only useful when dealing with locally transformation[24].

Fig. 5c shows an experimental result for similarity transform and similarity intensity histogram. From this result, it can be seen that, similarity transformation is not suitable for registration of images acquired from same sources and images with global geometric differences. Fig. 5e shows an experimental result for rigid transformation and it intensity histogram. Fig. 5g shows an experimental result for translation transformation and it histogram intensity. It can be seen that, a rationally good accuracy was attained at and near the control points. This can be concluded that translation basis is reasonably good for registering images with global geometric differences which is uniformly spaced control points and images acquired from same modality. Table 1 illustrates evaluation summary of each linear transformation types. Mean, standard deviation, mean square error (MSE) and Peak signal noise ratio (PSNR) were applied to show the results.



(a) (b)

(c). (d)



(e) (f)

(g) (h)

Fig. 5. (a) Affine Transformation., (b) Affine Intensity., (c) Similarity Transformation, (d) Similarity Intensity., (e) Rigid Transformation., (f) Rigid Intensity., (g) Translation Transformation. , (h)Translation Intensity.

B. *Quantitative Measures of Image Registration Accuracy*

To demonstrate the accuracy and robustness of each geometrical transformation model, it is necessary to evaluate registration accuracy. Measures techniques were used to evaluate the quality of the registration are mean, standard deviation, mean square error (MSE) and peak signal noise ratio (PSNR) also, charts are generated based on these values.

1) Mean square error (MSE): MSE values indicated a value of null for perfect alignment between both images. For similar alignment achieved, the MSE values range nearly from 0.10e4 to 0.12e8 and for misregistration, the MSE values estimated in the range of 5.9e6 to 6.4e6[22]. The values of this evaluation are illustrated as in Fig.6. Translation shows better results compared to affine, similarity and rigid.

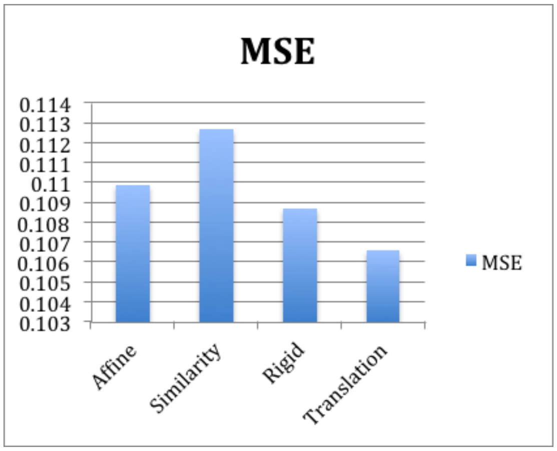


Fig. 6. Similarity Measure MSE Computed for Image Registration.

2) Peak signal noise ratio(PSNR): PSNR is identified as a numerical measure for image registration quality based on the pixel differences between two images[26]. The Mathematical equation is provided as in equation 12 based on [26]

$$PSNR=10*\log_{10}(MAX^{2)}/(MSE) \quad (12)$$

Where MAX is the maximum possible pixel value of the image and MSE is mean square error between reference and target image. Fig. 7 presents image registration accuracy using PSNR to evaluate the pixel difference between these images.

A higher PSNR value indicates a higher similarity between registered images contrarily, the smaller PSNR value indicates poor similarity between images [27].From this chart, it can be seen that, translation has achieved the highest ranking among the others. Hence, signifies better accuracy found via translation transformation.

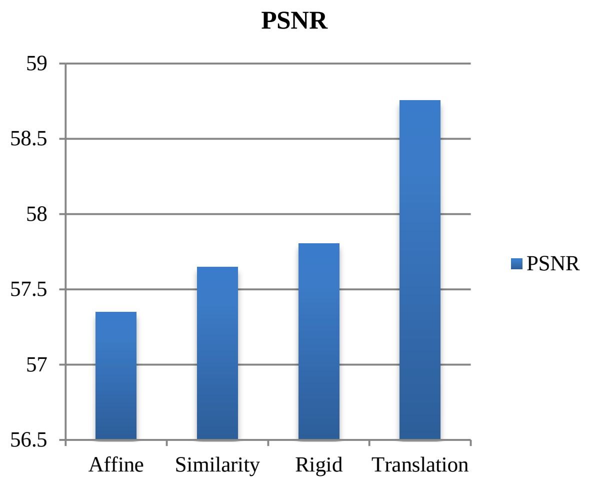


Fig. 7. Accuracy Meaurement using PSNR to Evaluate the Pixel differences between Images.

3) Result summary: Table 1 draws the registration accuracy using various measurement methods. Mean and standard deviation shows similar results for all transformation type while MSE and PSNR display dissimilar results for each transformation type.

TABLE I. SUMMARY OF LINEAR TRANSFORMATION RESULTS

Registration Accuracy				
Model/Measures	**Affine**	**Similarity**	**Rigid**	**Translation**
Mean	127.5	127.5	127.5	127.5
SD	74.05	74.05	74.05	74.05
MSE	0.1099	0.1127	0.1087	0.1066
PSNR	57.3470 dB	57.6465 dB	57.8023 dB	58.7549 dB

IV. Conclusions

Transformation types are crucial for the success of image registration. A transformation type is selected to reflect geometric differences between two images in image registration. Often, selection of the right transformation type for image registration is undeniably challenge. To make this selection as effective as possible, an experimental mechanism has to be done to determine suitability in particular image registration. Hence to understand of the above problem, a comparison between linear geometrical transformations on medical image registration is presented in this paper. The priciest alignment among all transformation types has been successfully investigated. In this study, intensity based image registration method has been used. Mean square error metric is used as similarity metric and regular step gradient descent optimizer is used as an optimization method. Data set has been acquired from Hospital Canselor Tuanku Muhriz. Based on the Mean Square Error similarity measure analysis, the translation transformation performed better in aligning the mono-modal MR images with MSE values lesser than other transformation functions. Translation control points distributed globally coverage. Thus higher similarity found between reference image and the transformed target image.

Acknowledgment

The authors would like to thank Dana Lonjakan Universiti Kebangsaan Malaysia (UKM) for financial support in publishing this paper, and administrations in Hospital Canselor Tuanku Muhriz and UKM for their good cooperation and helps in doing this project. An appreciation also goes to members of Medical Computing Lab for the ideas and fruit full discussions.

References

[1] U. Vovk, F. Pernuš, and B. Likar, "A Review of Methods for Correction of Intensity Inhomogeneity in MRI," *IEEE Trans. Med. Imaging*, vol. 26, no. 3, 2007.

[2] G. Lematre, "Computer-Aided Diagnosis for Prostate Cancer using Multi-Parametric Magnetic Resonance Imaging," 2016.

[3] L. Liu, Z. Tian, Z. Zhang, and B. Fei, "Computer-aided Detection of Prostate Cancer with MRI," Acad. Radiol., vol. 23, no. 8, pp. 1024–1046, 2016.

[4] A. Goshtasby, "Transformation functions for image registration," 2003.

[5] L. Liu, Z. Tian, Z. Zhang, and B. Fei, "Computer-aided Detection of Prostate Cancer with MRI: Technology and Applications," Acad. Radiol., vol. 23, no. 8, pp. 1024–1046, 2016.

[6] J. O. Barentsz et al., "ESUR prostate MR guidelines 2012," Eur. Radiol., vol. 22, no. 4, pp. 746–757, Apr. 2012.

[7] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," Comput. Biol. Med., vol. 60, pp. 8–31, May 2015.

[8] Noha Mohamed AbdelMaboud a, 1 *, Hytham Haroon Elsaid a, and 2 Essam Aly Aboubeih b, "The role of diffusion – Weighted MRI in evaluation of prostate cancer."

[9] E. de Kerviler, "[Imaging in prostate cancer]," Rev Prat, 2013. [Online]. Available: https://www.medscape.com/viewarticle/742986_3. [Accessed: 19-Oct-2018].

[10] T. Higaki et al., "ImagIng PhysIcs Introduction to the Technical Aspects of Computed Diffusion-weighted Imaging for Radiologists," RadioGraphics, vol. 38, pp. 1131–1144, 2018.

[11] M. Hara et al., “A new phantom and empirical formula for apparent diffusion coefficient measurement by a 3 Tesla magnetic resonance imaging scanner,” Oncol. Lett., vol. 8, no. 2, pp. 819–824, Aug. 2014.

[12] A. X-ray, “Matching in 2D,” 2000.

[13] A. Goshtasby, “Transformation functions for image registration,” 2003.

[14] C.-W. Wang and H.-C. Chen, “Bioimage informatics Improved image alignment method in application to X-ray images and biological images,” vol. 29, no. 15, pp. 1879–1887, 2013.

[15] E. Irmak and M. Burak Türköz 2, “A Useful Implementation of Medical Image Registration for Brain Tumor Growth Investigation in a Three Dimensional Manner,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 6, 2017.

[16] “Intensity-based automatic image registration - MATLAB &amp; Simulink.” [Online]. Available: https://www.mathworks.com/help/images/intensity-based-automatic-image-registration.html. [Accessed: 12-Oct-2018].

[17] A. Heidenreich et al., “Guidelines on Prostate Cancer. Update,” 2011.

[18] Y. Zhang, “Image processing using spatial transform,” 2009 Int. Conf. Image Anal. Signal Process., vol. 2, no. 1, pp. 282–285, 2009.

[19] S. Uchida, Image processing and recognition for biological images, vol. 55, no. 4. 2013.

[20] P. Wittman, Lec 12: Image Registration. The Citadel, 2014.

[21] A. A. Goshtasby, “2-D and 3-D Image Registration,” Wiley Press, 2004.

[22] R. Joshi, R. Cook, and C. It, “An Analysis of Rigid Image Alignment Computer Vision Algorithms,” 2012.

[23] E. IRMAK, E. ERÇELEBİ, and A. H. ERTAŞ, “Brain tumor detection using monomodal intensity based medical image registration and MATLAB,” Turkish J. Electr. Eng. Comput. Sci., vol. 24, no. January 2016, pp. 2730–2746, 2016.

[24] L. Lo Presti and M. La Cascia, “Multi-modal Medical Image Registration by Local Affine Transformations,” 2018.

[25] R. Panda, S. Agrawal, M. Sahoo, and R. Nayak, “A novel evolutionary rigid body docking algorithm for medical image registration,” Swarm Evol. Comput., vol. 33, pp. 108–118, 2017.

[26] Y. Tanabe and T. Ishida, “Quantification of the accuracy limits of image registration using peak signal-to-noise ratio,” Radiol. Phys. Technol., vol. 10, no. 1, pp. 91–94, Mar. 2017.

[27] M. Sen Pan, J. T. Tang, and X. L. Yang, “Medical image registration based on SVD and modified PSNR,” 2011.

Using Game Theory to Handle Missing Data at Prediction Time of ID3 and C4.5 Algorithms

Halima Elaidi[1], Zahra Benabbou[2], Hassan Abbar[3]
Laboratory of Information and Decision Support Systems
Hassan 1st University Settat, Morocco

***Abstract*—The raw material of our paper is a well-known and commonly used type of supervised algorithms: decision trees. Using a training data, they provide some useful rules to classify new data sets. But a data set with missing values is always the bane of a data scientist. Even though decision tree algorithms such as ID3 and C4.5 (the two algorithms with which we are working in this paper) represent some of the simplest pattern classification algorithms that can be applied in many domains, but with the drawback of missing data the task becomes harder because they may have to deal with unknown values in two major steps: at training step and at prediction step. This paper is involved in the processing step of databases using trees already constructed to classify the objects of these data sets. It comes with the idea to overcome the disturbance of missing values using the most famous and the central concept of the game theory approach which is the Nash equilibrium.**

Keywords—Decision tree; ID3; C4.5; missing data; game theory; Nash equilibrium

I. Introduction

Machine learning is a discipline where knowledge is created automatically from raw data. Several algorithms have been developed for this purpose. This knowledge is then exploited to make decisions. Naturally, good decisions are made when data is of a good quality. Even though decision trees have proved that they are efficient classification tools, they remain, just like any other machine learning technique, helpless in front of missing data. This paper proposes to employ the concept of Nash equilibrium which is a fundamental concept of the theory of non-cooperative games with perfect information, to put an end to the disturbance caused by missing data. We only consider trees constructed by the use of algorithms ID3 or C4.5, and we suppose that these trees are perfectly constructed, the reason why our proposed method intervenes in the step of utilizing the resulting decision rules (trees) to classify new data sets containing observations with missing values.

When data is missing, it does not mean that we are allowed to ignore the corresponding records or observations. Because if we do ignore them, we are immediately causing a partial loss of information about the population we are studying through this data set. On contrary, we should treat them very accurately and try to find some useful techniques we can use to deal with missing values in a given data set. As a result researchers have developed several methods to handle this problem [1], such as:

- Deleting the records with missing values.
- Allocating the missing value of an attribute by its amount if it is quantitative or by the most frequent value if it is qualitative.
- Looking for the maximum likelihood between the records, etc.

The imputation technique this paper proposes is based on a mathematical approach which can be considered as one of the most fundamental and important discoveries of the last century: game theory.

Furthermore, the technique we are proposing can be considered as an improvement of both algorithms ID3 and C4.5 at the same time. Because, the calculations based on the Nash equilibrium that we will present later on our paper might be added as instructions or steps to the algorithm structure. As a result, for a given training data, the algorithms ID3 and C4.5 with their new structures permit to produce trees that are able to deal with data sets containing attributes with some missing values, thing which makes it possible to classify their records without any problem and with no need to look for a method among those that already exist to handle or impute the missing data.

Our document will be organized as follows: we will start by presenting the theory of decision trees and its algorithms we are interested in for this research. Then, we will introduce in details the problem to which we are proposing a solution in this paper. Next, Section 4 will be about the game theory and Nash equilibrium concepts. Section 5 will present in details the proposed imputation method, which is at the same time a way to improve the performance of the algorithms discussed in Section 2. And finally we will conclude with a brief and concise discussion.

II. Theory of Decision Trees

Decision trees constitute simple tools for decision making [2]. Actually, they are used in various fields. Their form of graphical tree representation makes of them a very simple tool, but also a very powerful one. Decision trees are the result of a set of algorithms which identify different ways of dividing a database into branches called segments, these segments form a tree characterized by a root node at the top of it.

In the same paper, Quinlan claimed "Decision-makers need to make predictions...One sound basis for such predictions is an extrapolation of past, known cases" this population of known cases is called, in the field of machine learning, the training data. It represents the principal raw material of a decision tree. In fact, a decision tree describes how to divide a population into homogeneous groups depending on the discriminant variables, since each node is just a choice on an attribute. The method used to separate the training data differs from one algorithm to another. However they all aim at making the best separation possible at each node of the tree by testing the "goodness of split" of each attribute [3].

Decision tree learning is a powerful tool and one of the most widely used and practical methods in the domain of machine learning. It is one of the supervised methods whose idea consists of classifying objects according to their characteristics or attributes and then the way these classes are formed should be used so that the resulting decision tree learn how to classify the elements of every treated data-set [4]. Several decision tree learning algorithms have been developed, but in this paper we are going to be interested in the two famous algorithms of Quinlan ID3 and C4.5. Their process of construction is based on the concept of gain (Profit or benefit): ID3 uses "Information gain" as its attribute selection measure, while the C4.5 algorithm which is the successor of ID3 uses the "gain ratio" as its attribute selection measure. Let's briefly present an overview of each of the two algorithms.

A. The ID3 Algorithm

ID3 is the well-known decision tree algorithm [5]. It is based on a recursive top-down approach; Giving a training data in which each observation is described in terms of a set of attributes, the ID3 algorithm uses the information gain as an attribute selection measure in order to separate recursively that set of examples. The information gain is calculated using the entropy:

$$E(S) = -\sum_{i=1}^{k} p_i log_2(p_i) \qquad (1)$$

Where;

- **E** is the entropy function.
- **S** the set of examples that can be divided into classes $C_1, C_2, \ldots, C_k$
- $\mathbf{p_i}$ is the probability that a set of objects from S belongs to class C_i

The resulting entropy value for a treated attribute gives an idea about its randomness or uncertainty, in a way that the attribute with the smallest entropy value is the best to use in data separation. Contrariwise, the more information gain value is important, the more the tested attribute is gainful for the separation. This property of information gain to vary in the opposite direction of variation of entropy is explained by its formula:

$$\mathrm{Gain(A,S)} = \mathrm{E(S)} - \sum_{j=1}^{n} \frac{|S_j|}{|S|} \mathrm{E(S_j)} \qquad (2)$$

- **A** is the treated attribute.
- **n** the number of possible values of attribute A.
- $\mathbf{S_j}$ are the subsets of S containing objects with the same value of attribute A.

Even though the ID3 algorithm works well in some cases, it remains powerless with attributes having a significant number of values, continuous data and missing values. These limitations of ID3 were the reason why J.Ross Quinlan developed C4.5.

B. The C4.5 Algorithm

The C4.5 decision tree algorithm [6] was developed in order to overcome the limitations of ID3 mentioned previously. Just like the ID3 algorithm, C4.5 has as a starting point a given training data, but this time the measure used to split the data is the Gain Ratio which is none other than a normalized information gain. Its formula is written as follows:

$$\mathrm{GainRatio(A,S)} = \frac{\mathrm{Gain(A,S)}}{\mathrm{SplitInfo(A,S)}} \qquad (3)$$

where;

$$\mathrm{SplitInfo(A,S)} = -\sum_{j=1}^{n} \frac{|S_j|}{|S|} \log_2\left(\frac{|S_j|}{|S|}\right) \qquad (4)$$

The present gain formula (Gain Ratio) intervenes to put an end to the weakness of information gain in front of attributes with a large number of values, because the information gain used as a splitting measure by the ID3 algorithm favors attributes with a significant number of values. Furthermore, C4.5 is said to be more efficient than ID3 in view of the fact that it is able to overcome the problem of features with continuous values as well as missing data, which is not the case for the ID3 algorithm. Another advantage of C4.5 over ID3 is that it can produce pruned decision trees. Pruning technique aims at reducing the size of a tree that over-fits the training data, which allows decreasing the prediction error rate [7].

III. Problem to be Solved

Our paper assumes that a tree is perfectly constructed using a given training set and one of the algorithms we discussed previously (ID3 and C4.5). Since decision trees are developed for the purpose of making decisions and classifying data, the produced tree can be used to classify the elements of any data set structured in the same way of the training data i.e. a data set where the objects are described using the same attributes of the training data. The set of rules established after constructing this tree may be useless if the object we are trying to classify has one or several attributes with missing values.

This work comes to remedy that problem of unknown data by using the game theory approach (more precisely, the Nash equilibrium technique). In order to fully understand the way game theory works for helping an ID3 or C4.5 algorithm to overcome the missing data problem, we propose a simple example. Thus we will consider the same example (playing tennis) already introduced by the paper "Induction of decision trees" [4]. It is an ID3 algorithm example.

The training set of our example is presented by table1

By using this data set and applying the ID3 algorithm steps on it, we obtain the decision tree of figure 1:

"N" to indicate the decision "Not play" and "P" to indicate the decision "Play".

TABLE I. A TRAINING SET EXAMPLE

Day	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
5	rain	cool	normal	false	P
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

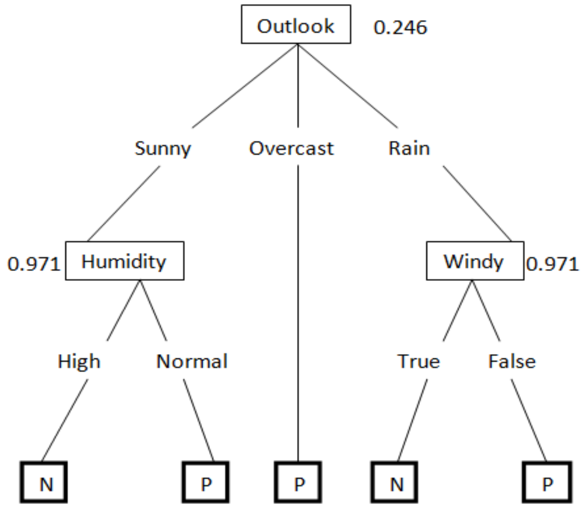


Fig. 1. The Resulting Tree using the ID3 Algorithm.

The classification rules of our example appear clearly on the established tree, we can then easily classify new observations described and defined by the use of the same set of attributes. But, suppose that while processing a data set, our classifier has encountered an observation like that of the 25th day appearing on table 2

It is absolutely clear that the decision rules of our classifier are helpless in such a case. That is why in the following parts of our paper, we will present the discipline (game theory) that will help us adjust this limitation.

TABLE II. EXAMPLE OF A RECORD WITH MISSING VALUES

Day	Outlook	Temperature	Humidity	Windy
..	..	..	..	..
25	**?**	**mild**	**?**	**false**
..	..	..	..	..

IV. GAME THEORY AND NASH EQUILIBRIUM APPROACHES

Despite of the fact that the fundamentals of game theory began to emerge earlier, this mathematical approach became more famous as a discipline only after publishing the book "Theory of Games and Economic Behavior" by J.V.Neumann and O.Morgenstern in 1944 [8]. And in spite of the presence of the word "Game" in the theory appellation, game theory remains very useful and helpful in plenty of domains which are crucial and of a major importance such as biology, economics and business, political science, engineering, computer science...and many others. The purpose of this paper is to find a solution for a problem often encountered in one of machine learning branches (decision trees).

In the field of game theory, the player is the principle element. His definition is extremely large: he can be an individual, a firm, a political party...In general, he is the decision-maker, conscious of his choices and their results, looking forward to ensure a gainful position in the game, and aware of the fact that his decisions and actions depend on those of other players i.e. he is supposed to be rational.

Thus, the main idea of game theory is to model the behavior of a set of players by observing and analyzing their strategic interactions. Usually, a game is defined by three elements: the set of players, the set of strategies (a strategy of each player is the set of his decisions) and utilities (the preference indicators of each decision) [9].

A. Mathematical Representation of a Game

First of all, note that our proposed game is a non-cooperative one. Because as we will discover later on this paper, the players of our game do not make any agreements that can bind them.

In the field of game theory, a normal form game is defined as follows:

$$G = (N, (S_i)_{i \in N}, (U_i)_{i \in N}) \quad (5)$$

where;

- **N** is the set of players (Card (N) = n).
- $\mathbf{S_i}$ is the strategies set of player i, namely the set of decisions the player i can make.
- $\mathbf{U_i}: S_1 \text{ X } S_2 \text{ X} \dots \text{ X } S_n \longrightarrow \mathbb{R}$ is the utility function of each player i (i=1,2, ..., n) called also the payoff function.

Concerning the game of this paper: The set of players is defined as the attributes with missing values of a given observation (or object). The strategy of each of the players is given by the set of values of each attribute. We will suppose that the utility or payoff of a player when making a decision, corresponds to the value of information gain (or Gain Ratio; depending on the algorithm constructing the decision tree) realized by the node that comes immediately after the branch representing the taken decision. But in this case, the utility value remains the same whatever the decisions made by the rest of players, which makes of the player a non-rational one. Because, the payoff of a rational player participating in a non-cooperative game as presented with (5) should depend not only on his own strategy, but also on the decisions of other players. Thus, for a player (attribute) making a decision, his utility is equal to the amount of information gain (Gain Ratio) provided by this decision (which is always the value of information gain mentioned on the node that comes immediately after the branch representing the decision) multiplied by a proportion that we determine with the help of the training data. Indeed, for all observations on training data that have the same decisive value for the attribute (player) in question, we must look for the proportion of records that respect each of the decisions made by the rest of the players. For instance, considering the same example presented by the tree of figure 1, assume that "outlook" and "humidity" are the attributes players of the game as shown on table 2: the utility of the attribute "outlook" when "sunny" is its decision and "normal" is the decision of "humidity" is equal to 0.971*(2/5) = 0.388. We can then conclude that the utility of a player i is calculated using the following formula:

$$\mathrm{u_i(s_i, s_{-i}) = \ G_{s_i}\ \frac{K_{s_i,s_{-i}}}{K_{s_i}}} \tag{6}$$

Where;

- $\mathbf{s_i}$ is the decision of player i.
- $\mathbf{s_{-i}}$ is representing the strategic profile of the rest of players.
- $\mathbf{K_{s_i,s_{-i}}}$ is the number of records from the training data whose attributes players have the profile of strategies $(\mathrm{s_i, s_{-i}})$.
- $\mathbf{K_{s_i}}$ is the number of records from the training data where player i plays his strategy si.

Intuitively, $\mathrm{G_{s_i}}$ is equal to 1 when the successor node of the branch representing the decision si of the player i is a leaf, because the information is fully provided. On the other hand, if an attribute does not appear at least in one of the classification rules of the established tree, $\mathrm{G_{s_i}}$ will be equal to 0 for all the possible values of this attribute, i.e. the attribute does not provide any information.

B. Nash Equilibrium

The notion of Nash equilibrium can be considered as the most brilliant as well as influential game theoretical concept that was invented by the "beautiful mind" John Nash. It is defined as a stable situation, where each player (from the set of players in interaction) is not ready to deviate of his decision. Because if he does, while the rest of players are keeping their strategies, his utility will immediately decrease, thing which is not gainful for a rational decision-maker (or simply player) [10].

1) Pure nash equilibrium: The normal form of a game as it is previously presented (5) is considered in the field of game theory as a game with pure strategies. The Nash equilibrium (the pure strategy Nash equilibrium) for a set of players participating in such a game is given by a profile of strategies $(\mathrm{s_1^*,\ s_2^*, \dots,\ s_n^*})$ where each $\mathrm{s_i^*}$ is representing the best decision made as a response to other players' strategies [9, 10]. Mathematically, this can be written as follows:

For all players i = 1, 2, ..., n

$$\mathrm{u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*)\ ;\ \forall s_i \in S_i} \tag{7}$$

where

$$\mathrm{s_{-i} = (s_1, s_2, \dots, s_{i-1}, s_{i+1}, \dots, s_n)}$$

2) Mixed Nash equilibrium: The concept of mixed strategy can be adjudged to be the generalization of pure strategy, which comes to provide a much clearer vision about the real behavior of a rational player. A mixed strategy is simply a pure strategy associated to a distribution of probability with which each player is making his choice.

So, for a set of pure strategies Si, let $\Delta(\mathrm{S_i})$ denote the set of probability distributions over it, so that:

$$\mathrm{P_i \in \Delta(S_i) \Longrightarrow \begin{cases} \mathrm{P_i: S_i \longrightarrow [0,1]} \\ \quad \mathrm{s_i \longrightarrow P_i(s_i)} \\ \sum_{\mathrm{s_i \in S_i}} \mathrm{P_i(s_i) = 1} \end{cases}} \tag{8}$$

Thus Pi which is a probability distribution over Si represents the mixed strategy of player i.

Therefore, in a case where the decisions of players are in the form of mixed strategies, the mixed Nash equilibrium is defined as a profile of mixed strategies$(\mathrm{P_1^*, P_2^*, \dots, P_n^*})$, such as:

For all players i=1, 2,..., n

$$\mathrm{u_i^{mixed}(P_i^*, P_{-i}^*) \geq u_i^{mixed}(P_i, P_{-i}^*); \forall P_i \in \Delta(S_i)} \tag{9}$$

Where; $\mathrm{u_i^{mixed}}$ is the expected utility function defined as follows:

$$u_i^{mixed}(P_i) = \ \sum_{s_i \in S_i}\left(\prod_{j \in N} P_j(s_j)\right) u_i(s_i) \tag{10}$$

V. Proposed Method

As it is already mentioned at the beginning of the paper, our work assumes that a decision tree is constructed using one of the famous algorithms of Quinlan (ID3 or C4.5). It also supposes that the stage of construction did not face any obstacle. As a matter of fact, the present work comes to fix a problem which appears during the use of the established tree for classifying new data, more precisely data with missing values. Our method assumes that, for a given observation, the imputation of its missing values is a strategic game of the form (5). According to the theorem of existence of the Nash equilibrium [10], the proposed game accepts at least a Nash equilibrium in mixed strategies, because the number of players is finite as well as the number of strategies for each player i.

Since the construction of decision trees using ID3 and C4.5 algorithms is a task which is mainly based on the theory of information, we then had the idea of using the same theory to impute the missing values of a data set at prediction time. In fact, Quinlan's splitting measures were based on the information theory of shannon in order to find how well each attribute by its own classifies the records of the training set, then the attribute with the highest value of information gain or gain Ratio is the one that can generate the best partition i.e. that attribute is providing the best quantity of information. Hence the inspiration of making from this concept of information gains a useful tool to handle missingness at prediction process.

The idea is to impute missing values of each observation by values maximizing the quantity of information, respecting of course the structure and the characteristics of the training data; to that end, we are suggesting the use of the game theory approach every time when encountering a data entry where at least two attributes' values are missing: each of those attributes is represented as a player and the strategy set of each one consists of the possible values in the range of the attribute. The payoffs correspond to that quantity of information we cited previously. As a matter of fact, the Nash equilibria would yield "balanced" ways of substituting the missingness: all values used for imputation had the same objective which is "maximization of information gain".

Using the example presented in section 3 (table 1), let's assume that the established decision rules (tree) (figure 1) are used to classify the elements of a given database containing an observation where the values of attributes "outlook" and "humidity" are missing (table 2). We can look for the Nash equilibrium in pure strategies for this game with two players: outlook and humidity. Notwithstanding this equilibrium point does not exist all the time, but if it does exist, the values of the corresponding attributes represent the ones that should be used to impute the missing values. If a Nash equilibrium in pure strategies does not exist, then we proceed to a Nash equilibrium in mixed strategies which always exists, depending on the theorem of Nah we discussed above.

It is to highlight that the job becomes difficult when the problem of missingness concerns continuous attributes (This difficulty is encountered only while working with the C4.5 algorithm). But as it is widely known, C4.5 converts continuous values to nominal ones by proposing to perform binary splits based on a threshold value. As a matter of fact two intervals should be obtained: [minimum value, threshold] and]threshold, maximum value]. Then we propose using the centers of those intervals to impute missing values.

A. Imputation by the use of Pure Nash Equilibrium

Always in the case of our explanatory example, the payoff matrix of the game is given by table 3:

According to the rule of determining the pure Nash equilibrium (7), we can deduce that this game admits two equilibrium profiles which are (high, sunny) and (normal, rain) ("humidity" is the first player and "outlook" is the second one). As a result, for the observation of table 2, attributes outlook and humidity can take respectively the values (high and sunny) or (normal and rain).

TABLE III. THE PAYOFF MATRIX OF THE GAME

	Outlook			
		sunny	**overcast**	**rain**
Humidity	**normal**	(0.286, 0.388)	(0.286, 0.5)	(0.429, 0.583)
	high	(0.429, 0.583)	(0.286, 0.5)	(0.286, 0.388)

TABLE IV. THE PAYOFF MATRIX OF THE NEW GAME

	Player 2 (Attribute 2)		
Player 1 (Attribute 1)		**C**	D
	A	(0.7 , 0.97)	(0.88 , 0.1)
	B	(0.81 , 0.5)	(0.6 , 0.92)

B. Imputation by the use of Mixed Nash equilibrium

In this subsection, we will treat another case in which we are confronting a situation where the game does not accept any Nash equilibrium in pure strategies, the reason why we are forced to look for Nash equilibrium in mixed strategies. Giving an example remains the best way to explain a technique. Thus, for a given training data (different from the one we worked with previously), assume that by using once again the ID3 (or C4.5) algorithm, we got a new tree and certainly different values of the players' utilities. Then suppose that while working with the obtained tree for the purpose of classifying new data, we have encountered an observation with two attributes whose values are missing. Therefore, the first thing to do is to construct the payoff matrix, which is given by table 4:

It is quite clear that the game in question does not admit a Nash equilibrium in pure strategies, thereby we will look for the mixed equilibrium. In fact, note α as the probability with which player 1 plays the strategy "A" and $(1 - \alpha)$ the probability with which he plays strategy "B". Similarly, player 2 plays strategy "C" with probability β and strategy "D" with probability $(1 - \beta)$.

According to these probability values, the expected utility of player 1 is written as follows:

- $0.7\ \beta + 0.88\ (1 - \beta) = 0.88 - 0.18\ \beta$; if player 1 chooses to play strategy A.
- $0.81\ \beta + 0.6\ (1 - \beta) = 0.6 + 0.21\ \beta$; if player 1 chooses to play strategy B.

Accordingly:

$u_1(P_1, P_2) = \alpha\ (0.88 - 0.18\ \beta) + (1 - \alpha)\ (0.6 + 0.21\ \beta)$

$= (0.6 + 0.21\ \beta) + \alpha\ (0.28 - 0.39\ \beta)$

Which is a function increasing in α if $(0.28 - 0.39\ \beta) > 0$ and decreasing if $(0.28 - 0.39\ \beta) < 0$.

Consequently, the strategy A constitute the best response of player 1 in mixed strategies if and only if $\beta < 0.72$, while B is the best response of player 1 in mixed strategies if and only if $\beta > 0.72$, but when β=0.72 he still indifferent between the two strategies.

Similarly, the expected utility of player 2 is written as follows:

- $0.97\ \alpha + 0.5\ (1 - \alpha) = 0.5 + 0.47\ \alpha$; if player 2 chooses to play strategy C.
- $0.1\ \alpha + 0.92\ (1 - \alpha) = 0.92 - 0.82\alpha$; if player 2 chooses to play strategy D.

Thus;

$u_2(P_1, P) = \beta\ (0.5 + 0.47\ \alpha) + (1 - \beta)\ (0.92 - 0.82\ \alpha)$

$= (0.92 - 0.82\ \alpha) + \beta\ (1.29\ \alpha - 0.42)$

Which is a function increasing in β if $(1.29\ \alpha - 0.42) > 0$ and decreasing if $(1.29\ \alpha - 0.42) < 0$. Therefore, C is the best response of player 2 in mixed strategies if and only if $\alpha > 0.33$ and D is the best response of player 2 in mixed strategies if and only if α < 0.33. But in the case where α = 0.33 the player 2 is indifferent between the two strategies C and D.

The equilibrium of a game in mixed strategies is established when the players are indifferent in their choices of strategies. Concerning our example, the equilibrium is presented as a profile of probabilities: [(0.33 , 0.67) ; (0.72 , 0.28)] where player 1 chooses strategy A with a probability of 0.33 and strategy B with a probability of 0.67, and similarly player 2 chooses 72% of the time the strategy C and 28% of the time the strategy D.

VI. Discussion and Conclusion

Just like any other machine learning algorithm, techniques used for classification tasks are all the time facing the problem of missing data. In fact, in real data applications, the presence of missing data is a general and challenging problem [11]. A decision tree classifier may encounter this problem in two contexts: values may be missing in the training data (at induction time) or while predicting the classes to which new records should belong (at prediction time) [12]. Concerning the method we are proposing, it aims at handling missing values at prediction time and it only concerns two types of algorithms constructing decision trees: ID3 and C4.5.

Specialists in the field of data missingness insist on the necessity of making assumptions about what caused the data to be unknown. Thereby, they identify three categories or types of missing data:

- **MCAR**: (Missing Completely At Random) refers to data that were collected randomly which means that for an observation where a feature's value is missing, that missingness does not depend on any variable of the data set.
- **MAR**: (Missing At Random) requires that the cause of missingness is not related to the unknown feature while it could be conditional on some of the rest of variables in the data set.
- **NMAR**: (Not Missing At Random) this case takes place when the missingness is not random and depending on the actual value of the missing data.

"When data are MCAR or MAR, the missing data mechanism is termed ignorable" [13], the approach we are proposing does not require any prior knowledge about the reasons of data missingness i.e. we are assuming data to be MCAR or MAR. Classification approaches and methods have proved their usefulness in many problem domains. However, they have to deal with the problem of missing data which is a common drawback when solving a real life classification task. A wide range of techniques were elaborated to handle the limitations caused by unknown values such as:

- Case deletion
- Mean imputation
- Multiple imputation
- Hot and cold deck imputation
- Maximum likelihood

As their being classifiers, ID3 and C4.5 can use such approaches to face unknown values while processing a data set for prediction. Of course each one of these methods has its advantages and disadvantages. But logically, each classification technique has some specific characteristics related to the way the classifier should be constructed, from this point we thought that the approach adopted by every classifier to handle the limitations of missing data should respect the characteristics and essential elements of the classifier. This work concerns two well-known classification algorithms (ID3 and C4.5) that are based on the notion of "quantity of information gained", for that reason a method respecting this notion should be used when handling data missingness. Thereby, our proposed method consists of maximizing the gain of information while imputing unknown values in treated observation: our method covers all the features of only one record at the same time. Thus, working on a data set for classification purpose, we propose to handle the unknown values as a first step by the use of our method based on game theory approach (as seen on section 5), then the classifier can be applied to determine the class of each record.

Note that we are proposing a method under two forms: imputation using pure Nash equilibrium and imputation using mixed Nash equilibrium. The first one seems to be easier, but the second one is the most relevant as it always gives results and it is more realistic. In fact, records having exactly the same features whose values are unknown will use the same form of game to impute missing data. And as it was already mentioned, the solution in mixed strategies comes in the form of probabilities. For instance, assume that an attribute is part of a game and that attribute can take two possible values A and B, at the Nash equilibrium in mixed strategies, A can be the best decision with a probability of P% and B can be the best decision with a probability of (1 - P)%. We can then conclude that for all records utilizing the same game, A will be assigned to P% of these observations and B will be assigned to (1 - P)% of that observations (N.B for an observation, when only one variable is missing, we assign to it the value with the maximum payoff).

In general, features are considered the fundamental elements for constructing classifiers such as ID3 and C4.5 and they still unchanged while processing data sets. Assume that we are working on a data set with N variables, the number of possible games that can be used to deal with missing data is$\sum_{k=1}^{N-1} C_N^k$; it is a finite number of cases. Consequently, it is a step which can be added to both of algorithms as an improvement of algorithms ID3 and C4.5.

References

[1] J.L. Schafer, J.W. Graham, "Missing Data: Our View of the State of the Art" Psychological methods 2002, Vol. 7, N° 2, American Psychological Association, PP. 147-177.

[2] J.R. Quinlan, "Decision trees and decision-making" IEEE Trans. Syst. Man Cybern. 1990 Vol. 20, N° 2, PP 339-346, DOI: 10.1109/21.52545.

[3] L. Breiman, J. Friedman, R. Olshen, C. Stone, "Classification and regression trees" Pacific Grove, CA: Wadsworth 1984.

[4] J.R. Quinlan, "Induction of Decision Trees" Machine Learning 1986, 1, 81-106.

[5] J.R. Quinlan, "Discovering rules by induction from large collections of examples". In D. Michie (Ed.), Expert systems in the micro electronic age 1979, Edinburgh University Press.

[6] J.R. Quinlan, “Comparing connectionist and symbolic learning methods”. In S. Hanson, G. Drastal, And R. Rivest (Eds.), Computational Learning Theory and Natural Learning Systems: Constraints and Prospects 1993, Cambridge, MA: MIT Press.

[7] H. Elaidi, Z. Benabbou, H. Abbar, “A comparative study of algorithms constructing decision trees: ID3 and C4.5” LOPAL’18 Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications 2018, DOI: 10.1145/3230905.3230916.

[8] J. von Neumann, O. Morgenstern, “Theory of Games and Economic Behavior” Princeton University Press, Princeton, New Jersey, 1944.

[9] M.J. Osborne, “An Introduction to Game Theory” Oxford University Press, Oxford, 2002.

[10] J. Nash, “Non-cooperative games” The Annals of Mathematics, Second Series 1951, Vol. 54, N° 2, PP. 286-295.

[11] Q. Song, M. Shepperd, “Missing Data Imputation Techniques” Int. J. Business Intelligence and Data Mining 2007, Vol. 2, N° 3, 261-291.

[12] M. SAAR-TSECHANSKY, F. PROVOST, “Handling Missing Values when Applying Classification Models” Journal of Machine Learning Research 2007, 8, 1625-1657.

[13] P.J. Garcia-Laencina, J.L. Sancho-Gomez, A.R. Figueiras-Vidal, “Pattern classification with missing data: a review” Neural Comput and Applic 2010, 19, 263–282, DOI: 10.1007/s00521-009-0295- 6.

Optimizing Power-Based Indoor Tracking System for Wireless Sensor Networks using ZigBee

Ahmad H. Mahafzah
Department of Computer Science
Faculty of Information technology
Middle East University Amman, 11831 Jordan

Hesham Abusaimeh
Associate Professor of Computer Science
Faculty of Information Technology
Middle East University Amman, 11831 Jordan

***Abstract*–Evolution of wireless and digital communication gives birth to the smaller but powerful battery-equipped devices which are easy to maintain and perform the desired tasks. ZigBee is a Wireless Personal Area Network (WPAN) used for home or indoor automation, collecting data for medical research by using the low power digital radios to handle the low data rate. In ZigBee network, sensor nodes are heterogeneously deployed and continuously moving. To detect and tracking of those sensor nodes are challenging in terms of accuracy, calculation time and energy consumption. In this paper, proposed system uses the Received Signal Strength Indication (RSSI) protocol for localization, trilateration for fetching the exact coordinates of sensor nodes and these protocols helps to overcome the problem which eventually led to prolonged sensor network, accurate localization.**

Keywords—Indoor tracking; ZigBee; wireless personal area network; localization; trilateration

I. INTRODUCTION

WSN consists of multiple heterogeneous and homogeneous sensor nodes [1]. The basic difference between these two are consists in their basic feature i.e. initial power, processing capability, signal strength etc. These nodes are used to gather the data from the surrounding. Usually, the nodes in range share data with each other to communicate better and transmit the gathered data to Sink Node (SN). Sink node is used as the intermediary or gateway node between the WSN and the researchers. As suggested by [2], sensor nodes store the data processing and communication protocols to process them when required.

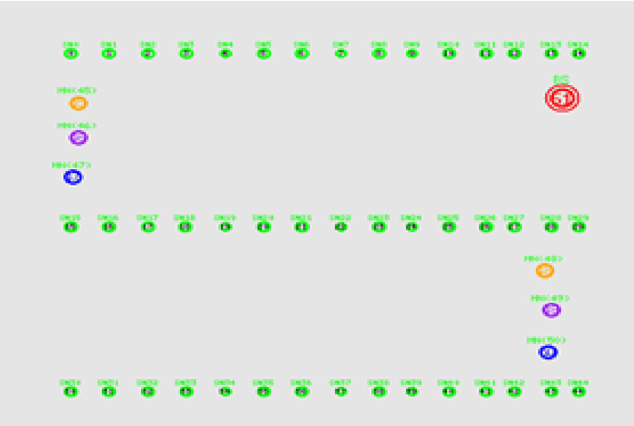


Fig. 1. ZigBee Wireless Sensor Network.

Sensor nodes are randomly deployed in a WSN as shown in figure 1. Each sensor node has a responsibility to collect data from sensor field, share it with other and finally route the collected data to the base station or sink node. As having a mesh network, collected data will be routed back to end user by a multi-hop infrastructure. And in the end, it is the responsibility of sink node to send the data to the user.

Each sensor node consists of battery, sensor unit, memory unit, transceiver (omniAntenna) and processor unit. These parameters have their own unique function to perform in the network. Battery-equipped sensor nodes are used to provide the power; sensor unit uses the omniAntenaa for transceiver for sensing the environment and collect the desired data packets. After processing the data, memory unit will help them to keep the data safe. As sensor nodes has small size memory unit, they cannot save it for long and sensor nodes have to override the memory segments which may cause data loss. So, nodes will transmit the data packets to sink node very frequently to save the data and avoid the anomalies [3].

The WSN is available to provide many services like indoor tracking mobile object through ZigBee network standard. ZigBee network provides remarkable communication protocols using low range and low-power digital radios based on standard IEEE 802.15.4 for Wireless Personal Area Networks. It is originally designed to provide short-range communication services that provides lower data rate. ZigBee is developed on the network and application layer in IEEE 802.15.4 standard. Topologies is considered as an acceptable three main of configuration, they are tree, star and mesh topologies. It is built on top of IEEE 802.15.4 standard, which defines the characteristics of the physical and Medium Access Control (MAC) layer on protocol stack for WPAN [4].

Node localization is defined as determining the position of a sensor node in the concerned network area with respect to origin. Localization techniques for WSN consist of the algorithms that estimate the locations of sensors with initially unknown position information normally using available information about the absolute positions of a few other sensors and position measurements. Sensors with known location information are called beacons or anchors. The anchors define the local coordinate system to which all other sensors are referred. The coordinates of the sensors with unknown location information also called blind or non-anchor nodes, it is estimated by various sensor network localization techniques. The most popular method in distance estimation for wireless systems is the Received Signal Strength Indicator (RSSI) technique, which is based on the physical fact of wireless communication that theoretically, the signal strength is inversely proportional to the squared distance between the transmitter and receiver [5].

II. Related Work

Chen et al. [6] presented a survey with reference to RSSI solutions on indoor localization and proposed a Closer Tracking Algorithm (CTA) to locate a mobile user in the house. The proposed CTA was implemented by using ZigBee CC2431 modules. The proposed CTA shows at least 85% precision when the distance is less than one meter.

Larranaga et al. [7] discussed different environmental factors that affect the RSSI values measured to calculate the position of the device. Proposed system consists of two main phases: calibration and localization. The use of a central processing server allows the implementation of complex algorithms, while the ZigBee network allows collecting signal level values and at the same time, it is used to provide data to the central server for localization computation. To locate blind nod the system performs the calibration, so that changes in the environment are taken into account in the localization phase, and thus making the system more robust and accurate.

Chu et al. [8] discussed the two novel techniques to improve tracking in indoor system using ZigBee by improving positioning accuracy. Authors combined the Neighbour Area Majority Vote Priority Correction and Environment Parameter Correction. The experiment results demonstrate that proposed methods can largely increase accuracy of ZigBee positioning and provide useful personnel tracking technology.

Chandane et al. [9] analysed performance of IEEE 802.15.4 using Qualnet 4.5. Author designed star topology, multi-hop peer to peer network. MANET routing protocols such as AODV, DSR and DYMO are used for analysis of Quality of Service (QoS) parameters like throughput, packet delivery ratio, average end-to-end delay, jitter, total energy consumption, and network scalability as the performance metrics. Results show that AODV outperforms other two in star topology whereas DSR has slightly upper hand in multi-hop topology for varying traffic loads and in beacon-enabled mode.

Obaid et al. [10] briefly explained an overview of ZigBee technology and its application in wireless home automation systems. The performances of the ZigBee based systems have also been compared with those of other competing technologies-based systems. In addition, some future opportunities and challenges of the ZigBee based systems have been listed Oracevic &Ozdemir (2014) targeted different target tracking methods via focusing on security. Author explained the important protocols in each category and described which security properties they provide. Paper presented that most of the research in target tracking focused on the accuracy or energy efficiency and there is limited amount of work that considers security. Author presented a table that summarizes the state of the art in the target tracking area.

Vancin & Erdem [11] in this study utilize the IEEE 802.15.4/ZigBee, which has advantages than other types of technology with respect to parameters such as; use of the battery in addition to low consumption of power. In this study, OPNET simulator is used to achieve the required results. The behaviour of mobile node and network fixed has been compared based on the quality of the end-to-end delay parameters and traffic received through the destination.

Shinde et al. [12] created a study in order to develop a low cost, indoor location positioning system that can be utilized to find the indoor moving position of the object. This paper portrays a novel low-cost system for two diminution (2D) locations monitoring utilizing indoor ZigBee Technology through location fingerprinting technique to estimate position. Use of Mesh networking of ZigBee assesses in making the network more scalable and also in increasing the coverage area of the system.

TABLE I. Simulation Parameters

Serial No.	***Parameter Name***	***Parameter Value***
1.	Area of flat grid	400m*400m
2.	No. of sensor nodes	45
3.	No. of mobile nodes	6
4.	No. of Base Station	1
5.	MAC Layer Protocol	MAC/802.15.4
6.	Simulation time	100 s
7.	Antenna type	OmniAntenna
8.	Radio model	TowRayGround
9.	Transmission range	30m
10.	Routing protocol	AODV
11.	energyModel	100 Joules
12.	T_x power	1.0 w
13.	R_x power	1.0 w
14.	DataSize	1000 Byte
15.	DataRate	5K bits/s

III. Experimental Design

Existing system lacks the accuracy while fetching the coordinates of sensor nodes and very prone to choke the battery very fast. Proposed model removes these errors and drawbacks by the combination of trilateration, localization and energy consumption. Trilateration is used to get the locations of different sensor nodes and store the coordinates into a trace file. This will help the sensor network to know the initial location of nodes. After initiating the simulation, mobile nodes 45to 47 will start moving randomly towards the sink and mobile nodes 48 to 50 away from the sink node. This will provide robust scenario of mobility in both directions. Just after initiating the simulation to transfer data packets, a hello packet is broadcasts to every node in the network and simultaneously RSSI procedure starts working. RSSI computes the distance between mobile node and sensor nodes and stores them into a trace file at every 0.5 seconds time interval. The rest of simulation parameters are defined such as mentioned in the following Table 1.

IV. Algorithm

The proposed model uses the localization and distance techniques to get the exact location and distance among them and storing the location coordinates and distance into a trace file.

Figure 2 describes the working of proposed model. These two tables are used to send the data from mobile nodes to sink node.

To transfer the data uninterruptedly, they use the nearby sensor nodes as intermediates and send it. Now sink node has the sensed data of network and the location of nodes that is stored in different tables.

Proposed Algorithm

1. Node N starts up.

2. Start mobility of Mobile nodes (from 45 to 50).
   $MN = \{node\_N\ \ X_{coordinate}\ \ Y_{coordinate}\ \ Speed\}$

3. Get location using trilateration

4. Calculate Distance between sensor node and Mobile node

   $$d = \sqrt{(\mathrm{X2}-\mathrm{X1})^2 + (\mathrm{Y2}-\mathrm{Y1})^2}$$

   Where *X1* and *X2* is X-axis of sensor nodes and *Y1* and *Y2* are Y-axis of sensor nodes

5. Build distance table

6. Calculate remaining energy (*RemEng*) of mobile node $Pt = PtInit \times \frac{\mathrm{RemEng}}{InitEng}$

   $$\mathrm{RemEng} \quad = Pt \times \mathrm{GTX} \times \mathrm{GRX}\left(\frac{\lambda}{4\pi d}\right)^2$$

   Where *Pt* is transmission power; *PrInit* is the initial reception power, *RemEng* is the remaining energy of the node, and *InitEng* is the initial full energy of the node

7. Build Residual Energy table.

8. Compute the Received Signal Strength Indicator (RSSI). $RSSI = 10 \times LOG\frac{\mathrm{RemEng}}{\mathrm{PRef}}$, $[RSS = dBm]$

   Where *RemEng* is remaining energy, *PRef* is reference power.

9. Update distance table

10. Start data transmission between nodes.

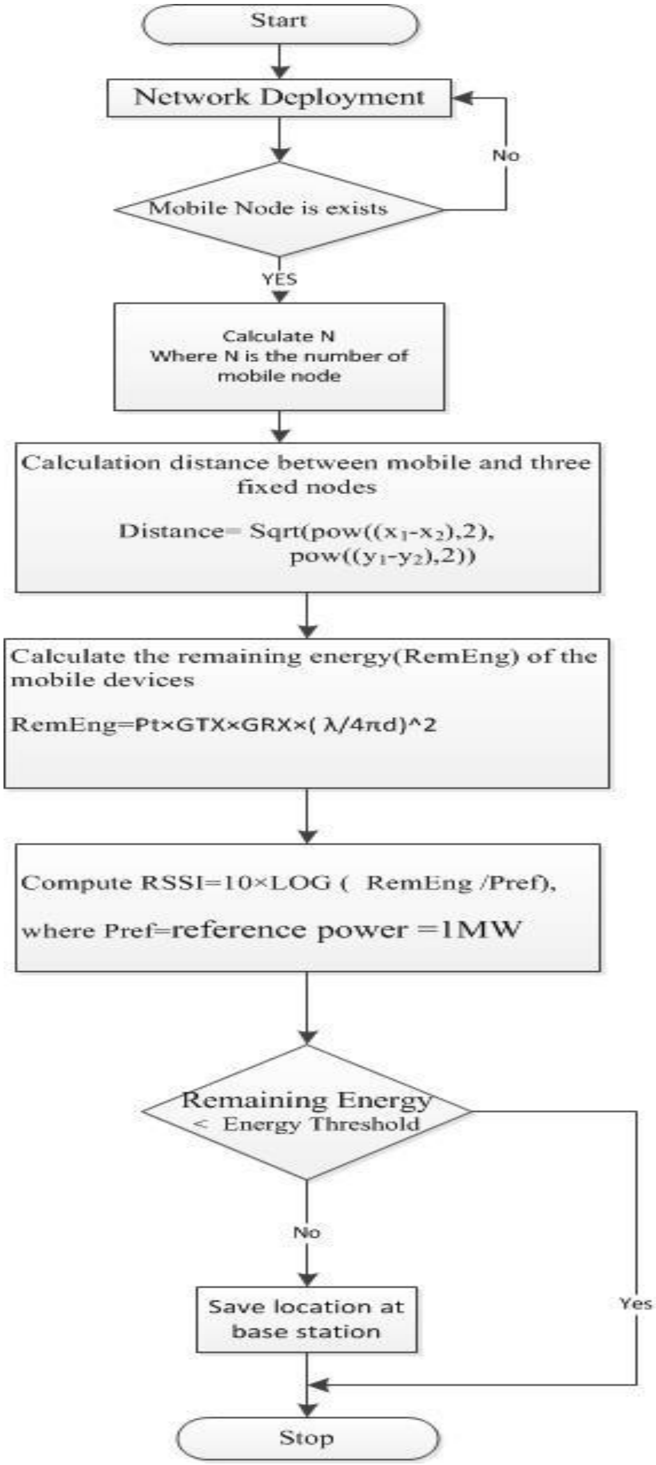

Start
Network Deployment
No
Mobile Node is exists
YES
Calculate N
Where N is the number of mobile node
Calculation distance between mobile and three fixed nodes
Distance= Sqrt(pow((x1-x2),2), pow((y1-y2),2))
Calculate the remaining energy(RemEng) of the mobile devices
RemEng=Pt×GTX×GRX×(λ/4πd)^2
Compute RSSI=10×LOG (RemEng /Pref),
where Pref=reference power =1MW
Remaining Energy < Energy Threshold
No
Yes
Save location at base station
Stop

Fig. 2. Flowchart of Proposed Model.

V. Results

The proposed model has been implemented using the Network Simulator version 2.35 (NS2). There are 52 sensor nodes deployed randomly in the wireless network where nodes ranging from 0 to 44 are sensor nodes, nodes ranging from 46 to 50 are mobile nodes and node 51 is programmed as sink node. Where sensor nodes are static in nature and mobility is defined randomly. The sensing and transceiver range is fixed to 30 meter as the property of ZigBee.

Table 2 describes the average results and performance evaluation of existing and proposed systems. This clearly shows that the proposed system is better in every aspect and provides more reliable sensor network.

TABLE II. Performance Evaluation of Proposed and Existing System

Parameters	***Proposed system (average)***	***Existing system (average)***	***Improvement (approx.)***
Tracking Error	0.73	0.79	8%
Throughput(bit/second (bps))	91095	63020	45%
Delay(millisecond(ms))	2	3.1	36%
PDR(bit/second (bps))	2107.1	1400	73%
Residual Energy (joules)	74.7	69.9	7%

A. Throughput

Throughput can be computed in Gbps, Mbps and Kbps. This shows the successfully transferred number of packets from source to destination.

Figure 3 shows that the packets throughput during the simulation is very high as compared to existing system because of the localizing the sensor nodes. This allows them to know the exact location of each node and reduce the time and connection packets to delivery.

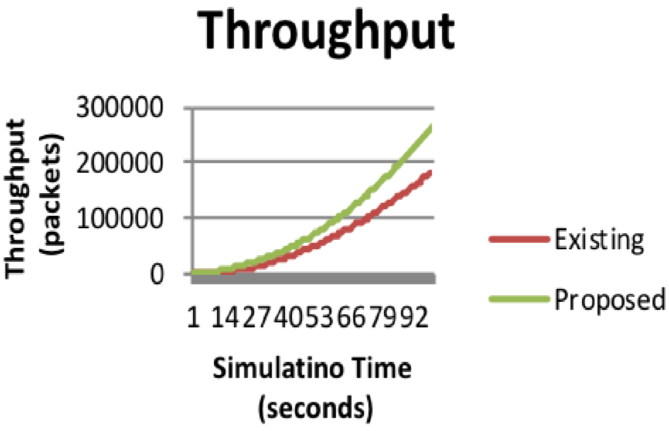


Fig. 3. Throughput of Proposed System.

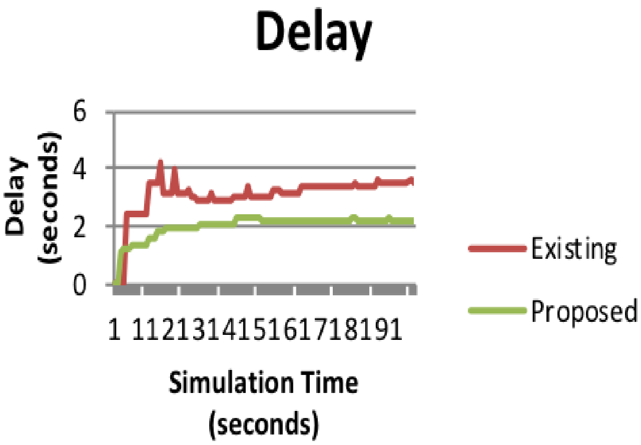


Fig. 4. Delay of Existing Proposed System.

B. Delay

Delay shows the delay of packets from source to destination. This can be caused because of queuing of packets, route discovery, low transmission rate etc. only successful data packets are considered during simulation.

Figure 4 shows that the delay packets is reduced by the proposed model and controlled after a sudden hike into it. The delay can be caused by multiple factors like sending sensing signals to gather information about neighbours, storing them into a table, parsing the table to get the nearest node to transfer data bytes.

C. Packet Delivery Ratio

The packet delivery ratio can be defined as the ratio of packets successfully received to the total sent. Figure 5 shows the efficiency of the network.

Proposed system has high Packet delivery ratio than existing because of their applied algorithms that proposed system don’t have to create connection every time they want to send packets. There are distance and connection trace files available eases the path formation procedure.

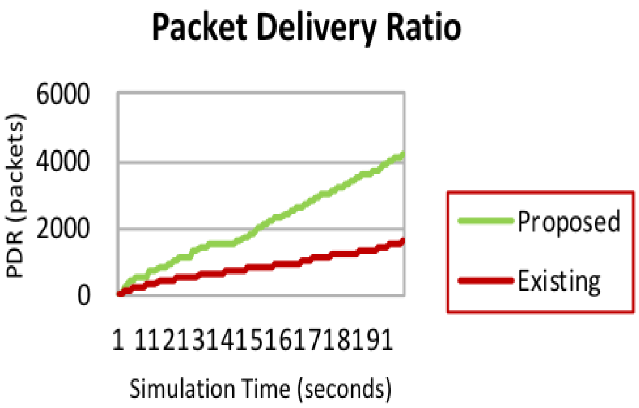


Fig. 5. Packet Delivery Ratio of Proposed System.

D. Energy Consumption

Residual energy is the remaining energy of any sensor node. Figure 6 will show that the lifetime of node or any network.

The topography patter of both system are different nature and existing system used less number of sensor nodes and the deployment of nodes are also half the size of proposed system. Due to this topography and nodes deployment of both systems, we can deduce that proposed system can survive twice the existing system.

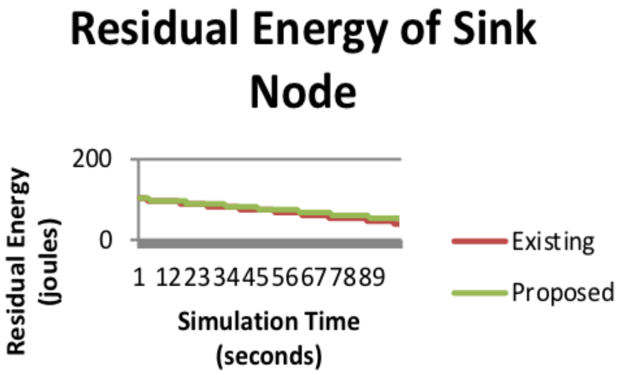


Fig. 6. Residual Energy of Sink Node.

E. Accuracy

Accuracy of finding the location it is measured by taking the averaged Euclidean distances between the real coordinates and estimated coordinates. It presents the uncertainty of having the correct target's location.

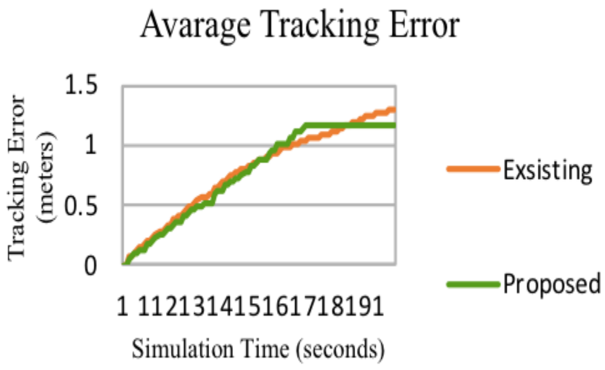


Fig. 7. Average Accuracy of all Mobile Nodes.

The accuracy of finding the distance is reduced by considering the trilateration algorithm. Figure 7 shows the accuracy of proposed system increases gradually because of the proposed algorithm.

VI. Conclusion

This thesis presents Indoor tracking using ZigBee Wireless Sensor Network. The proposed algorithm is designed after considering all the loopholes which may affect the environment. Heterogeneous sensor nodes are deployed in the ZigBee wireless sensor network to measure the effectiveness and efficiency of the proposed system. The proposed system can divide the implantation of algorithm into mainly three parts i.e. ensuring the availability of sensor nodes while connection and path formation by considering the residual energy of sensor node and store the energy into a trace file; fetching the locations of sensor nodes, mobile nodes and sink node using trilateration technique and store the location into a trace file and in the end, using Received Signal Strength Indicator (RSSI) technique determine the distance between each and every sensor node, mobile node and sink node and store the distance into a trace file. This process iterates itself after every 0.5 seconds. The trace tables of energy and distance are updated and provide the exact figures throughout the simulation. The proposed model is designed to eradicate the drawbacks of the existing system by merging the numerous algorithms together to develop a new and improved solution. The proposed model improved the indoor tracking systems lifetime using ZigBee wireless sensor network. This will also help to prolong the existence of network and work to sense and transmit data for a long time. The proposed model works on the principle of collecting information first rather than sending it. The calculation of numerous parameters like localization or distance and building the centralized trace table for the sensor nodes to provide the identical information to each node helps the proposed system to prolong the lifetime of the sensor network and more importantly increases the accuracy of the system. To get the assurance of working efficiency, effectiveness and accuracy of the proposed model, it is examined for various parameters like throughput, delay, residual energy and accuracy. The combination of examined collected data, trilateration and RSSI for tracking the indoor scenarios makes it more adaptable and effective than others.

A performance result of this method has been measured and compared with the original RSSI method using NS-2 simulator. All the achieved simulation results have shown better performance for the proposed methods as compared to the original method. The Tracking Error improved by 8% and power consumption improved by 7%. Furthermore, the new network performance measurements have obtained a good value compared with the other measures in term of the network performance measurements such as packet delivery ratio (PDR), delay and throughput improved by 73%, 36% and 45%, respectively.

Acknowledgment

The authors are grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fees of this research article.

References

[1] M Wu, Wenlan, Xianbin Wen, Haixia Xu, Liming Yuan, and Qingxia Meng. "Efficient range-free localization using elliptical distance correction in heterogeneous wireless sensor networks." International Journal of Distributed Sensor Networks 14, no. 1 (2018): 1550147718756274.

[2] Sohraby, Kazem, Daniel Minoli, and Taieb Znati. Wireless sensor networks: technology, protocols, and applications. John Wiley & Sons, 2007.

[3] Akyildiz, I.F., & Vuran, M.C., (2002), Wireless sensor networks, 1st ed, UK: John Wiley & Sons.

[4] Abusaimeh, H., & Yang, S. H. (2012). Energy-aware optimization of the number of clusters and cluster-heads in WSN. In Innovations in Information Technology (IIT), 2012 International Conference on (pp. 178-183). IEEE.

[5] Abusaimeh, h., (2014), "Balancing the Network Clusters for the Lifetime Enhancement in Dense Wireless Sensor Networks" in Arabian Journal for Science and Engineering, 2014. DOI 10.1007/s13369-014-1059-x.

[6] chen, y., & yango, c., (2006), A RSSI-Based Algorithm for Indoor Localization Using ZigBee in Wireless Sensor Network, Int. J. Digit. Content Technol. it's Appl. Digit. Content Technol. its Appl., 5(7), 407–416.

[7] Larranaga, J.,& Muguira, L.,& jyyh & Lopez-Garde, J. M. & Vazquez, J. I. ,(2010), An Environment Adaptive ZigBee-Based Indoor Positioning Algorithm, 2010 Int. Conf. Indoor Position, Indoor Navi.

[8] Chu, C. H., & Wang, Liang, C. H. C. K. W. & Ouyang, J. H. & Cai, & Chen, Y. H., (2011), High-accuracy indoor personnel tracking system with a ZigBee wireless sensor network, 7th Int. Conf. Mob. Ad-hoc Sens. Networks, MSN 2011.

[9] Chandane, M. M., (2012), Performance Analysis of IEEE 802 . 15 . 4, Int. J. Comput. Appl., 40(5), 23– 29.

[10] Obaid, T.,& Abou-Elnour, A., & Rehan, M., & Muhammad Saleh, M. & Tarique, M., (2014), Zigbee Technology and Its Application in Wireless Home Automation Systems: a Survey, Int. J. Comput. Networks Commun, 6(4), 115–131.

[11] VANCIN, S. & ERDEM, E., (2015), Design and Simulation of Wireless Sensor Network Topologies Using the ZigBee Standard. International Journal of Computer Networks and Applications. 2(3), 125-143.

[12] Shinde, V. & Panchal, R. & Panchal, J., (2016), ZigBee based indoor location tracking system. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. 5(4).

Heterogeneous Ensemble Pruning based on Bee Algorithm for Mammogram Classification

Ashwaq Qasem[1], Shahnorbanun Sahran[3]
Centre of Artificial Intelligence,
Faculty of Information Science and Technology,
Universiti Kebagsaan Malaysia,
43650 Bangi, Selangor, Malaysia.

Siti Norul Huda Sheikh Abdullah[2]
Center for Cyber Security,
Faculty of Information Science and Technology, Universiti
Kebagsaan Malaysia, 43650 Bangi, Selangor, Malaysia

Dheeb Albashish[4]
Computer Science Department, Prince Abdullah Bin Ghazi
Faculty of Information Technology,
Al-Balqa Applied University, Jordan

Rizuana Iqbal Hussain[5], Shantini Arasaratnam[6]
Department of Radiology,
Universiti Kebangsaan Malaysia Medical Centre,
56000, Cheras, Kuala Lumpur, Malaysia[5]
Department of Radiology, General Hospital of Kuala
Lumpur, Malaysia[6]

***Abstract*—In mammogram, masses are primary indication of breast cancer; and it is necessary to classify them as malignant or benign. In this classification task, Computer Aided Diagnostic (CAD) system by using ensemble learning is able to assist radiologists to have better diagnosis of mammogram images. Ensemble learning consists of two steps, generating multiple base classifiers and then combining them together. However, combining all base classifier in the ensemble model increases the computational cost and time. Therefore, ensemble pruning is an important step in ensemble learning to select the ensemble's members. Due to huge subsets of combination in the ensemble, selecting the proper ensemble subset is desirable. The objective of this paper is to select the optimal ensemble subset by using Bee Algorithm (BA). A pool of different classifier models such as Support vector machine, k-nearest neighbour and linear discriminant analysis classifiers have been trained using different samples of training data and 12 groups of features. Then, by using this pool of classifier models, BA was used to exploit and explore random uniform combination subsets of the trained classifiers. As a result, the best subset will be selected as the optimal ensemble. The mammogram image dataset that was used to test the model has been collected from Hospital Kuala Lumpur (HKL) and consists of 263 benign and malignant masses. The proposed method gives 77 % of Area Under Curve(AUC), 83% of accuracy, 93% of specificity and 60% of sensitivity.**

Keywords*—Ensemble learning; ensemble pruning; bee algorithm; mammogram; breast cancer

I. Introduction

Breast cancer is the most widely recognized dangerous cancer among ladies and the second driving reason for death [1]. According to the study from the World Health Organization, breast cancer is the most diagnosed cancer after lung cancer (10.9% of cancers in men and women is breast cancer), and the second driving reason for death [1], i.e., in 2008, 458 503 women worldwide died because of breast cancer [2].

Although no effective ways to prevent breast cancer exist, the early discovery of breast cancer is deemed significant for the decrease of associated death. Therefore, considerable effort is focused on cancer diagnosis at early stages. One of the most useful tools in early detection and diagnosis of breast cancer is the imaging technologies and mammography [3].

Mammography is the popular technique designed to image the breast [3]. In mammography, the most critical symptoms of breast cancer are masses. Mass type diagnosis (benign or malignant) in mammography images is a challenging issue for radiologists.

CAD helps in diagnosis and identification of masses. Typical CAD system consists of three stages as shown in Fig. 1; Pre-processing, which includes image segmentation, feature extraction, and classification into normal, benign, and malignant. In image segmentation stage, a mammogram will be segmented to extract Region of Interest (ROI). ROI maybe extracted manually [4]–[9] or it can be extracted automatically using any segmentation method [10]–[14].

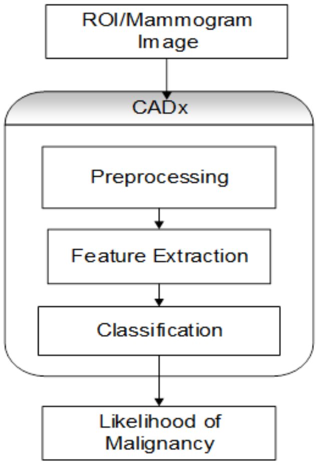


Fig. 1. General Steps of CADx.

The second stage is the feature extraction, which plays an important role for achieving high performance in the classification stage. This can be only achieved by extracting the suitable features that describe whether the suspicious region is benign or malignant. One of the most popular types of extracted features is the texture features.

In classification stage, the appropriate classifier model is trained using the training samples with the extracted features and then used to predict the class of the unseen pattern. Classifier design in the classification stage of CAD system is one of the key steps to get higher performance. Most of researchers have been focus in using a single classifier such as the Support Vector Machine (SVM) [6], [7], [15], Multi-Layer Perceptron (MLP) [5], [7] , Linear Discriminant Analysis (LDA) [4], [7], Decision Tree (DT) [6], and K-Nearest Neighbour (KNN) [6]. However, due to the heterogeneity in the mammogram, single classifiers make errors on different samples. Thus, the ensemble learning in machine learning is used to improve the diagnosis of the mammogram. It has shown its ability in solving different classification problems as histopathology image grading [16]–[18], intrusion detection system [19] and breast cancer detection and diagnosis [20], [21].

Ensemble learning has become important because of its ability to improve the performance of single classifier system in theory and practice [22]. Generally, ensemble model consists of a group of independent trained base classifiers that are aggregated together in order to classify new samples. Fusing the outputs of all these base classifiers to get a final ensemble output is based on the aggregation rule. Using the appropriate base classifiers helps to improve the overall ensemble results. Fig. 2 is showing the different between single classifier model and ensemble learning.

Combining all base classifiers in the ensemble model will increase the computational cost and time. Using subset of the ensemble can outperform the whole ensemble. Therefore, selecting the member of ensemble subset (ensemble pruning) is an important step.

Although, choosing the member of ensemble is an important issue, the challenge here is how to decide the criteria for selecting ensemble members [23]. This means, to have a successful ensemble, balancing between diversity and accuracy of the ensemble members should be achieved. To choose the best ensemble, optimization methods like genetic algorithm, artificial bee colony have been implemented [19], [24]–[27]. Fig 3 is showing ensemble selection using GA.

Bee Algorithm (BA) was firstly proposed by Pham et al. [28].It is a population based algorithm inspired by natural behaviour of food searching of honeybee. Generally, BA has three important steps; initialization, local search, and global search. At the beginning, scout bees are distributed in random uniform in the search space starting to explore it. Then, recruited bees start exploiting and searching in the most promising areas identified by the scout bees. BA was applied and showed a good performance to problems in many fields, such as multi-level thresholding [29].

In this study, optimization of ensemble learning is emphasized. Two criteria of ensemble are used; 1) The generation of multiple heterogeneous base classifier, so that diversity among them is ensured, and 2) The use of BA optimization method in pruning ensemble.

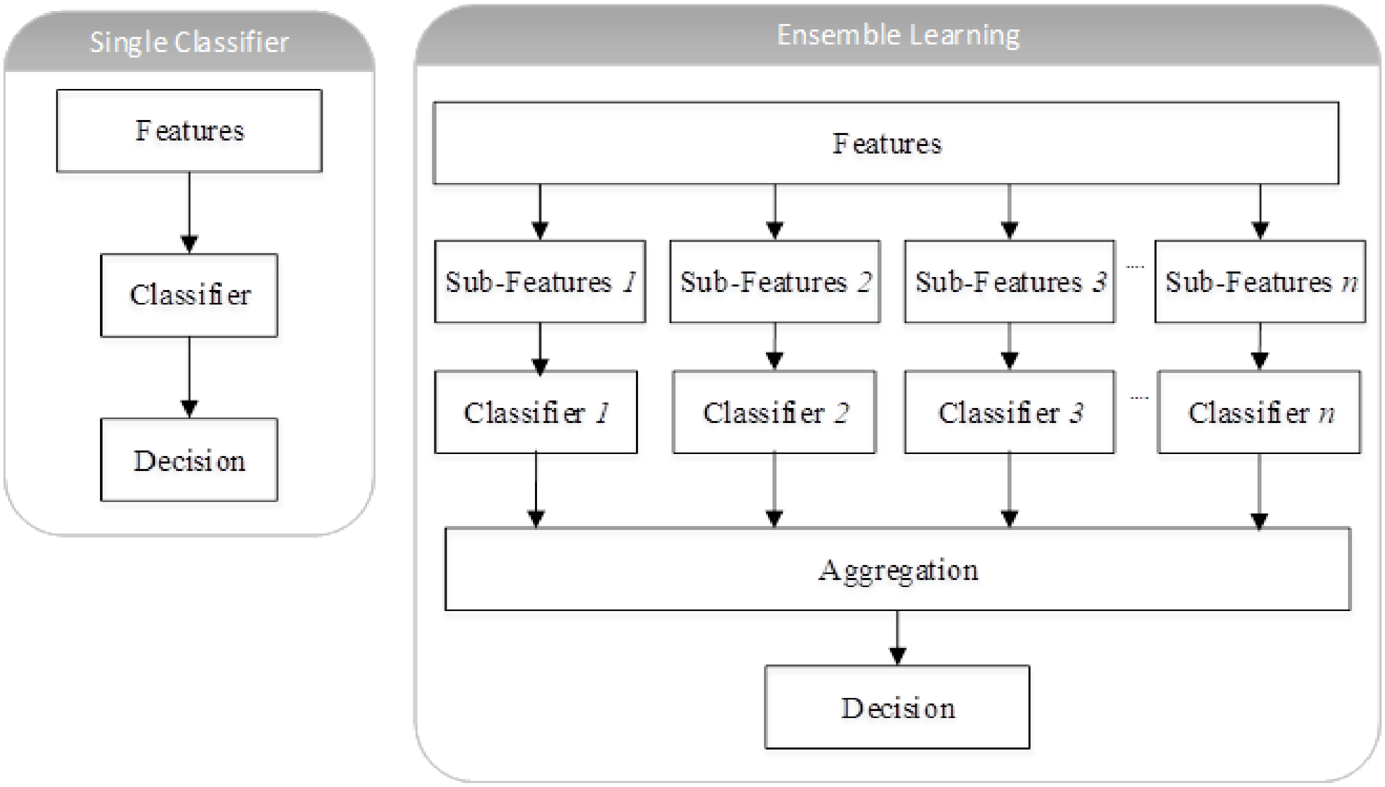


Fig. 2. Ensemble Learning Vs Single Classifier.

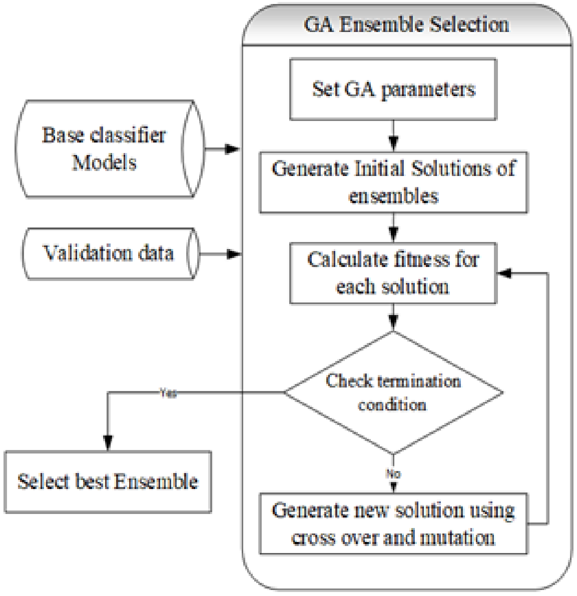

GA Ensemble Selection
Set GA parameters
Base classifier Models
Generate Initial Solutions of ensembles
Validation data
Calculate fitness for each solution
Check termination condition
Yes
No
Select best Ensemble
Generate new solution using cross over and mutation

Fig. 3. Ensemble Selection using GA.

II. Material and Method

In this paper an ensemble pruning method using bee algorithm for breast cancer classification is proposed. First of all, the mammogram Region of Interest (ROI), where the mass is in its centre, is manually cropped and resized into 512×512. Different features extracted from each ROI. Then, about twelve categories of features has been extracted as shown in Table I.

The extracted features were used to build the ensemble model. Our ensemble classifier system consists of three phases; pool generation, ensemble pruning, and ensemble testing. The key here is to select the optimal number of ensemble committee members by using optimization method. BA is used in order to select the best number of classifier that used to build an ensemble that ensure to achieve the better performance

TABLE I. Features Groups

	Features	**No. of Features**
1	Gray Level Co-Occurrence Matrix (GLCM)	38
2	histogram of oriented gradients (HOG)	81
3	Local Binary Pattern (LBP)	256
4	First Order Statistics (FOS)	5
5	Haralick Spatial Gray Level Dependence Matrices (SGLDM)	25
6	Gray Level Difference Statistics (GLDS)	4
7	Neighborhood Gray Tone Difference Matrix (NGTDM)	5
8	Statistical Feature Matrix (SFM)	4
9	Laws Texture Energy Measures (TEM)	6
10	Fractal Dimension Texture Analysis (FDTA)	4
11	Fourier Power Spectrum (FPS)	2
12	Shape	2

A. Pool Generation

Firstly, best base classifier is generated using all groups of features. Unlike Choi et al. [21], who depended on single base classifier in the generation process, multi-based classifier was considered. Fig. 4 shows how to generate multiple based classifier.

In details, dataset, after extracting features, is divided into three sets; 1) training set, which is used in generation phase; 2) validation set, which is used in the selection phase, and; 3) testing set, which is used to evaluate the selected ensemble.

In the generation phase, feature groups, as shown in table I, were used to build classifier model using different based classifiers. SVM, KNN, and LDA are considered here as base classifiers. SVM has been chosen due to its robustness in handling overfitting when there is no balancing between the training sample and features**.** KNN which is considered as a lazy classifier, is used here because it is simplicity and low bias, while LDA is good for correlation.

Assuming we have k group of features and n base classifiers then k×n classifier models are built. Then the best model is selected at each iteration based on the data sampling process. The outcomes of this phase are M best classifier models which will be the input of the pruning phase.

Let F represents the features pool $F = \{f_i, i = 1, \dots k\}$, C represents the classifier pool $C = \{c_i,\ c = 1 \dots n\}$, T represents the training samples $T = \{(x_i, l_i)\}_{i=1}^{N}$, where class label is $l_i \in \{0,1\}$ and N is the number of training sample, and M is the number of boosting rounds.

Based on Freund and Schapire [30], the initial distribution $D_0(x_i) = 1/N$, whereas for each m-th round, the distribution $D_m(x_i)$ on the training sample x_i can be calculated as follow:

$$D_m(x_i) = \frac{w_{m,i}}{\sum_{i=1}^{N} w_{m,i}}\ for\ i = 1 \dots N \tag{1}$$

where $w_{m,i}$ is the weight for the i-th training sample at the *m-th* boosting round.

We apply random resampling with Uniformed training data for obtaining different subsample in each M round.

Let $C_{m,k,n}$ is the n-th base classifier c_n trained with k-th feature f_k and m-th resample subset. So for each m round we are going to select the best base classifier C_m which gives the most accurate result among the weighted training set as follow:

$$C_m = min_{h_{m,k,n}} \varepsilon_{h_{m,k,n}} \tag{2}$$

$$\varepsilon_{C_{m,k,n}} = \sum_{i=1}^{N} D_m(x_i) \left| C_{m,k,n}(x_i) - l_i \right| \tag{3}$$

where, D is a weighted distribution, N number of training samples, l is the training label and C is the classifier output.

The idea here is that different classifier trained by different features will make the mass that cannot be recognized by one representation easily detected by another. Using multiple features with different base classifiers increases ensemble diversity between ensemble members and avoid making coincident errors.

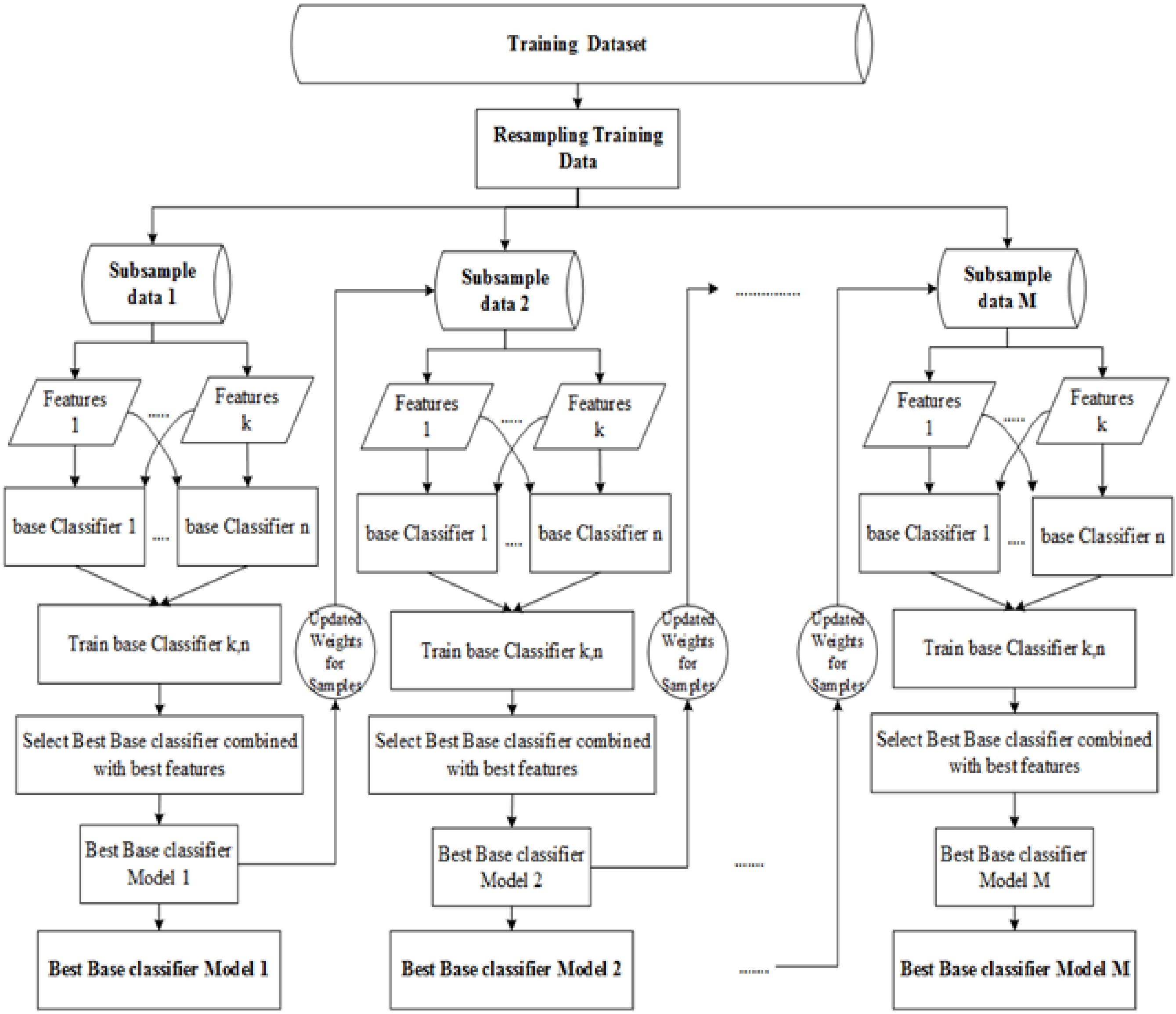

Training Dataset
Resampling Training Data
Subsample data 1
Subsample data 2
Subsample data M
Features 1
Features k
base Classifier 1
base Classifier n
Train base Classifier k,n
Select Best Base classifier combined with best features
Best Base classifier Model 1
Best Base classifier Model 2
Best Base classifier Model M
Updated Weights for Samples

Fig. 4. Adapted Ensemble Generation.

$$\beta_t = \varepsilon_t/(1-\varepsilon_t) \tag{4}$$

where ε is the error of the best classifier.

$$a_t = \frac{1/\beta_t}{\sum_{t=1}^{M} 1/\beta_t} \tag{5}$$

B. Ensemble Pruning (Selection)

In the selection phase, M best classifier models will be used to build the ensemble using the validation dataset. BA was used for selecting models to be included in the ensemble. The output of this phase is the best ensemble solution and it will be evaluated using the testing data set in the testing phase.

Fig. 5 illustrates the combination the generated based classifiers to build the ensemble models and to select the optimal ones.

Assuming that there are n trained based classifier in the pool, therefore, the search space is 2^n. Corresponding to that, BA was used.

1) BA for ensemble pruning

The first step in BA is the initialization via randomly uniform. This initialization returns continuous values between [0,1]. To convert it to discrete, thresholding is applied. For each scout bee, initialization as an ensemble randomly created by combining various number of based classifier. This random combination is done based on the discretization process for the random generated numbers. Fig. 6 shows the steps of BA. For example, initial position $ipos$ is created by generating n random numbers in between [0,1], where n is the pool size. Then, discretization is applied using threshold $0 < \text{Th} < 1$ using equation 6:

$$DePos_i = \begin{cases} 0 & ipos < Th \\ 1 & ipos \geq Th \end{cases} \tag{6}$$

where $DePos_i$ is the descritized position of i based classifier, $i \in [1,2,3,\ldots.n]$, $ipos$ is the initial position and Th is the threshold value.

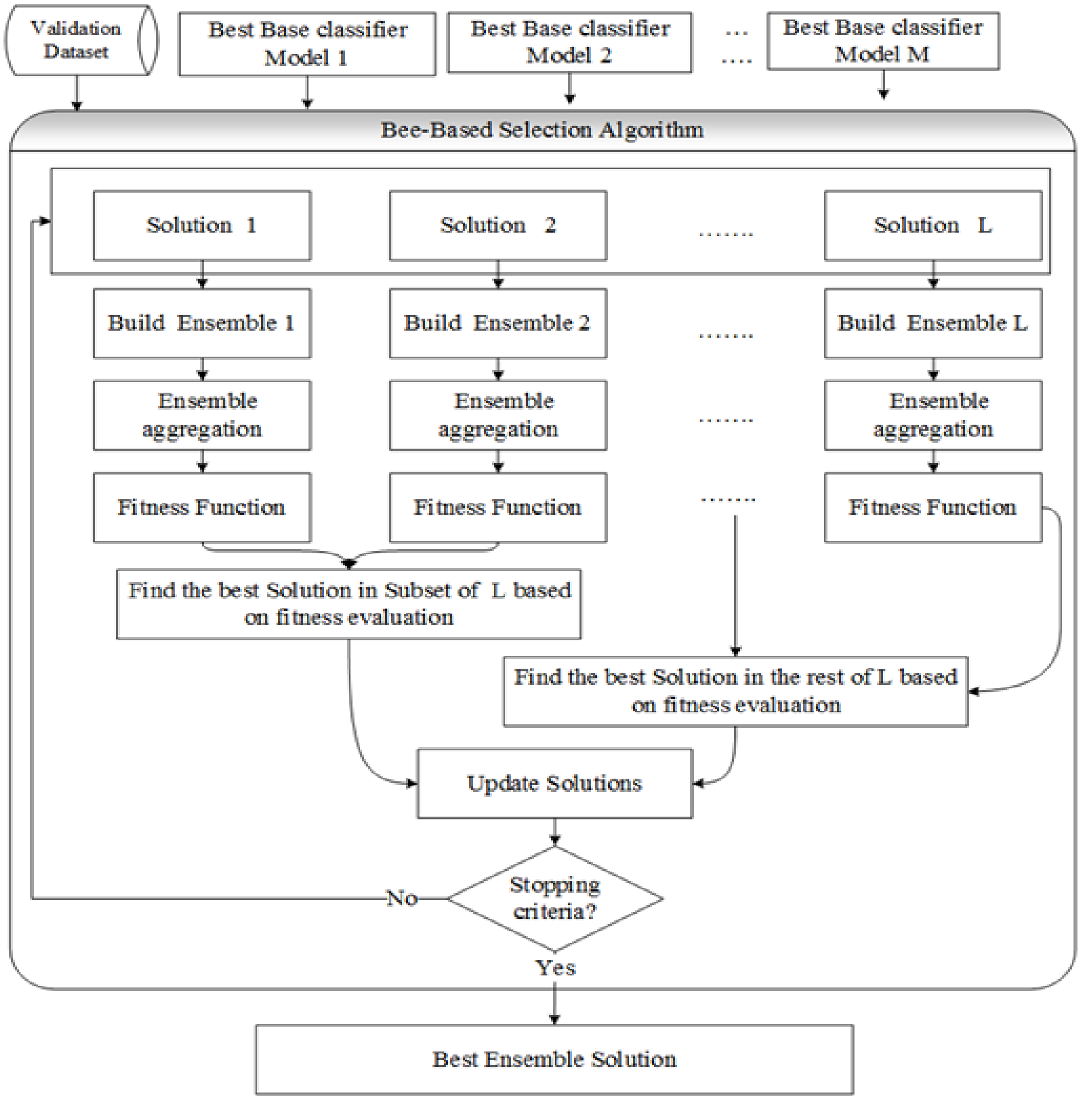
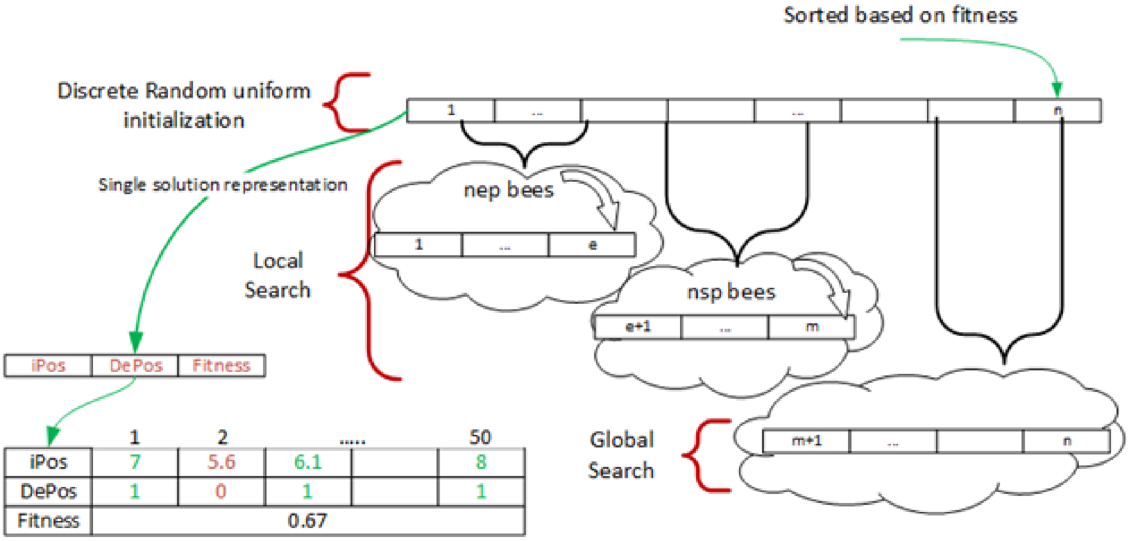


Fig. 5. The Proposed Ensemble Pruning using BA.



	1	2		.....	50
iPos	7	5.6	6.1		8
DePos	1	0	1		1
Fitness	0.67				

Fig. 6. The Proposed Ensemble Pruning Process using BA, where N is Number of Scout Bees, M is the Selected Sites, E is the Elite Sites, Nsp is Recruited Bees for Selected Sites and Nep is Recruited Bees for Elite Sites.

Eventually, The ensemble is created based on the $DePos$ vector. For each based classifier in th pool C_i

$$C_i \rightarrow \begin{cases} Part\ of\ ensmeble & DePos_i = 1 \\ Out\ of\ ensemble & DePos_i = 0 \end{cases} \tag{7}$$

So that the initial ensemble for each scout bee can be define as below

$$Initial\ Ensemble = \{C_i\}_{i=1}^{M} \tag{8}$$

where.$DePos_{i}{}_{i=1}^{M} = 1$, M is a variable number representing the number of ensemble members.

After initial ensemble is created, fitness function of each ensemble is calculated based on the performance of the ensemble. The performance of the ensemble is considered by both AUC and diversity of the ensemble. As stated in literature, there is no unique measure for diversity, in this work, calculation of ensemble fitness is done based on Choi at el. [21] equation as follow:

$$Ensemble\ finness = \overline{AUC} - \lambda\overline{Div} \tag{9}$$

where λ is the controlling variable, $\overline{AUC}$ and $\overline{Div}$ are calculated as:

$$\overline{AUC} = (\mathrm{AUC} - AUC_{min})/(AUC_{max} - AUC_{min}) \tag{10}$$

$$\overline{Div} = (\mathrm{Div} - Div_{min})/(Div_{max} - Div_{min}) \tag{11}$$

where, AUC_{min} and AUC_{max} are the minimum and maximum AUC of ensemble members, Div_{min} and Div_{max} are the minimum and maximum diversity between pairs members in the ensemble. AUC is calculated as:

$$AUC = AUC_{ens} + (AUC_{ens} - AUC_{mean}) \tag{12}$$

where, AUC_{ens} is the ensemble AUC based on Majority Voting (MV) aggregation and,

$$AUC_{mean} = \frac{1}{M}\sum_{i}^{M} AUC_i \tag{13}$$

where, M is the number of selected based-classifier in the ensemble. AUC_i the value of AUC obtained using *i-th* classifier.

Div is calculated based on three diversity pair wised measures, $Q\ statistic$, correlation coefficient and double-fault as follow:

$$Div = Q_{avg} + \rho_{avg} + DF_{avg} \tag{14}$$

where, Q_{avg} , ρ_{avg}, and DF_{avg} are the average values of the $statistic$, correlation coefficient and double-fault respectively and could be calculated as:

$$(Q,\rho,DF)_{avg} = \frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(Q,\rho,DF)_{\mathrm{i,j}} \tag{15}$$

TABLE II. CONTINGENCY TABLE

	C_i **correct**	C_i **wrong**
$C_j$correct	a	b
C_j wrong	c	d

The Q,ρ and DF are calculated between two paired classifiers C_i, and C_j based on Table II. The contingency table is used when calculating diversity between two classifiers i,j for the same testing data. The values a,b,c and d have different meaning where a is the number of instance in the testing data that correctly classified by both classifier i,j, while d is the number of instance in the testing data that are incorrectly classified by both classifier i,j; c is the number of instance in the testing data that are correctly classified by classifier i and misclassified by classifier j, and b is the number of instance in the testing data the misclassified by classifier i and correctly classified by classifier j.

The $Q\ statistic$ for two classifiers C_i andC_j is:

$$Q_{t,m} = \frac{\bar{a}\bar{d}-\bar{b}\bar{c}}{\bar{a}\bar{d}+\bar{b}\bar{c}} \tag{16}$$

where, $\bar{a}$ is the propablity that both classifiers classify the instance correctly. $\bar{b}$, $\bar{c}$ and $\bar{d}$ can be defined in the same way:

$$\bar{a} = {}^{a}/_{N}\, ,\ \bar{b} = {}^{b}/_{N}\, ,\ \bar{c} = {}^{c}/_{N}\, ,\ \bar{d} = {}^{d}/_{N},$$

where, N, is number of testing instance.

While, the correlation of two classifiers C_i andC_j is:

$$\rho_{t,m} = \frac{\bar{a}\bar{d}-\bar{b}\bar{c}}{\sqrt{(\bar{a}+\bar{b})(\bar{c}+\bar{d})(\bar{a}+\bar{c})(\bar{b}+\bar{d})}} \tag{17}$$

The double-fault measures the misclassified instances by both classifiers C_i C_j. It is calculated as below:

$$DF_{t,m} = \bar{d} \tag{18}$$

After initialization of BA, local and global searches are started. The output of BA is the best *k* ensembles where *k* is the number of BA generation.

C. Ensemble Testing

After the selection process is finished, multiple solution is presented. The final solution is identified as the best among the best. This final ensemble will be tested using the test dataset. The aggregation of ensembles is done using MV.

III. Result and Discussion

A. Dataset Description

The dataset used here is a self-collected dataset from Hospital Kuala Lumpur (KHL). It is consisting of 236 mammogram ROI images. All images have either benign or malignant mass. The expert radiologist has gone through all images to evaluate them and identify the correct place of mass in each.

B. Experiment Setup

The experiment was run 10 times to see the stability of the model. In each run, double cross validation is applied to the data with k=15. In the generation stage, the pool size*(M)* was set to 50. So that 50 base classifiers are generated. Sampling with r=30 is applied to the training data.

The based classifier parameters for both single classifier approach and ensemble approach were similar. First, SVM tested by LibSVM toolbox with radial basis function kernel.

C and γ was calculated using grid search with range $[2^{-5}, 2^{10}]$ [31]. Then, KNN classifier was used with the default k =1. Regulation used in LDA with γ interval [0,1].

For BA pruning ensemble, parameters set to their default and most used values; 10 iterations, 30 scout bees, 15 selected site, 12 elite sit, 15 recruited bees for selected sites and 30 recruited bees for elite sites. The threshold value (Th) for discretization of BA initialization was set to 6. GA parameters are set as A. Ekbal and S. Saha [27] with population size of 50 and generation is set to 100. Same fitness function used for GA and BA based on performance and diversity of the ensemble as equation (9) with controlling variable $\lambda = 0.25$ as mentioned in Choi et al. [21]. MV is used in ensemble aggregation.

C. Evaluation Matrices

To evaluate the proposed method, confusion matrix was calculated in order to calculate the accuracy, specificity and sensitivity.

Sensitivity is the ability of the model for detecting malignant masses, and it can be calculated based on the following equation:

$$Sensitivity(SE) = \frac{TP}{TP+FN} \times 100 \tag{19}$$

where TP is the number of corrected malignant diagnostics instance and FN is the malignant instance that was wrongly diagnosed as benign.

Specificity is representing the ability of the model to identify the benign masses, and it is calculated as:

$$Specificity(SP) = \frac{TN}{TN+FP} \times 100 \tag{20}$$

where, TN is the number of benign instance that classified correctly and FP is the the number of benign instance that was misclassified.

The accuracy of the model is describing its ability of detecting both benign and malignant.

$$Accuracy(AC) = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \tag{21}$$

On the other hand, AUC calculated the area under Receiver operating characteristic ROC curve, which draws the relation between sensitivity rate and specificity rate.

D. Result of Proposed Heterogeneous BA Ensemble Pruning

The objective of this experiment is to show the output of the proposed ensemble. Table III shows the average result with standard deviation of 10 run of the proposed BA heterogeneous ensemble pruning using different aggregation method.

TABLE III. RESULT OF PROPOSED BA ENSEMBLE PRUNING FOR ONE RUN USING DIFFERENT AGGREGATION METHOD

	MV	**Product**	**Max**	**Sum**	**Min**	**Average**
AUC	0.79 ±0.13	0.79 ±0.08	0.79 ±0.8	**0.80** ±0.12	0.79 ±0.08	**0.80** **±0.12**
AC	**0.85** **±0.09**	0.74 ±0.09	0.74 ±0.09	**0.85** ±0.09	0.74 ±009	**0.85** ±0.09
SP	**0.94** **±0.05**	0.66 ±0.13	0.66 ±0.13	**0.94** ±0.06	0.66 ±0.13	**0.94** ±0.06
SE	0.64 ±0.25	**0.92** ±0.09	**0.92** ±0.09	0.66 ±0.24	**0.92** ±0.09	0.66 ±0.24

TABLE IV. EXAMPLE RESULT OF THE ENSEMBLE MEMBER AUC, CLASSIFIER AND FEATURE GROUP OF THE PROPOSED BA ENSEMBLE PRUNING OF ONE CROSS VALIDATION ITERATION

	Ensemble Members														
	Mem 1	Mem 2	Mem 3	Mem 4	Mem 5	Mem 6	Mem 7	Mem 8	Mem 9	Mem 10	Mem 11	Mem 12	Mem 13	Mem 14	Mem 15
Single AUC	0.86	0.90	0.90	0.95	0.90	0.80	0.90	0.90	0.90	1.00	0.90	0.90	0.90	0.90	0.90
classifier	LDA	LDA	SVM	KNN	LDA	LDA	LDA	SVM	KNN	KNN	LDA	LDA	LDA	LDA	SVM
Feature	SFM	GLDS	LBP	LBP	TEM	GLCM	TEM	LBP	GLDS	GLCM	GLCM	TEM	GLDS	GLCM	LBP

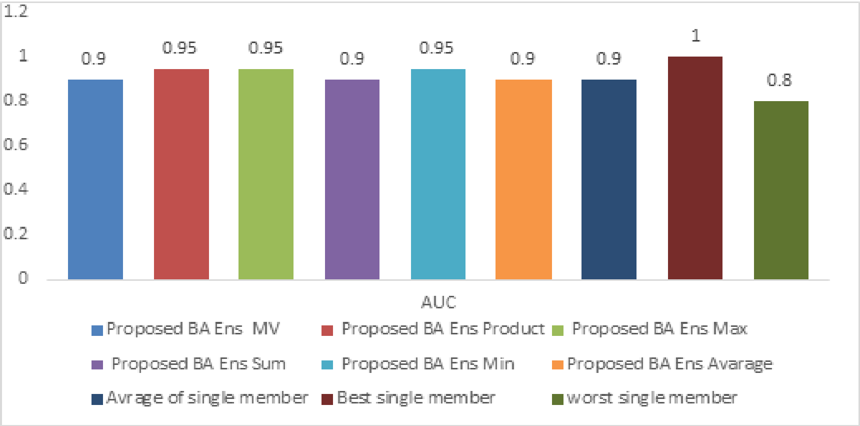


Fig. 7. Comparison between AUC of Proposed BA Ensemble and its Members AUC of a One Cross Validation Iteration.

Table IV is showing an example of an ensemble member after applying the proposed BA pruning method.

We can notice in Table IV the ensemble contains various classifiers as well as features groups. It might be noticed that there is a repetition in term of similar classifier using the same feature group. However, each classifier was actually trained by different training samples. The subsampling steps applied in the generation process guaranteed that all the trained classifiers inside the pool are different from each other's in terms of classifier model or feature group or training sample. This variation will maintain the diversity of the ensemble members.

Moreover, Fig. 7 is comparing with the AUC of the proposed ensemble framework using the 6 aggregation method compared to the AUC of ensemble members; best, average, and worst. The performance of the proposed BA ensemble can outperform the average of its members.

As this is heterogeneous ensemble, Fig. 8 shows comparison between the proposed ensemble frameworks with single classifier framework. The single classifier framework was built by using the three based classifiers used in the ensemble; SVM, KNN, and LDA. For the single classifier, all features used in the ensemble were combined together. The proposed BA ensemble outperforms the three single classifiers in terms of average AUC and AC.

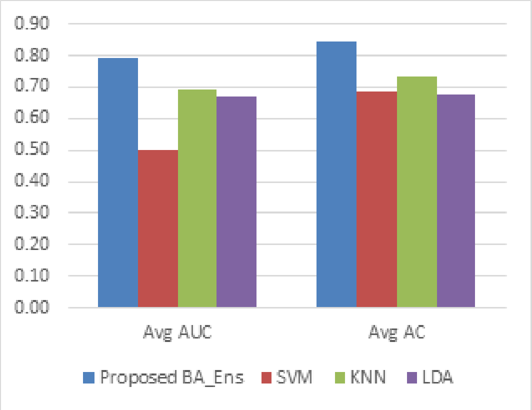


Fig. 8. Comparison between the Result of the Proposed BA Ensemble Pruning Framework and Single Classifier System.

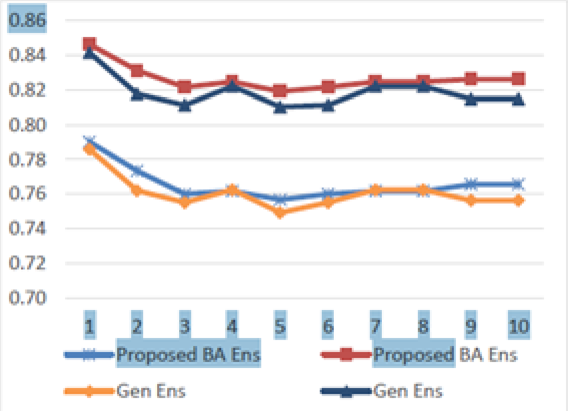


Fig. 9. Comparison of 10 Runs between Proposed Ensemble Pruning using BA and Ensemble Pruning using GA.

Furthermore, to show the effectiveness of using BA ensemble pruning, comparison between the proposed framework and ensemble pruning using genetic algorithm (GA) was performed. Fig. 9 shows that using BA outperformed GA in selection ensemble members. T-Test analysis result showed that the using of BA in ensemble pruning is statically significant than GA with $P<0.00037$.

IV. Conclusions

In this paper, a novel ensemble pruning method based on BA optimization algorithm is presented to select the optimal ensemble base classifiers. This method assists the radiologists for diagnosing the malignity and benignity of the masses. The ROI was manually cropped. Ensemble generation process was done using three different classifiers and twelve group of features extracted directly from the ROI. Thresholding was applied to discretize BA initialization. AUC of the ensemble and diversity between ensemble's members were used to calculate the fitness function of BA. Different aggregation functions were used to test the result of the proposed method. The result showed that the proposed framework is stable. And comparing to single classifier frameworks and GA in ensemble pruning, using BA in ensemble pruning enhanced the quality of the ensemble learning.

Acknowledgment

We would like to thank the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for providing the facilities and financial support under "ETP-2013-053 diagnostic services nexus for breast cancer". We would also like to acknowledge Human Life Advancement Foundation (HLAF) for the financial support to the main researcher. We have obtained Ethics approval with number "UKM 1.5.3.5/244/FTSM-002-2015" entitled "Diagnostic services nexus for breast cancer" from UKM Medical Center, Malaysia, for collecting and conducting experiments on breast cancer patient records. Here, we would like and acknowledge Prof Fuad Ismail from Department of Radiotherapy and Oncology, UKM medical centre and Prof Norlia Abdullah form Department of Surgery, UKM medical centre for their guidance in preparing and applying ethical proposal. Finally, we got an ethical approval from Medical Research and Ethics Committee entitles of "Diagnostic services nexus for breast cancer" with registration No "NMRR-15-21800-27949(IRR)" to allow us collecting mammogram data from HKL.

References

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-tieulent, and A. Jemal, "Global Cancer Statistics, 2012," *CA a cancer J. Clin.*, vol. 65, no. 2, pp. 87–108, 2015.

[2] X. Liu and J. Tang, "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method," IEEE Syst. J., vol. 8, no. 3, pp. 910–920, 2014.

[3] D. L. Monticciolo, M. S. Newell, R. E. Hendrick, M. A. Helvie, L. Moy, B. Monsees, D. B. Kopans, P. R. Eby, and E. A. Sickles, "Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging," J. Am. Coll. Radiol., vol. 14, no. 9, pp. 1137–1143, 2017.

[4] R. Rabidas, A. Midya, J. Chakraborty, and W. Arif, "A Study of Different Texture Features Based on Local Operator for Benign-malignant Mass Classification," Procedia Comput. Sci., vol. 93, no. September, pp. 389–395, 2016.

[5] A. Tahmasbi, F. Saki, and S. B. Shokouhi, “Classification of benign and malignant masses based on Zernike moments,” Comput. Biol. Med., vol. 41, no. 8, pp. 726–735, 2011.

[6] M. Z. Do Nascimento, A. S. Martins, L. A. Neves, R. P. Ramos, E. L. Flores, and G. A. Carrijo, “Classification of masses in mammographic image using wavelet domain features and polynomial classifier,” Expert Syst. Appl., vol. 40, no. 15, pp. 6213–6221, 2013.

[7] M. Abdel-Nasser, H. A. Rashwan, D. Puig, and A. Moreno, “Analysis of tissue abnormality and breast density in mammographic images using a uniform local directional pattern,” Expert Syst. Appl., vol. 42, no. 24, pp. 9499–9511, 2015.

[8] Y. a S. Duarte, M. Z. Nascimento, and D. L. L. Oliveira, “Classification of mammographic lesion based in Completed Local Binary Pattern and using multiresolution representation,” J. Phys. Conf. Ser., vol. 490, p. 012127, 2014.

[9] N. Azizi, Y. Tlili-guiassa, and N. Zemmal, “A Computer-Aided Diagnosis System for Breast Cancer Combining Features Complementarily and New Scheme of SVM Classifiers Fusion,” Int. J. Multimed. Ubiquitous Eng., vol. 8, no. 4, pp. 45–58, 2013.

[10] B. W. Hong and B. S. Sohn, “Segmentation of regions of interest in mammograms in a topographic approach,” IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 1, pp. 129–139, 2010.

[11] A. Rojas Domínguez and A. K. Nandi, “Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection,” Comput. Med. Imaging Graph., vol. 32, no. 4, pp. 304–315, 2008.

[12] W. Xie, Y. Li, and Y. Ma, “Breast mass classification in digital mammography based on extreme learning machine,” Neurocomputing, 2015.

[13] D. Saraswathia and E. Srinivasan, “A high- sensitivity computer-aided system for detecting microcalcifications in digital mammograms using curvelet fractal texture features,” Comput. Methods Biomech. Biomed. Eng. Imaging Vis., no. September, 2016.

[14] A. Qasem, S. Norul, H. Sheikh, and F. I. Shahnorbanun Sahran, Tengku Siti Meriam Tengku Wook, Rizuana Iqbal Hussain, Norlia Abdullah, “Breast Cancer Mass Localization Based on Machine Learning,” in 2014 IEEE 10th International Colloquium on Signal Processing & its Applications (CSPA), 2014, pp. 31–36.

[15] A. Qasem, S. Norul, H. Sheikh, S. Sahran, and F. Ismail, “An Accurate Rejection Model for False Positive Reduction of Mass Localisation in Mammogram,” Pertanika J. Sci. Technol., vol. 25, no. S6, pp. 49–62, 2017.

[16] D. Albashish, S. Sahran, A. Abdullah, N. A. Shukor, and S. Pauzi, “Ensemble Learning of Tissue Components for Prostate Histopathology Image Grading,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1134–1140, 2016.

[17] D. Albashish, S. Sahran, and A. Abdullah, “Lumen-Nuclei Ensemble Machine Learning System for Diagnosing Prostate Cancer in Histopathology Images,” Pertanika J. Sci. Technol., vol. 25, pp. 39–48, 2017.

[18] S. Sahran, D. Albashish, A. Abdullah, N. A. Shukor, and S. Hayati Md Pauzi, “Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading,” Artificial Intelligence in Medicine, vol. 87. pp. 78–90, 2018.

[19] A. A. Aburomman and M. Bin Ibne Reaz, “A novel SVM-kNN-PSO ensemble method for intrusion detection system,” Appl. Soft Comput. J., vol. 38, pp. 360–372, 2016.

[20] Y. Zhang, N. Tomuro, J. Furst, and D. S. Raicu, “Building an ensemble system for diagnosing masses in mammograms,” Int. J. Comput. Assist. Radiol. Surg., vol. 7, no. 2, pp. 323–329, 2012.

[21] J. Y. Choi, D. H. Kim, K. N. Plataniotis, and Y. M. Ro, “Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography,” Expert Syst. Appl., vol. 46, pp. 106–121, 2016.

[22] N. V Chawla and J. Sylvester, “Exploiting Diversity in Ensembles: Improving the Performance on Unbalanced Datasets,” Int. Work. Mult. Classif. Syst. Springer Berlin Heidelb., vol. 4472, pp. 397–406, 2007.

[23] H. Parvin, M. Mirnabibaboli, and H. Alinejad-Rokny, “Proposing a classifier ensemble framework based on classifier selection and decision tree,” Eng. Appl. Artif. Intell., vol. 37, pp. 34–42, 2015.

[24] A. Peimankar, S. J. Weddell, T. Jalal, and A. C. Lapthorn, “Ensemble Classifier Selection Using Multi-Objective PSO for Fault Diagnosis of Power Transformers,” Evol. Comput., vol. IEEE Congr, pp. 3622–3629, 2016.

[25] E. Parhizkar and M. Abadi, “BeeOWA: A novel approach based on ABC algorithm and induced OWA operators for constructing one-class classifier ensembles,” Neurocomputing, vol. 166, pp. 367–381, 2015.

[26] L. Shi, L. Xi, X. Ma, M. Weng, and X. Hu, “A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization,” Appl. Soft Comput. J., vol. 11, no. 8, pp. 5674–5683, 2011.

[27] A. Ekbal and S. Saha, “Classifier Ensemble Selection Using Genetic Algorithm for Named Entity Recognition,” Res. Lang. Comput., vol. 8, no. 1, pp. 73–99, 2010.

[28] D. T. Pham, A. Ghanbarzadeh, E. Koç, S. Otri, S. Rahim, and M. Zaidi, “The Bees Algorithm - A Novel Tool for Complex Optimisation Problems,” Intell. Prod. Mach. Syst. - 2nd I*PROMS Virtual Int. Conf. 3-14 July 2006, no. August 2015, pp. 454–459, 2006.

[29] W. A. Hussein, S. Sahran, and S. N. H. S. Abdullah, “A fast scheme for multilevel thresholding based on a modified Bees Algorithm,” Knowledge-Based Syst., vol. 101, pp. 114–134, 2016.

[30] Y. Freund and R. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” Comput. Syst. Sci., vol. 55, pp. 119–139, 1997.

[31] C.-J. L. Chang, Chih-Chung, “LIBSVM: a library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27, 2011.

CNNSFR: A Convolutional Neural Network System for Face Detection and Recognition

Lionel Landry SOP DEFFO[1], Elie TAGNE FUTE[2]
Department of Computer Engineering, Department of Mathematics and Computer Science,
{University of Dschang, University of Buea} Cameroon

Emmanuel TONYE[3]
National Advanced School of Engineering, Department of Electrical Engineering
{University of Yaounde I} Cameroon

***Abstract*—In recent years, face recognition has become more and more appreciated and considered as one of the most promising applications in the field of image analysis. However, the existing models have a high level of complexity, use a lot of computational resources and need a lot of time to train the model. That is why it has become a promising field of research where new methods are being proposed every day to overcome these difficulties. We propose in this paper a convolutional neural network system for face recognition with some contributions. First we propose a CRelu module, second we use the module to propose a new architecture model based on the VGG deep neural network model. Thirdly we propose a two stage training strategy improved by a large margin inner product and a small dataset and finally we propose a real time face recognition system where face detection is done by a multi-cascade convolution neural network and the recognition is done by the proposed deep convolutional neural network.**

Keywords—Convolutional neural network; face recognition; VGG model; CReLU module; deep learning; architecture

I. Introduction

High-quality cameras in mobile devices have made facial recognition a viable option for authentication as well as identification. However, the used multimedia computational devices cannot act as well human being does. That is why studies have tried to mimic the behavior of human brain to approximate artificially the results obtained by a human being: it is the notion of deep learning. In the mid-1960s, scientists began work on using the computer to recognize human faces. Since then, facial recognition software has come a long way.

In 1966, Bledsoe [1], [2] developed a system that could classify photos of faces by hand using what’s known as a RAND tablet, a device that people could use to input horizontal and vertical coordinates on a grid using a stylus that emitted electromagnetic pulses

In 1987, Sirovich and Kriby [3], were able to show that feature analysis on a collection of facial images could form a set of basic features. They were also able to show that less than one hundred values were required in order to accurately code a normalized face.

In 1991, Turk and Pentland [4] expanded upon the Eigen face approach by discovering how to detect faces within images. This led to the first instances of automatic face recognition

From 1993 to 2000 the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology rolled out the Face Recognition Technology (FERET) program [5] which consists of creating a database of facial images. The database was updated in 2003 to include high-resolution 24-bit color versions of images. Included in the test set were 2,413 still facial images representing 856 people.

From 2005, the Face Recognition Grand Challenge (FRGC) [6] consisted of progressively difficult challenge problems was launched. It includes sufficient elements to overcome the lack of data. The set of defined experiments assists researchers and developers in making progress to meet the new performance goals.

The year 2010 was marked with a great change in the social media platforms all over the world and has leaded researchers to develop photo tagging feature for its user. However the accuracy was not that satisfying that is why technologies using deep learning such as deep face where born [7]. His tools identify human faces in digital images. It employs a nine layer neural network with over 120 million connection weights and was trained on four million images uploaded by Facebook users.

Many other models have been developed over years and two of the most popular are Facenet network [8] and VGG network [9]. They propose a deep architecture that is able to deal with the complexity of classification problem. However, these architectures generally need a very huge date set and a lot of iterations to have good results which if often difficult to have in some cases.

This paper presents a convolutional Neural Network System for Face Recognition based on VGG model and has four proposed contributions. First we propose a CRelu module that has proved to be efficient in enhancing computations; second we use the module to propose a new architecture of VGG network. Thirdly we propose training strategy that needs small dataset and we prove that it leads to good results and finally we propose a real time face recognition system where face detection is done by a multi-cascade convolution neural network and the recognition is done by the proposed deep convolutional neural network.

The rest of the paper is organized as follows: Section 2 presents the details on the proposed approach. In Section 3, the training methodology is presented. Section 4 presents the implementation, analysis and results interpretations included. Finally, Section **5** concludes the work by doing an appraisal and by proposing amelioration perspectives.

II. Method

In this section, we present our proposed model for face recognition based on the VGG [10] deep convolutional neural network. It is a deep architecture that has been developed by the visual geometric group of the University of Oxford in 2015. It has proven to be very efficient in the image recognition task. In addition we have noticed that the deeper the network, the better are the results for more coefficients are used to compute the expected results. Also, we have noticed that the choice activation function is also crucial when designing the network and commonly for convolutional neural networks, the used function is ReLU (Rectified Linear Unit, Rectifier) which is an activation function for Neural Network, known as a ramp function and applied to computer vision and speech recognition. It has been used with some success in restricted Boltzmann machines for computer vision tasks [11].

Several variations have been proposed, like ELU [12] (Exponential linear unit), PReLU [13] (Parametric rectified linear unit), LReLU (Leaky ReLU) [14] and RReLU [15] (randomized ReLU). In contrast to ReLU, in which the negative part is totally dropped, leaky ReLU assigns a noon-zero slope to it .In PReLU, the slopes of negative part are learned from data rather than predefine and has prove to be a key factor of surpassing human-level performance on ImageNet classification task. ELU speeds up learning and alleviates the vanishing gradient problem however; it positive part has a constant gradient of one so it enables learning and does not saturate a neuron on that side of the function. In ReLU, the slopes of negative parts are randomized in a given range in the training, and then fixed in the testing and could reduce over-fitting due to its randomized nature.

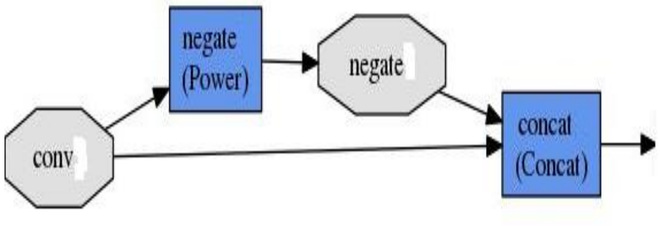


Fig 1. Proposed CRelu Module.

In this view, we propose a simple CReLU, where the idea in general is to concatenate a ReLU which selects only the positive part of the activation with a ReLU which selects only the *negative* part of the activation. Note that as a result this non-linearity doubles the depth of the activations [16]. This is with the knowledge that CReLU increases the quality of the result as proven in [17]. We therefore propose a simple CReLU shown in Figure 1.

We can see how we have connected the output of the convolution to the negation of the same output. The next step is to replace every activation functions, here is ReLU and PreLU essentially and it gives rise to a new architecture of the VGG model.

A. The Presentation of the VGG Model

It is a deep convolutional network for object recognition developed and trained by Oxford's renowned visual geometric group (VGG), which achieved very good performance on the ImageNet dataset. It is quite famous because not only it works well, but the Oxford team has made the structure and the weights of the trained network freely available on-line.

The idea of the VGG group members was to give an answer to “how to design the network structure”. Among many choices, they has adopted the simplest. Only 3x3 convolutions and 2x2 pooling are used throughout the whole network. They have also used the fact that the depth of the network plays an important role. Deeper networks give better results.

Figure 2 gives the structure of the model, which takes input image of size 224 * 224 * 3 (RGB image), built using Convolutions layers (used only 3*3 size), max pooling layers (used only 2*2 size), a fully connected layers at end and has a of total 16 layers. Below is the description of each layer.

1) Convolution using 64 filters
2) Convolution using 64 filters + Max pooling
3) Convolution using 128 filters
4) Convolution using 128 filters + Max pooling
5) Convolution using 256 filters
6) Convolution using 256 filters
7) Convolution using 256 filters + Max pooling
8) Convolution using 512 filters
9) Convolution using 512 filters
10) Convolution using 512 filters + Max pooling
11) Convolution using 512 filters
12) Convolution using 512 filters
13) Convolution using 512 filters + Max pooling
14) Fully connected with 4096 nodes
15) Fully connected with 4096 nodes
16) Output layer with Softmax activation with 1000 nodes

B. The Proposed Model

We have already explained in details the proposed CRelu module and presented in details the architecture of the VGG chosen model. It is therefore important to mention that the model uses the ReLu activation function and is used 15 times in the network. Our proposed model will therefore replace these ReLu function by the the proposed module. Also in the last layer (layer 16) the softmax inner product is replaced by the combination of Large Margin Inner Product and Softmax with Loss. It is usually called L-Softmax loss [18] built for convolutional neural networks and this loss can greatly improve the generalization ability of CNNs, so it is very suitable for general classification, feature embedding and biometrics. This gives rise to the architecture presented on Figure 3.

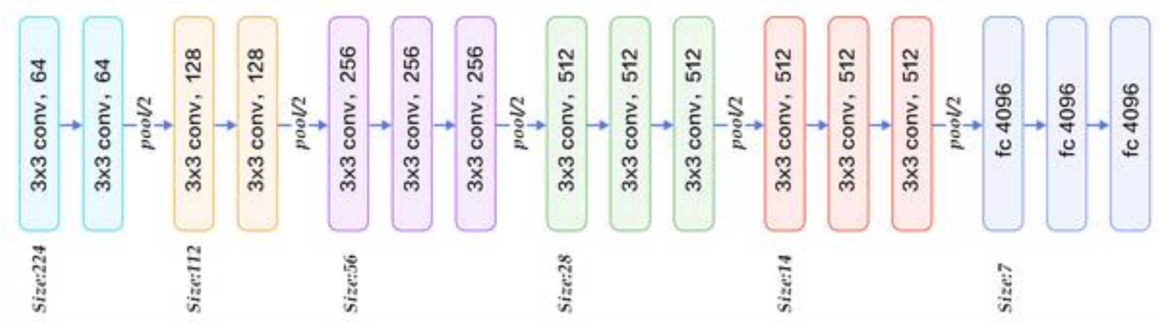


Fig 2. VGG Model.

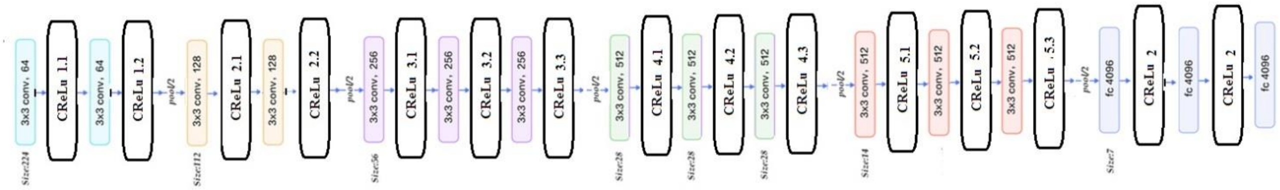


Fig 3. Proposed VGG Model.

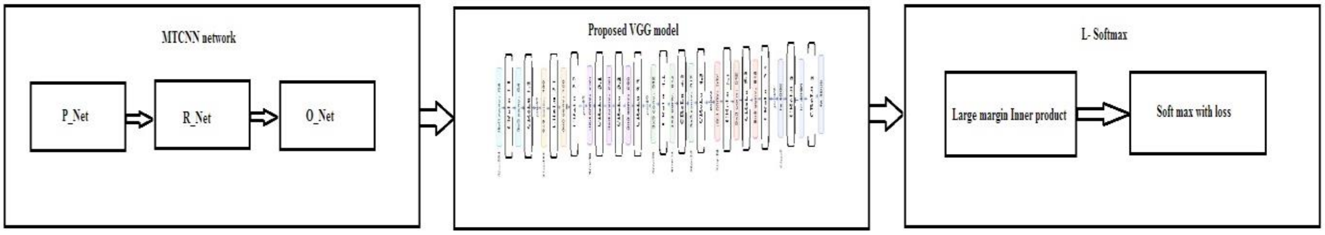


Fig 4. Final Architecture of the System.

C. The Real Time System

Now that we have proposed the recognition model we combine it to a detection model to produce our final framework. It is well known that a face recognition system passes through a detection phase before recognition. However the proposed approaches in the literature usually use face cascade detection which is relatively old. We have decided to use the MTCNN (multi cascade convolutional neural network). It is based on:

- A Proposal Network (P-Net) used to obtain the candidate facial windows and their bounding box regression vectors. Then candidates are calibrated based on the estimated bounding box regression vectors. Finally a non-maximum suppression (NMS) is employed to merge highly overlapped candidates.
- A Refine Network (R-Net), which further rejects a large number of false candidates, performs calibration with bounding box regression, and conducts NMS.
- An O-Net network which aims to identify face regions with more supervision. In particular, the network will output five facial landmarks positions.

We combine this multi cascade neural network with the proposed VGG model and finally we add an L-softmax module which is as stated earlier composed of a large margin inner product and a softmax with loss. Not that this module replaces the last inner product layer of the proposed VGG model. This give rise to the architecture presented on Figure 4.

III. Training Methodology

To train our model, we perform the following steps:

- We gather images that will be used for training and divide them into train set and test set with the ratio 2/3 and 1/3. These images are usually taken from public datasets where each identity has at least 80 images.
- Each identity is assigned a label, consequently all the images of one identity is assign a unique number. This leads to the creation for each identity a file containing the names of all it images zit hot corresponding label
- We gather into one file all the names and label and shuffle the obtained results such a way that all images names of one identity should not be adjacent. Note that it is better for each name of image to be written with it absolute path

- Divide the images into train set and test set with ratio 2/3 and 1/3 indicating that the number of images for one identity in the training set should be the double of the one present in the test set.
- When all this are done you can use that information to train your network. But first of all a training is done using the model and no initial weigh value then the, obtained weigh values are used to fine-tuned the same work. In our case we have decided to take 1000 iterations for the first training and 9000 iterations to fine-tune. This has proven that it is more efficient than the one using the previous approach.

IV. Results

A. *Training Results*

We choose Caffe [19] to implement our solution. It is Caffe a deep learning framework made with expression, speed, and modularity in mind. It is developed by Berkley AI Research (BAIR) and by community contributors. The choice is motivated by: Expressive architecture, Extensible code, Speed and Community.

It is written in C/C++ and has a python interface. The parameters used to train our model are listed in table 1.

Since we are working in CPU mode, it was almost impossible to work on large dataset for the resources are limited in that mode. The GPU memory of the machine used is only of 2GB so was unable to do more than 10 steps with a memory dump. For that reason we have chosen 7 identities from the pub83 [20] dataset and the last one is that of the second author. Each of these identities has at least 80 pictures for the training set and at least 20 pictures for the testing set which gives a total of about 100 images per persons.

We had the following results during training.

Figure 5 shows the variation of loss during training as well of that of the accuracy. We can notice that the accuracy tend to increase and decrease later on. For the lost it seems to increase only.

TABLE I. PARAMETERS USED IN THE TRAINING PROCESS

Parameters	Values
Number of iterations	10 000
Initialization method	Xavier method
Propagation algorithm	Stochastic gradient descent (SGD)
momentum	0.9
Weigh decay	0.0005
Batch size	8
Learning rate	0.001
Test interval	20
Step size	2000

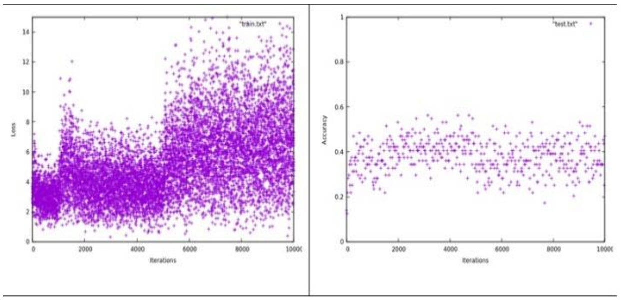


Fig 5. Lost and Accuracy Evolution during Training with the Original Model.

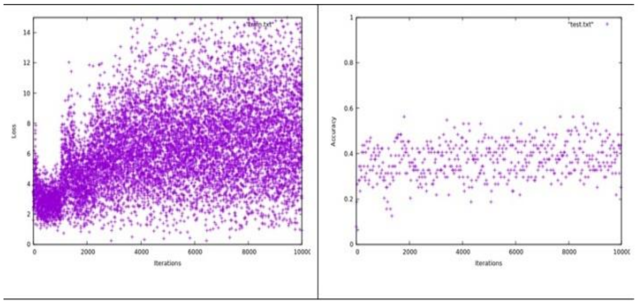


Fig 6. Lost and Accuracy Evolution during Training with the Proposed Model.

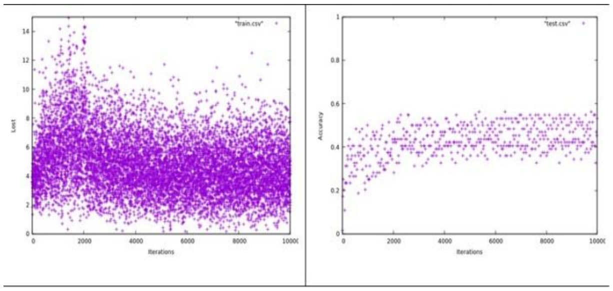


Fig 7. Effect of Large Margin on Training.

TABLE II. RESULTS OBTAINED DURING THE TEST PROCESS

Identity	Out put probability of the original model	Out put probability of the proposed model
SOP DEFFO	0.99	0.99
Angeline Jolie	0.958	0.959
Barack Obama	0.013	0.997
Beyonce Knowles	0.0003	0.2
Brad Pitt	0.044	0.05
Christina Ricci	0.0001	0.03
Georges Clooney	0.0003	0.001
Halle Berry	0.0019	0.17

A different observation can be made on figure 6. The loss in increasing and when closer to the end of the training it starts decreasing. For the accuracy, it increases gradually which is a good result.

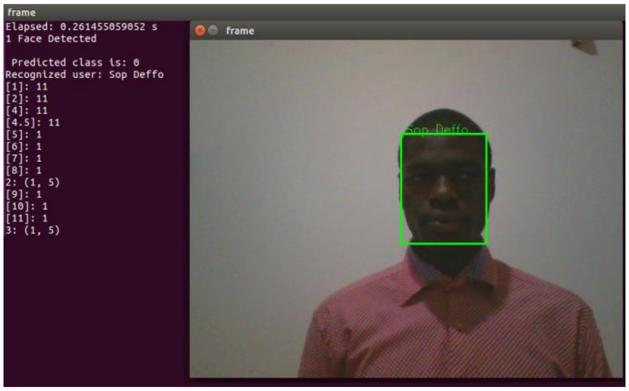
Finally on Figure 7, the convergence of loss is more visible thanks to the large margin module. In addition, the evolution of accuracy is more perceptible which means the result is getting better.

After these observations during training let us see the effect on the output probabilities for each identity presented in table II

We see that our results are better than the one obtained using the original VGG model. With the original model, we observe a high output probability for only 2 identities and a very low one for the other but with our model we have high values for 3 identities and acceptable one for the rest.

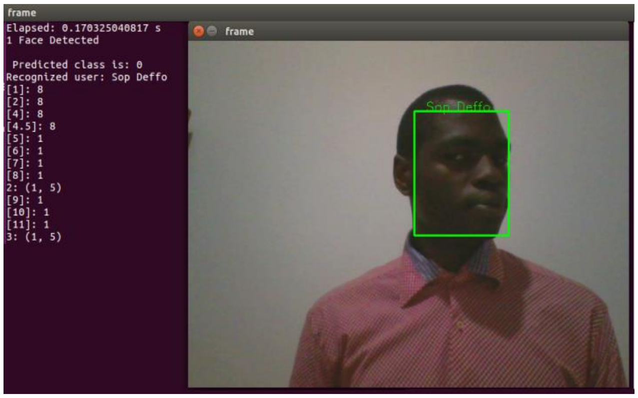
B. Real Time Detection and Recognition

Figure 8 and Figure 9 present a real time detection and recognition by the proposed system. It can be seen how the face is first of all detected by the bounding box then recognized later on. The label of the detected person is written on top of the image. This shows that the system is really working.



```
Elapsed: 0.261455059052 s
1 Face Detected

 Predicted class is: 0
Recognized user: Sop Deffo
[1]: 11
[2]: 11
[4]: 11
[4.5]: 11
[5]: 1
[6]: 1
[7]: 1
[8]: 1
2: (1, 5)
[9]: 1
[10]: 1
[11]: 1
3: (1, 5)
```

Fig 8. Detection and Recognition of Identity Sop Deffo.



```
Elapsed: 0.170325040817 s
1 Face Detected

 Predicted class is: 0
Recognized user: Sop Deffo
[1]: 8
[2]: 8
[4]: 8
[4.5]: 8
[5]: 1
[6]: 1
[7]: 1
[8]: 1
2: (1, 5)
[9]: 1
[10]: 1
[11]: 1
3: (1, 5)
```

Fig 9. Detection and Recognition of identity Sop Deffo.

V. Conclusion

We presented in this paper a convolutional neural network system for face detection and recognition. In this system the detection is done by a multi cascade convolutional neural network system and recognition by deep proposed neural network architecture. The proposed model is based on the deep VGG neural network, a large margin inner product and a proposed CRelu function. The results obtained have proven to be better than the one obtained using the original model. For future work we intend to find mechanism to increase the size of the dataset in order to be able to recognize many persons.

References

[1] Bledsoe, W.W,"The model method in facial recognition," Panoramic Research Inc., Palo Alto, CA, Rep. PRI:15, August 1966

[2] Bledsoe, W.W, "Man machine facial recognition, Panoramic Research Inc., Palo Alto, CA, Rep. PRI:22, August 1996

[3] L. Sirovich and M. Kirby Low-Dimensional procedure for the characterization of human faces. Journal of optical society of America Vol 4 page 519 March 1987

[4] Mattew A Turk and Alex P. Pentland, Face recognition using Eigen faces, vision and modeling group, The media laboratory , Massachusetts Institute of Technology, 1991

[5] Jonathon Phillips, Patrick J. Rauss, and Sandor Z. De, FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results, Army Research Laboratory (ARL), October 1996

[6] P. Jonathon Phillips, Patrick J. Flynn Todd Scruggs Kevin W. Bowyer, William Worek, Preliminary Overview of the Face Recognition Grand Challenge, IEEE Conference on Computer Vision and Pattern Recognition 2005

[7] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, Deep Face Recognition, Visual Geometry Group, Department of Engineering Science, University of Oxford

[8] West, J (2017) History of Face Recognition – Facial recognition software [online] FaceFirst Face Reconition facial recognition software available on https://www.facefirst.com/blog/brief-of-face recognition-software/ [Accessed 15 Oct. 2018

[9] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge CoRR, abs/1409.0575, 2014

[10] Karen Simonyan and Andrew Zisserman, very deep convolutional network for large-scale image recognition, ICLR conference 2015

[11] Vinod Nair and Geoffrey Hinton, Rectified linear Units improve Reststricted Boltzmann Machines. ICML 2010

[12] Clevert D.A., Unterthiner T. &Hochreiter S. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)

[13] Maas, Andrew L, Hannun, Awni Y, and Ng, Andrew Y. Rectifier nonlinearities improve neural network acoustic models. In ICML, volume 30, 2013

[14] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. arXiv preprint arXiv:1502.01852, 2015.

[15] Wang, Naiyan, Li, Siyi, Gupta, Abhinav, and Yeung, Dit-Yan. Transferring rich feature hierarchies for robust visual tracking. arXiv preprint arXiv:1501.04587, 2015

[16] Wenling Shang, KihyukSohn, Diogo Almeida, Honglak Lee, Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units, arXiv:1603.05201

[17] Shifeng Zhang Xiangyu Zhu Zhen, Lei Hailin, Shi Xiaobo and Wang Stan Z. Li FaceBoxes: A CPU Real-time Face] Detector with High Accuracy, arXiv: 1708.05234v2 [cs.CV] 19 Aug 2017.

[18] Weiyang Liu, Yandong Wen, Zhiding Yu and Meng Yang Large-Margin Softmax Loss for Convolutional Neural Networks Proceedings of The 33rd International Conference on Machine Learning. 2016: 507-516.

[19] Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor, Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv preprint arXiv:1408.5093, 2014

[20] Becker, B. C. and Ortiz, E.G. "**Evaluating Open-Universe Face Identification on the Web**," In CVPR 2013, Analysis and Modeling of Faces and Gestures Workshop.

A Proposed Methodology on Predicting Visitor’s Behavior based on Web Mining Technique

Abdel Karim Kassem[1]
University of Angers
Angers, France

Bassam Daya[2]
Lebanese University
Saida, Lebanon

Pierre Chauvet[3]
Université Catholique de l’Ouest
Angers, France

***Abstract*—The evolution of the internet in recent decades enlarge the website's reports with the records of user’s activities and behaviors that registered in the web server which can be created automatically in the web access log file. The feedback concerning the user’s activities, performance and any problem that may be occur including the cyber security approaches of the web server represents the principal raison of applying the web mining technique. In this paper, we proposed a methodology on predicting users behavior based on the web mining technique by creating and executing analysis applications using a Deep Log Analyzer tool that applied on the web server access log of our faculty website. Furthermore, an associated programmed application has been developed which employs the extracted data into dynamic visualizations reports(tables, graphs, charts) in order to help the web system administrator to increase the web site effectiveness, we had creating a suitable access patterns that permits to identify the interacting users behaviors and the interesting usage patterns such as the occurred errors, potential visitors, navigation activities, behavioral analysis, diagnostic study, and security alerts for intrusion prevention. Moreover, the obtained results achieved the aim of producing a dynamic monitoring by extracting investigation summaries which analyses the discovered access patterns that registered in the faculty web server in order to improve the web site usability by tracking the user’s behaviors and the browsing activities. Our proposed tool will highlight providing a security alerts against the malicious users by predicting the malicious behaviors taking into consideration all the discovered vulnerabilities by detecting the corrupted links used by the abnormal visitors.**

Keywords*—*Web server; log file; web mining; behavior; pattern; web usage mining; visualizations; vulnerabilities; security

I. Introduction

Web Usage mining is the strategy of applying web mining techniques to discover and analyze in real time clickstreams usage patterns and related data generated as a result of user interactions with one or multiple web sites. Specifically, web usage mining is the process of grabbing and extracting valuable information in order to find patterns relating to user’s behavior of a specific web based system that can determine: who they are, and what they tend to do. Web usage mining techniques consists of the following sections: pre-processing, pattern discovery, pattern analysis.

When a user requested specific and particular resources of web server, each request will be recorded and stored in the web log. This record is referring to the browsing behavior of the user. In Web Usage Mining, data can be collected from multiple resources such as: files (image, sound, video and web files), operational databases and server log files that can include web server access logs and application server logs. Otherwise the collected data in the web log file will be an unstructured format and it can’t be used directly for mining purposes, many techniques should be applied on it, the Pre-processing technique play the role of converting the data into suitable and organized form that can helps to precise the pattern discovery and to provide accurate, appropriate and summarized information for data mining intent. Data pre-processing, includes data cleaning, user identification, user sessions identification, path completion and data integration. Pattern discovery benefit from the preprocessing results in order to offer some techniques such as statistical analysis, sequential pattern analysis, association rules, clustering and other techniques. The pattern analysis should be executed and performed by the following techniques: visualization techniques, OLAP techniques and usability analysis.

Aside from detecting the visitors’ activities and their behavior, web usage mining can be effectively used to detect existing weaknesses on the web server components and analyzing audit results for anomalous patterns detection.

This research is divided into two parts, the first one by proposing a methodology based on the web usage mining technique that can easily detect the visitor behavior by analyzing the registered visitors’ activities on the log file and exporting analysis results to describe the usability of the faculty website, the second one is to discover the cyber-attacks by monitoring the visitors through the links sent to our web server.

To achieve our target, we apply the web usage mining by selecting the data type from our university apache web server which generates the web log file that used for mining purposes, these techniques are used to facilitate the determination of the user behavior and their activities on the web server by creating the rule of the access patterns selection, Furthermore, we will focus on generating some summaries in order to highlight the occurred errors that can be happened on our faculty website, analyzing the traffics, controlling the accessed web server resources, and detecting the illegal activities for expected visitor by controlling the accessed links to discover the web page vulnerabilities which it is a weakness that can be exploited by a threat attacker in order to perform cybercrime actions on the web server.

This paper is organized as follows. Related work in Section 2. Section 3 define the web usage mining methodology and providing an overviews about its types. Section 4 presents the web usage mining techniques according to the behavioral detection approaches. Section 5 describes our proposed methodology of the detection, followed by Section 6 where we state our experimental results and the extracted analyses summaries. Section 7 conclude the research work which its supported by a proposed perspective that can involve this research topic.

II. Related Works

In this section, we reveal the related work concerning our research study area. The daily web usage of websites with the big amounts of data resulted every second derive us to conclude that much attention has been drawn to the web usage mining that represents one of the popular research areas.

In web usage mining, data analyzation is essential for tracking the user behavior in order to serve the users in efficient way.

Several researchers that are shown in [1][2] developed preprocessing data model; they collected the data related to user ID, path completion, session ID, transaction ID etc. In this way they improve the organization by facilitating the determination of particular clients, products marketing plans and other promotional goals, etc.

According to [3], the authors presents the web log data files and their data difficulties. In addition, the author highlighted about the lessons and metrics based on e-commerce and about the web server's insufficiencies then he introduces some statistical graphs to find the fitting solutions and cover the resulted issues.

The authors in paper [4] presented a technique for detecting the interests of the visitors according to a study of the site-keyword graph. This technique can extract sub-graphs to reveal the major interests of the users taken from the site-keyword-graph were the data is collected from the log data of the website. According to [5] the authors described a mining algorithm for incremental web traversal pattern, this algorithm employs the mining results and predicts another patterns using the deleted or inserted data parts of the logs in the websites like the mining duration that may be reduced. The authors present in [6] an analysis on the web log data via a method for statistical analyzation. Moreover, this author clarifies a recommended tool for efficient realization and interpretation of the preprocessed statistical results taken from log file.

According to [7], the authors worked on this research topic by abstracting the log lines to log event types in order to mine the system logs, this work has been accomplished by presenting a technique based on clustering using the simple log file clustering tool to abstract the logs; moreover, this technique is useful when we cannot access the source code of the application. This research was done by the virtual computing lab at the university of North Carolina state.

These papers [8][9] explore the user session by applying detailed characterization study, after that the authors preview the results for several views such as each user requests per session, page number requested per each session, the session length.

III. Web Usage Mining

Web mining consist of three categories: Web content mining, web structure mining and web usage mining. The concept of Web usage mining is to gather data and information generated by the web. While the concept of the web content and structure mining is to apply the primary data on the web, moreover web usage mining will mine the secondary data obtained by the interactions of the multiple users in the web[10]. One of the functions of the web usage mining is to include the data from the web server access logs, browser logs, proxy server logs, registration data, user profiles and sessions, user queries, cookies, mouse clicks and scrolls, bookmark data and other detailed data as interaction results.

The web usage mining technique can be declared by three steps process: data pre-processing, pattern discovery and pattern analysis as we shown in the Fig. 1.

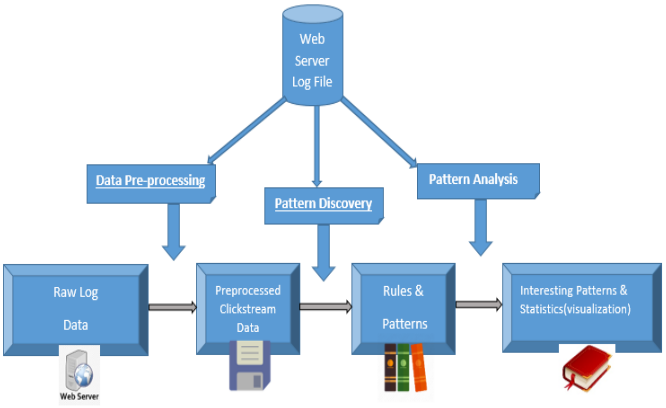


Fig. 1. Web usage Mining Process.

A. Data Preprocessing

By accessing any website, actually the user's behaviors[11] will be stored in the web server log file in unclear and unorganized form. As a definition, data preprocessing is the process for converting the raw data presented in log files into suitable form such as data base or different data store type which contribute effectively when applying the data mining algorithm. Since the main log file cannot be directly used in the web usage mining process, due to the large amount of irrelevant entries in the log file and difficulties and many reasons. Hence, web log file's preprocessing becomes essential and significant. Nowadays, many researches centers are interested in data preprocessing of Web Usage Mining methodology.

Thus, data preprocessing plays an essential role in increasing the mining accuracy in order to improve the data quality for further usage.

B. Pattern Discovery

Pattern discovery employs the preprocessing results to offer some techniques such as statistical analysis, association rules, sequential pattern analysis, dependency modeling, classification and clustering to capture beneficial useful information. The results that has been grabbed can be represented and employed in several ways such as graphs, charts and tables, etc. for example the visitor's location can be specified using his own IP address. Therefore, by discovering the web visitors[12], the web server administrator can detect the most active countries who's visiting a certain website or any web page that can provide the useful information relevant to the specific country.

C. *Pattern Analysis*

Pattern analysis can be classified as the final step in the Web Usage Mining process. The main purpose of applying the pattern analysis[13] is to filter out the unusable and the non-beneficial rules and patterns from the set that has been found in the pattern discovery phase. Most Pattern analysis techniques are used to attain the above mentioned purpose.

One of the above techniques is the knowledge query mechanism like SQL which is a standard language for storing, retrieving and manipulating data in databases[14]. Another method is called (OLAP) which is an operation to load usage data into a data cube in order to perform Online Analytical Processing. Visualization techniques is the process of conveying information in a way that the information can be quickly and easily digested by the viewer or the analyzer such as graphing patterns by assigning colors to a specific value in order to highlight overall patterns in the data. Content and structure information are used to extract patterns that contain several pages of a certain usage type that can match with a certain hyperlink structure.

IV. Web usage Mining and Behavioral Detection Approaches

Web mining is an application of data mining methodology that discovers the usable patterns from the internet according to the World Wide Web protocol. As the name inspires, by using the web mining techniques, this information will be gathered from the internet. This technique uses automated devices that reveal and extract data from the web servers and much reports on the internet that permits the companies and educational organizations to extract structured, semi structured and unstructured data from browser actions, server logs, website, web page's contents, page Links and another sources [15]. Web mining techniques[16] can be applied also to detect the user activities as shown in the Fig. 2; this can be reached when we employ their techniques to discover the user behavior as well as it is used to handle the problems presented in the databases and the cyber security troubles through analyzing the illegal and the irregular user activities.

Web usage mining is the practice of extracting valuable information from the server logs in order to find and conclude what visitors are looking for in the interconnected networks(internet), after that the discovered knowledge by the visitors are taken to roam and navigate via the websites [17]. In this paper, we proposed a "mixture approaches" the concept of web usage mining is used intended for the visitor's behavior detection. We can discover the web visitors' information that derive us to identify the user's movements and activities in order to detect and analyze the web traffics, the occurred errors, the users' activities, the abnormal and illegal actions and the security approaches.

The main advantage of the web usage mining technique is to propose a series of those combined approaches that exclusively save the time as well as decrease the estimated cost. Using this kind of techniques, the web administrator will dynamically analyze the user activities and the human efforts to extract the desired reports will be reduced and there will be no need to hard physical potential during the detection.

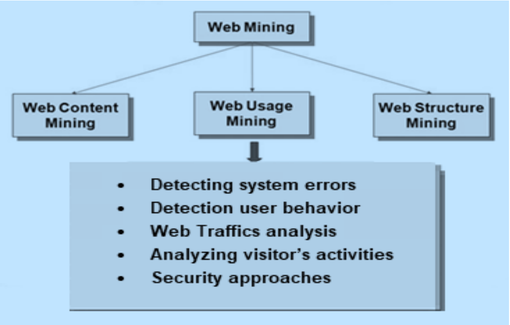


Fig. 2. The Behaviral Detection Approches based on the Web usage Mining Technique.

V. Proposed Methodology

Web Usage mining is the process of applying web mining techniques to discover the approaches of usage patterns from the extracted Web data. The web usage mining is one of the significant and fast developing zone of web mining that it is considered as an important part of the advanced technology (web mining) to discover the user's behaviors events. In this research paper, we developed and applied the Deep Log Analyzer tool associated with a programmed application that requires the web server log file to create a suitable pattern according to the visitor's behaviors by generating statistical and web usage mining reports which can analyze all the detected behaviors approaches.

In this section, we propose the used methodology that assists the web administrator to analyze the occurred system errors, security alerts and user's activities by detecting their behaviors on the web server logs. The steps bellow is included in the proposed methodology.

A. *Data Collection*

In this section, we present the data collection that applied in our research that has been extracted from our faculty web server access log. The web log stores the visitor's activities per each user visit and hit. The collected data was extracted from log file during a period of four days on February 2018 as shown in the TABLE Ibelow.

TABLE I. Details of the Input Data (Access Log File)

Access Log File Details	
File Name	iut.ul-iut.net-Feb-2018
Period	23 Feb 2018 – 26 Feb 2018
Size(KB) after preprocessing	1852.8
Number of entries	6742

B. Data Selection

In this section, we present the data selection concept that we used. Absolutely the web mining methodology has three kinds of data: the server side data, the middle data (proxy side) and the client side data. In our work, we employed the case of web server use.

C. Web Server Log

A web server refers to computer or to server software or both of them working together to transfer web pages. The web server uses HTTP (Hypertext Transfer Protocol) in order to serve the web server files that form Web pages to web users directly in response to achieve their requests, which are forwarded by their HTTP clients the main log file cannot be directly used in the web usage mining process. Log files[18] are files which are composed, established and maintained in a web server. Every hit to the Web site by the users, including each view of HTML documents, images or any other object will be logged. The raw web log file format is ultimately formed of single line text for each interaction, mainly it is a hit related to the web page interactions. The log files have the capability to maintain different types of information [19] and it will be presented in the log file and should summarized who, where and when the users visited the website [19], and it will serve to discover their behaviors and movements. Moreover, when the users communicate and interact with any website, the interaction’s details and the request activity resulted by the web visitor events will be automatically recorded and stored in the web server log file [20][21].

The basic information recorded and discovered in the log file can be shown as

- Username: This identifier will discover who visits the website. The identification of the user principally would be the IP address.
- Visiting Path: The path that the user typed while visiting the website.
- Path Traversed: it will distinguish the path taken by the user via different links.
- Time stamp: The time duration when the user spends on each web page while surfing through the website, this record recognized as a session.
- Last visited Page: The visited web page by the users before the leaving.
- Success rate: The number of downloads made and the number of replicating activities experienced by the user that can specifies the success rate of the website.
- User Agent: This is the browser that can indicate from where the user sends the request to the web server. It will be formed as a string that characterizes the type and the version of browser software being used.
- URL: It will be the resource of the user access. It may be an HTML page, a CGI program, or a script.
- Request Type: The method chosen for transferring data such as GET, POST

D. Tool Selection

Most of the valuable information about any website visitor stored in the log file on the web server, after analyzing these data we can generate beneficial reports as summaries, graphs and analytical figures by using the web usage mining technique which it can be done using various tools. A variety of tools are available in the internet assists the web administrator to apply analysis tasks by accessing the web server log files which produces effective web usage mining reports as output. Some of the most widely used tools are: Google analytics, webalizer, W3Perl, and AWStats. In this paper, we select our deep log analyzer with an analytical application to analyze the desired goal by examining the log file in order to achieve the target, this can facilitate of obtaining an output as reports about the accessed information, user behavior analysis, system errors, threatened links, security approaches, user identity, time, zone, URL, browser and OS of the users. Unlike other tools, our tool has the ability to analyze different types of logs including FTP logs. It can analyze the web site visitors' behavior to get the complete usage statistics that improve the usability and stability of our web site and provide an analytical protective studies in order to avoid the web vulnerabilities that actually occur on the web server.

We can study the extracted results and generate the following reports according to its own features

- View reports about accessed site's resources
- View reports about the most visited links
- View investigation report of the user behavioral activity
- Monitor and control the illegal visitors'
- View the abnormal sites that refers to the web traffic
- Reveal the search queries and search engine spiders
- Reveal the user browsers and operating systems
- View the web server errors
- Apply comparative reports in different time periods
- Analyze the log file with respect to all popular web servers such as IIS on Windows or Apache on Unix/Linux

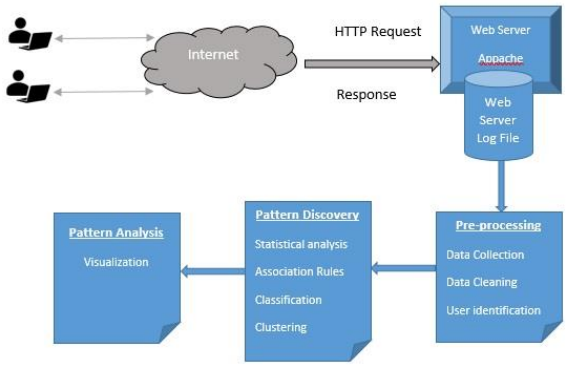


Fig. 3. The Proposed Methodology.

E. Methodology Implementation

Usually, by clicking on a web link or any click stream by the visitors, the web server stores and generates these actions in the log file. Log file consists of multiples raw records about all web pages that provide the discovery detection[22] of the user's behaviors. This paper sheds the light on pattern analysis of the visitor taken from the log data of our university web server.

Throughout this research paper, we can illustrate our framework in the Fig. 3 by analyzing the proposed methodology that permits to understands and evaluate the web visitor's behavior. Hence the user uses the internet service to serve web pages until he/she reach our faculty's website whether directly, or by using the search engines or through referrals resources. The user's actions on the website will be stored on apache server log file.

By applying the Web usage mining, we can collect and investigate the recovered data from it. Furthermore, the next step is to deal with user's interaction through the website in order to infer their behavioral patterns and profiles.

The main purpose of our research is to detect the information with respect to the visitor's behaviors. The extracted Information from the log file will be employed in our tool in order to extract web usage mining statistical results

The most important results will be displayed as listed below:

- Links and Resources Analysis
- Server Content Analysis
- Brower Analysis
- Web Page Analysis
- Security Approaches
- Operating System Analysis
- Time and Place Analysis

VI. Experimental Results and Analysis

The main objective of the web usage mining technique is to generate statistical reports as output results that can be used to detect some valuable information after analyzing them, in this paper we focused on the data extraction from our faculty's web server log file as an input concerning the visitors and the user's behaviors in order to generate an investigation reports with respect to the web server status.

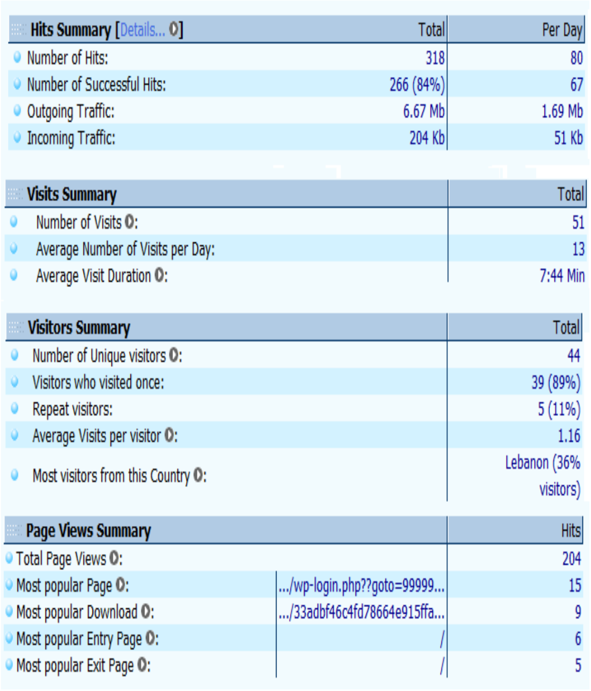
Our research will display and discuss several experimental results as:

1) General activity: the main general activities of our faculty web sites are shown below in the fig.4. which clarifies a brief summary about visitor's information during selected dates.

a) Selection information summary during selected dates

The Fig. 4 illustrates a summary report that will be explained below concerning the statistical results with respect to the number of hits, visits, visitors and page views of the faculty's web site.

- Hits summary that includes the number of hits, the number of successful hits and the outgoing and incoming traffic (as total or per day).
- Visits summary that includes the total number of user visits, the average number of visits per day and the average visit duration.
- Visitors' summary that includes the number of unique visitors, the visitors who visited once, the repeated visitors, the average visits per visitor and the most visitors from this country.
- Page views summary that includes the total page views, the most popular page, the most popular downloaded file, the most popular entry page and the most popular exit page.



Hits Summary [Details... ⓘ]	Total	Per Day
Number of Hits:	318	80
Number of Successful Hits:	266 (84%)	67
Outgoing Traffic:	6.67 Mb	1.69 Mb
Incoming Traffic:	204 Kb	51 Kb

Visits Summary	Total
Number of Visits ⓘ:	51
Average Number of Visits per Day:	13
Average Visit Duration ⓘ:	7:44 Min

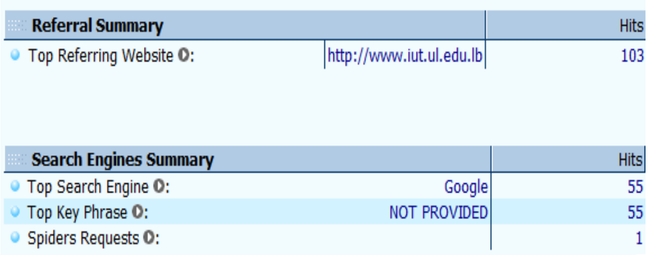
Visitors Summary	Total
Number of Unique visitors ⓘ:	44
Visitors who visited once:	39 (89%)
Repeat visitors:	5 (11%)
Average Visits per visitor ⓘ:	1.16
Most visitors from this Country ⓘ:	Lebanon (36% visitors)

Page Views Summary		Hits
Total Page Views ⓘ:		204
Most popular Page ⓘ:	.../wp-login.php??goto=99999...	15
Most popular Download ⓘ:	.../33adbf46c4fd78664e915ffa...	9
Most popular Entry Page ⓘ:	/	6
Most popular Exit Page ⓘ:	/	5

Fig. 4. The Proposed Methodology:A Brief Summary about the Visitor's Information.

b) Referral summary information

The Fig. 5 as we shown below represents and concludes the referral and search engine summaries.



Referral Summary		Hits
Top Referring Website ⓘ:	http://www.iut.ul.edu.lb	103

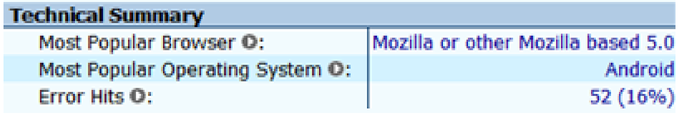
Search Engines Summary		Hits
Top Search Engine ⓘ:	Google	55
Top Key Phrase ⓘ:	NOT PROVIDED	55
Spiders Requests ⓘ:		1

Fig. 5. The Referral Information about the Website.

- Referred summary includes top referring websites on the web server.
- Search engine summary includes top search engine that provide the users to access the university website, top key phrase and spider requests on the search engine provider.

c) Technical information

The Fig. 6 reveals a technical summary that contains the most popular browser, the most popular operating system and the error hits that happened on the web server.

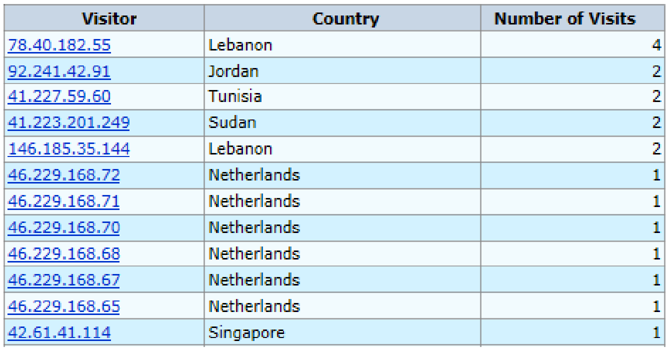


Technical Summary	
Most Popular Browser:	Mozilla or other Mozilla based 5.0
Most Popular Operating System:	Android
Error Hits:	52 (16%)

Fig. 6. Technical Summary for the Website.

2) Visitors activities: by controlling the visitor's activities on the web server, many difficulties can be encountered to detect the visitor's behavior and their purposes. After employing the web usage mining techniques. We achieve the target and we will be able to detect valuable information about the top visitors with their countries and the number of visits that contacts the concerned website as well as the daily and hourly user activities facts that occurred on the web server log file.

a) Selection information summary during selected dates:



Visitor	Country	Number of Visits
78.40.182.55	Lebanon	4
92.241.42.91	Jordan	2
41.227.59.60	Tunisia	2
41.223.201.249	Sudan	2
146.185.35.144	Lebanon	2
46.229.168.72	Netherlands	1
46.229.168.71	Netherlands	1
46.229.168.70	Netherlands	1
46.229.168.68	Netherlands	1
46.229.168.67	Netherlands	1
46.229.168.65	Netherlands	1
42.61.41.114	Singapore	1

Fig. 7. Top Detected Visitors Of The Faculty Website.

The Fig. 7 represents the most active visitors identified by their IP addresses, the countries and visit's numbers of the website.

b) Visitors spending Time

The graph bellow represented by the Fig. 8 determines the spending period time of the visitors in our faculty website. The x-axis represents the spending average time of the visits. however, the y-axis indicates the total number visits of each visitor. We can conclude from this statistical graph that the spending time is continued as long as the web server receives hits from that visitor.

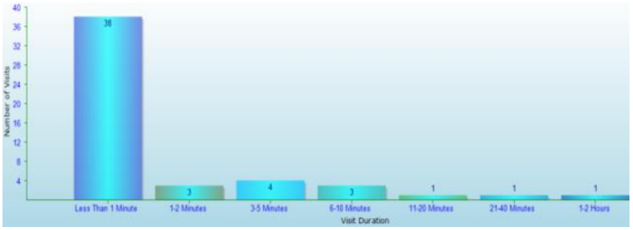
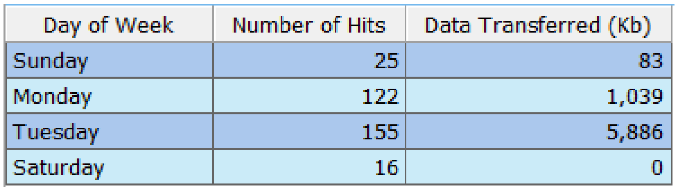


Fig. 8. Visitor's Spending Time on the Faculty Website.



Day of Week	Number of Hits	Data Transferred (Kb)
Sunday	25	83
Monday	122	1,039
Tuesday	155	5,886
Saturday	16	0

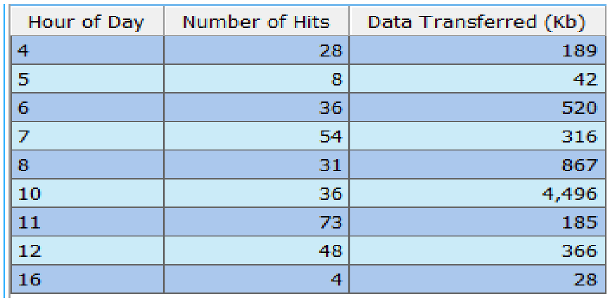
Fig. 9. Popular Days of the Week with the Number of Hits and Data Traffic.

c) Visitors daily activity

The Fig. 9 gives a clear image how the traffic may vary from a day to another in the same week. The traffic is presented by the hits number of each visitor measured in Kb as the transferred data related to the users, this figure will reveal about the days that the website achieves the traffic as a quantitative indicator about the exchanging data in the web server.

d) Hourly rate activity

The Fig. 10 below displays the traffic on the website that can be changed depending on the daily traffic time, we can find out the hourly time of a day of each hit on the website measured in Kb to display the transferred data related to the users.

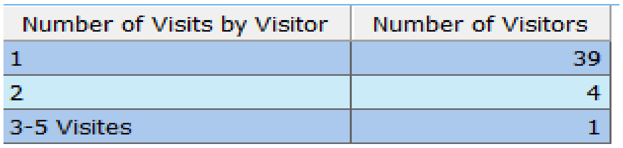


Hour of Day	Number of Hits	Data Transferred (Kb)
4	28	189
5	8	42
6	36	520
7	54	316
8	31	867
10	36	4,496
11	73	185
12	48	366
16	4	28

Fig. 10. Popular Hours of the Day Ranked by the Transferred Data.

e) Number of the visits by the visitor

The Fig. 11 shows the number of visits for each visitor in order to concluded the visitors' loyalty and interest according to the number of visitors.

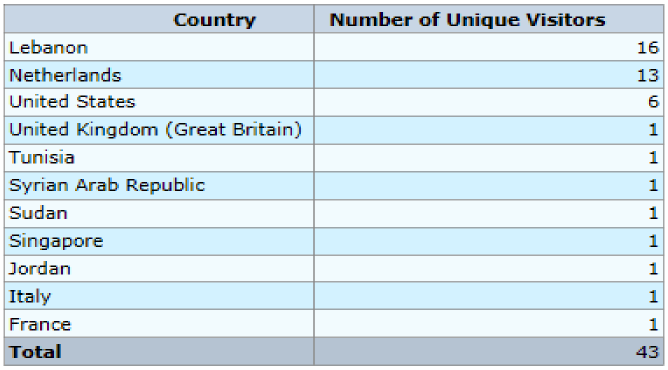


Number of Visits by Visitor	Number of Visitors
1	39
2	4
3-5 Visites	1

Fig. 11. The Number of the Visits Per Visitor.

f) Top visitors countries

The Fig. 12 illustrates the origin countries of the university web visitors ranked by the number of unique visitor from that country.

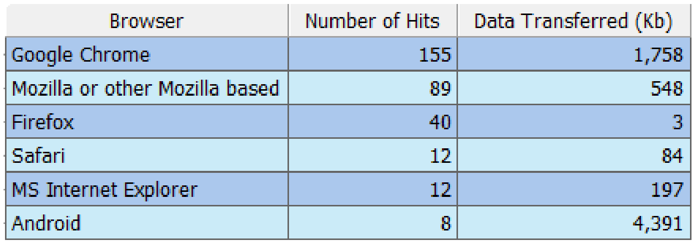


Country	Number of Unique Visitors
Lebanon	16
Netherlands	13
United States	6
United Kingdom (Great Britain)	1
Tunisia	1
Syrian Arab Republic	1
Sudan	1
Singapore	1
Jordan	1
Italy	1
France	1
Total	43

Fig. 12. Top Visitor’s Countries.

g) Browsers

The Fig. 13 displays the web browsers types employed by the visitors ranked by number of hits for each browser that identify the most used ones while accessing our web faculty. Furth more, the data Transferred column in the figure below shows the transferred amount traffic in KB’s from each web browser.

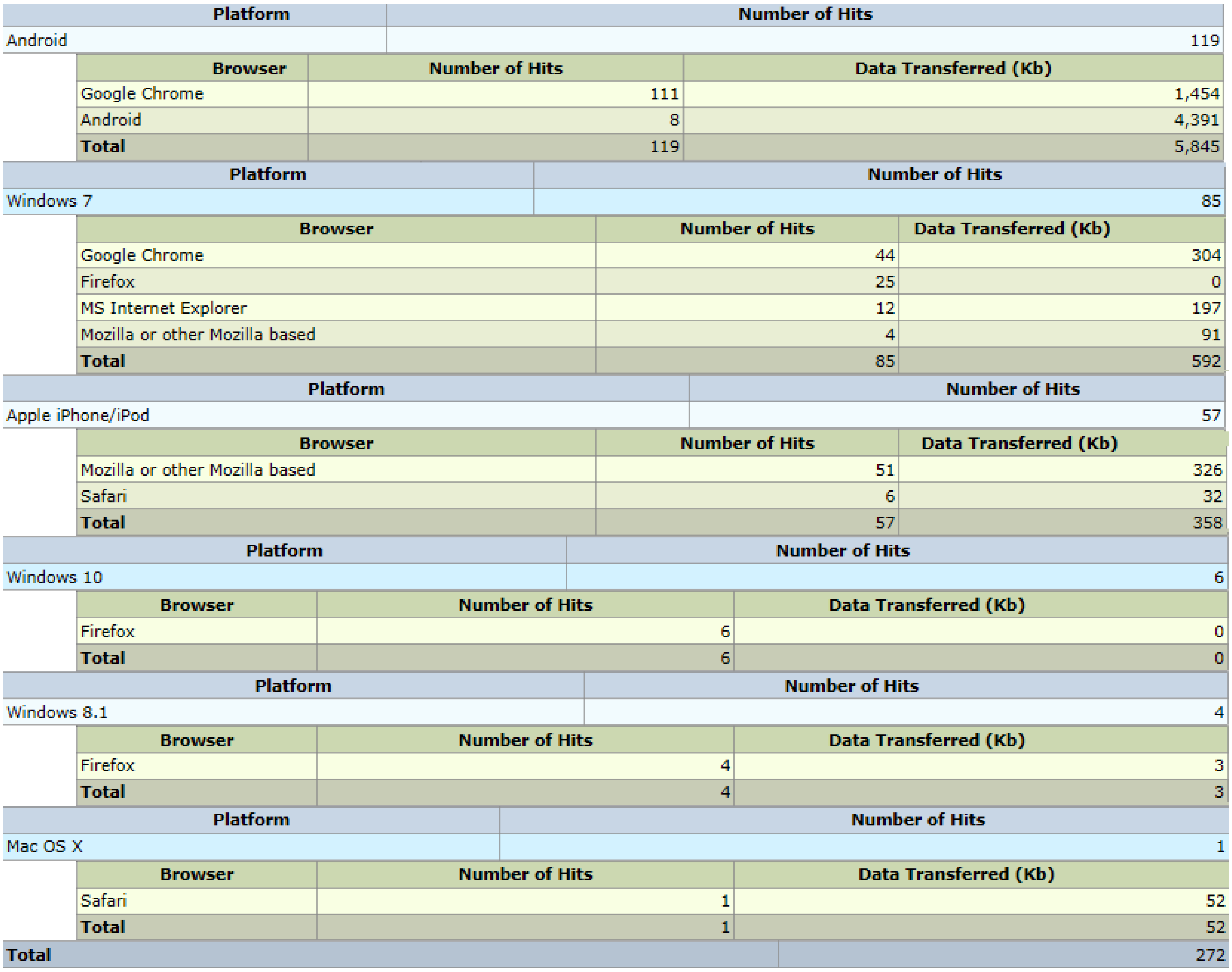


Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	155	1,758
Mozilla or other Mozilla based	89	548
Firefox	40	3
Safari	12	84
MS Internet Explorer	12	197
Android	8	4,391

Fig. 13. The Most used Browsers.

h) Operating system

The report bellow represented by the Fig. 14 illustrates the most used browser with the operating system platform which used to access the web faculty. The installed operating system platforms on the visitor’s computer should be ranked by the number of hits from each OS. On another hand, the data transferred column shows the traffic amount in KB’s transferred to the visitors.



Platform	Number of Hits
Android	119

Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	111	1,454
Android	8	4,391
Total	119	5,845

Platform	Number of Hits
Windows 7	85

Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	44	304
Firefox	25	0
MS Internet Explorer	12	197
Mozilla or other Mozilla based	4	91
Total	85	592

Platform	Number of Hits
Apple iPhone/iPod	57

Browser	Number of Hits	Data Transferred (Kb)
Mozilla or other Mozilla based	51	326
Safari	6	32
Total	57	358

Platform	Number of Hits
Windows 10	6

Browser	Number of Hits	Data Transferred (Kb)
Firefox	6	0
Total	6	0

Platform	Number of Hits
Windows 8.1	4

Browser	Number of Hits	Data Transferred (Kb)
Firefox	4	3
Total	4	3

Platform	Number of Hits
Mac OS X	1

Browser	Number of Hits	Data Transferred (Kb)
Safari	1	52
Total	1	52

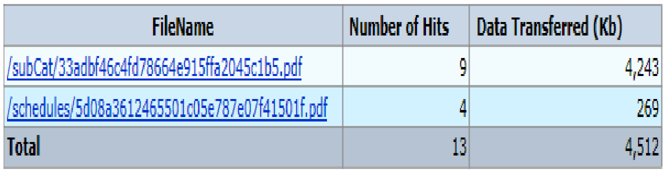
Total	272

Fig. 14. The used Operating Systems Accessed by the Visitor’s Web Browsers.

1) Access resources control and security approaches:

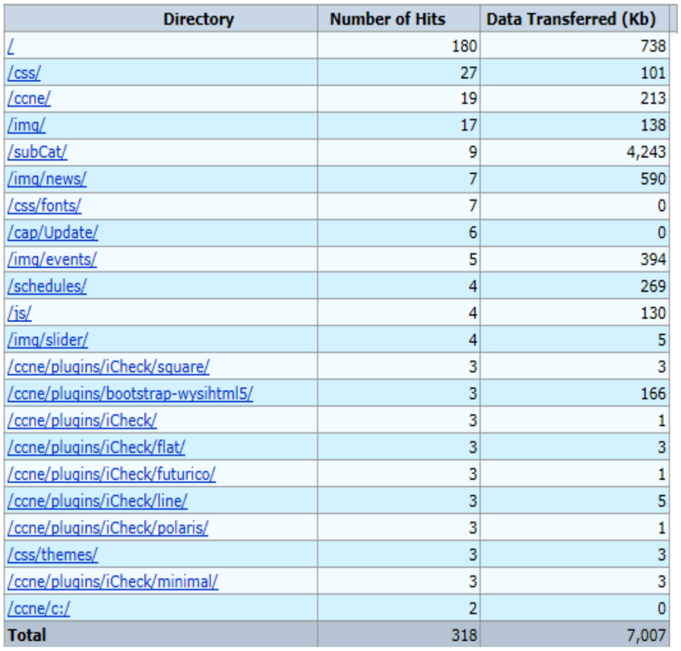
a) Top downloaded files

The 0 shows the popularity of the downloaded files from the faculty website. Downloads are ranked by the number of files that requested by the visitors (number of hits). This figure shows the downloaded files with their specific extensions. For example, these extensions include zip, exe, rar and tar, etc for compressed file, graphics (gif, jpg, etc.), sound (wav, mp3 ...) and video (avi, mpg, mp4...) otherwise the files that are not considered as downloaded file will not appear in this report.



FileName	Number of Hits	Data Transferred (Kb)
/subCat/33adbf46c4fd78664e915ffa2045c1b5.pdf	9	4,243
/schedules/5d08a3612465501c05e787e07f41501f.pdf	4	269
Total	13	4,512

Fig. 15. Top Downloaded Files.



Directory	Number of Hits	Data Transferred (Kb)
/	180	738
/css/	27	101
/ccne/	19	213
/img/	17	138
/subCat/	9	4,243
/img/news/	7	590
/css/fonts/	7	0
/cap/Update/	6	0
/img/events/	5	394
/schedules/	4	269
/js/	4	130
/img/slider/	4	5
/ccne/plugins/iCheck/square/	3	3
/ccne/plugins/bootstrap-wysihtml5/	3	166
/ccne/plugins/iCheck/	3	1
/ccne/plugins/iCheck/flat/	3	3
/ccne/plugins/iCheck/futurico/	3	1
/ccne/plugins/iCheck/line/	3	5
/ccne/plugins/iCheck/polaris/	3	1
/css/themes/	3	3
/ccne/plugins/iCheck/minimal/	3	3
/ccne/c:/	2	0
Total	318	7,007

Fig. 16. The Top Accessed Directories.

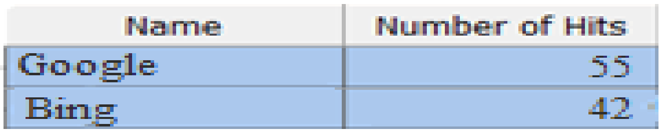
b) Accessed directories

The popularity of the web server directories is declared as shown in the figure below [Fig. 16]. This report is ranked by the number of visitors that requested the web pages or any file located in that directory.

The Data Transferred column shows the total number of Kb's transferred by the visitors of the web server according to the visited directories.

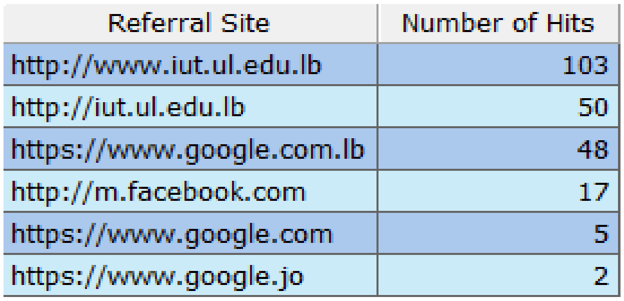
c) Search engines

When a user executes an online search query, the search engine will explore via its searchable index and will returns the results that are related to the desired searcher's query. The outputs are ranked based on the popularity of the website that provides the information. The value and the importance of a website is specified by several factors such as the keywords appearance on the web page, the relevancy of the web page content, the quality of hyperlink, the related social elements (such as Facebook, Instagram, Tweeter likes or shares), and other factors. Therefore, the value of studying the requested search engine is to know the access methods to a website that it is very influential in discovering the effective factors in the website search engine optimization. The figure below [Fig. 17] shows a list of search engines requested by the visitors to find the faculty web site ranked by the number of referrals (Number of Hits column) for each search engine.



Name	Number of Hits
Google	55
Bing	42

Fig. 17. Top used Search Engines.



Referral Site	Number of Hits
http://www.iut.ul.edu.lb	103
http://iut.ul.edu.lb	50
https://www.google.com.lb	48
http://m.facebook.com	17
https://www.google.com	5
https://www.google.jo	2

Fig. 18. Referring WebSites.

d) Referrals website

The Fig. 18 displays the referrer websites that may help to drive the external visitors to our website. These websites ranked by the number of hits received from that referrer.

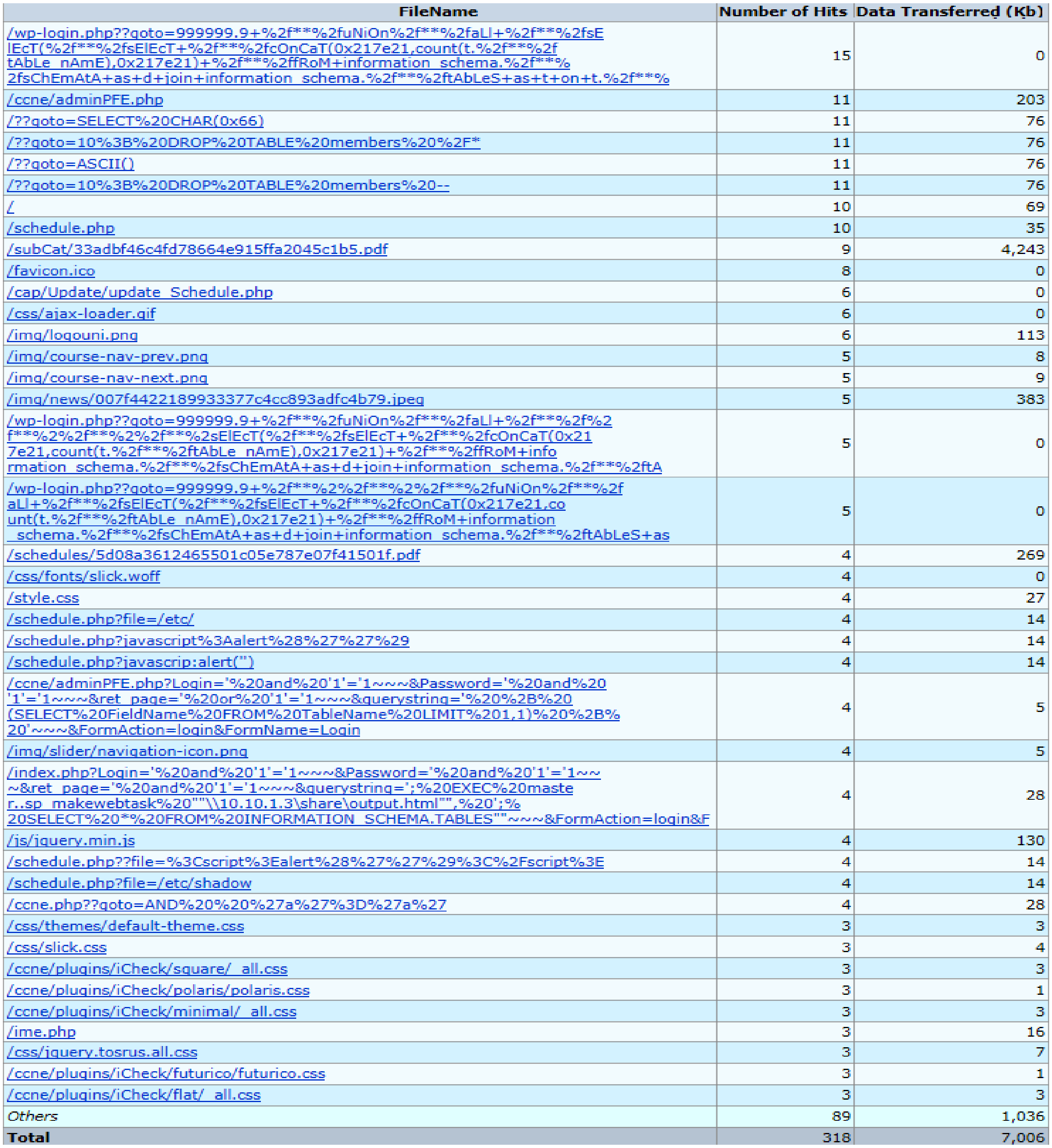
e) Security alert

Providing a website security, mostly controlling the user behavior has become one of the most important concerns of the technological research centers over the past few years. Many academic companies are joining the game in hopes of capitalizing from the research centers to have a secured web server by controlling the accessed resources in it. One of the essential vectors to provide a fundamental security is the Access Resources Control. When we talk about the access control, the researchers must be concerned with respect to the mechanisms to restrict access to a resource. We have to take into consideration about who are the visitors that connect to our website in order to detect the visitors behavioral by controlling the viewed and visited pages as well as all the accessed resources. The figures as shown below 0 and Fig. 20 will detect the popularity of the viewed and visited web pages that ranked by the number of hits and the transferred data of requested pages by the visitors that will highlight the importance of controlling the URL type and the structure form in order to detect the irregular resources (url, page, directory)accessed by the user according to the main resources, as well as extracting a summary about the quality of the visited resources concerning the web pages in order to classify the fearing of that visitors were its behavior can be detected and determined by studying the irregular URL cases (Sql injection, XSS, SSRF, Directory Traversal) ranked by the detected behavior type.

a) Diagnostic

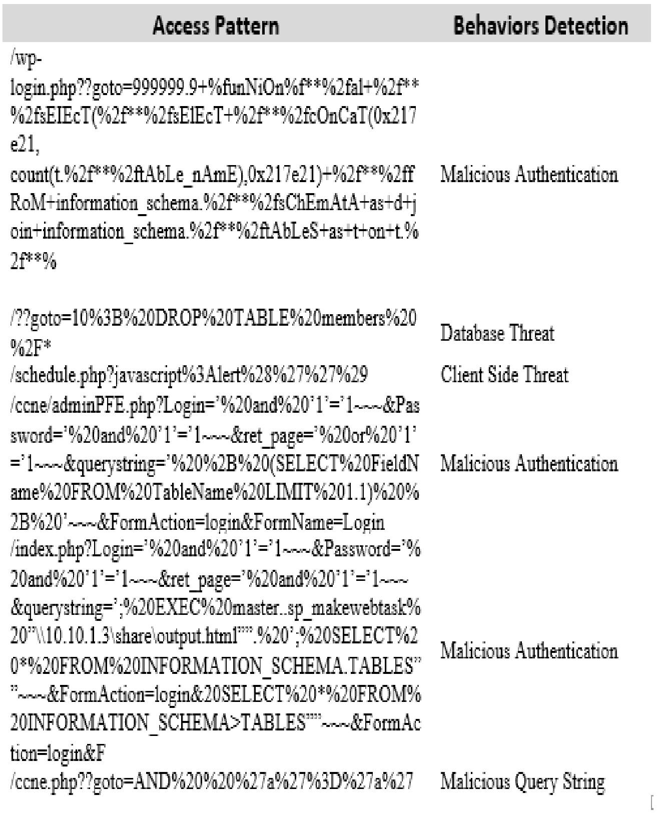
Mainly, the practices of the web usage mining techniques play an essential methodology in tracking the visitor's activities and its relation with respect to the other networks. Web system administrator employs this kind of techniques in the log file in order to monitor the desired network and the web server errors that can permits the identifying of the vulnerabilities that may happen in the web server to access critical and important information known as the cyber security attacks. Moreover, our proposed tool plays the role of detecting the occurred errors using the regular expression technique After analyzing the presented errors, we are able to identify who can play the illegal activities on the web server.

We can conclude from the figure below [Fig. 21] that the error "404" is the most error that occurred on the web server; moreover, we can observe the targeted pages in order to determine and find the best solution to fix the discovered vulnerabilities.



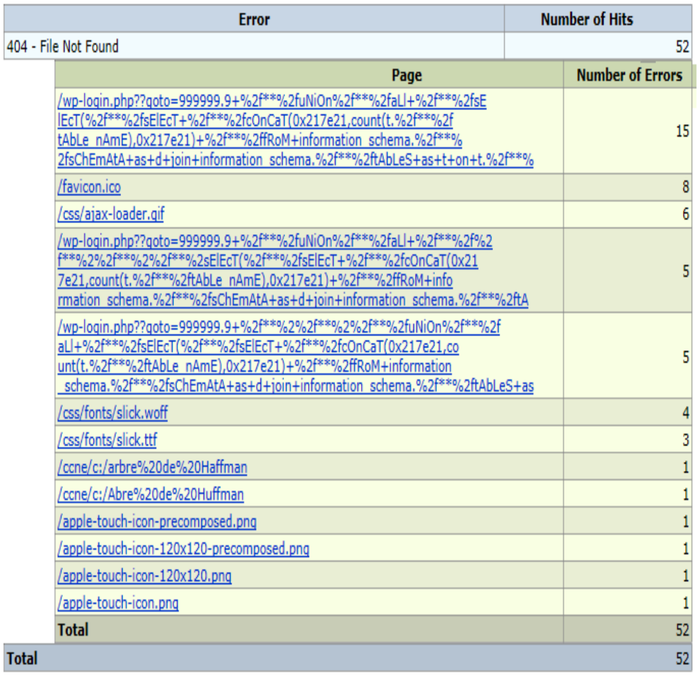
FileName	Number of Hits	Data Transferred (Kb)
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	15	0
/ccne/adminPFE.php	11	203
/??goto=SELECT%20CHAR(0x66)	11	76
/??goto=10%3B%20DROP%20TABLE%20members%20%2F*	11	76
/??goto=ASCII()	11	76
/??goto=10%3B%20DROP%20TABLE%20members%20--	11	76
/	10	69
/schedule.php	10	35
/subCat/33adbf46c4fd78664e915ffa2045c1b5.pdf	9	4,243
/favicon.ico	8	0
/cap/Update/update_Schedule.php	6	0
/css/ajax-loader.gif	6	0
/img/logouni.png	6	113
/img/course-nav-prev.png	5	8
/img/course-nav-next.png	5	9
/img/news/007f4422189933377c4cc893adfc4b79.jpeg	5	383
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2f%2f**%2%2f**%2%2f**%2sElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftA	5	0
/wp-login.php??goto=999999.9+%2f**%2%2f**%2%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as	5	0
/schedules/5d08a3612465501c05e787e07f41501f.pdf	4	269
/css/fonts/slick.woff	4	0
/style.css	4	27
/schedule.php?file=/etc/	4	14
/schedule.php?javascript%3Aalert%28%27%27%29	4	14
/schedule.php?javascrip:alert('')	4	14
/ccne/adminPFE.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20or%20'1'='1~~~&querystring='%20%2B%20(SELECT%20FieldName%20FROM%20TableName%20LIMIT%201,1)%20%2B%20'~~~&FormAction=login&FormName=Login	4	5
/img/slider/navigation-icon.png	4	5
/index.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20and%20'1'='1~~~&querystring=';%20EXEC%20master..sp_makewebtask%20""\\10.10.1.3\share\output.html"",%20';%20SELECT%20*%20FROM%20INFORMATION_SCHEMA.TABLES""~~~&FormAction=login&F	4	28
/js/jquery.min.js	4	130
/schedule.php??file=%3Cscript%3Ealert%28%27%27%29%3C%2Fscript%3E	4	14
/schedule.php?file=/etc/shadow	4	14
/ccne.php??goto=AND%20%20%27a%27%3D%27a%27	4	28
/css/themes/default-theme.css	3	3
/css/slick.css	3	4
/ccne/plugins/iCheck/square/_all.css	3	3
/ccne/plugins/iCheck/polaris/polaris.css	3	1
/ccne/plugins/iCheck/minimal/_all.css	3	3
/ime.php	3	16
/css/jquery.tosrus.all.css	3	7
/ccne/plugins/iCheck/futurico/futurico.css	3	1
/ccne/plugins/iCheck/flat/_all.css	3	3
Others	89	1,036
Total	318	7,006

Fig. 19. Top Accessed Pages and Resources.



Access Pattern	Behaviors Detection
/wp-login.php??goto=999999.9+%funNiOn%f**%2fal+%2f**%2fsEIEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	Malicious Authentication
/??goto=10%3B%20DROP%20TABLE%20members%20%2F*	Database Threat
/schedule.php?javascript%3Alert%28%27%27%29	Client Side Threat
/ccne/adminPFE.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20or%20'1'='1~~~&querystring='%20%2B%20(SELECT%20FieldName%20FROM%20TableName%20LIMIT%201.1)%20%2B%20'~~~&FormAction=login&FormName=Login	Malicious Authentication
/index.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20and%20'1'='1~~~&querystring=';%20EXEC%20master..sp_makewebtask%20"\\10.10.1.3\share\output.html"".%20';%20SELECT%20*%20FROM%20INFORMATION_SCHEMA.TABLES""~~~&FormAction=login&20SELECT%20*%20FROM%20INFORMATION_SCHEMA>TABLES""'~~~&FormAction=login&F	Malicious Authentication
/ccne.php??goto=AND%20%20%27a%27%3D%27a%27	Malicious Query String

Fig. 20. Some of the Detected Behaviours about the Malicious Visiotrs.



Error	Number of Hits
404 - File Not Found	52

Page	Number of Errors
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	15
/favicon.ico	8
/css/ajax-loader.gif	6
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2f%2f**%2%2f**%2%2f**%2sElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftA	5
/wp-login.php??goto=999999.9+%2f**%2%2f**%2%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as	5
/css/fonts/slick.woff	4
/css/fonts/slick.ttf	3
/ccne/c:/arbre%20de%20Haffman	1
/ccne/c:/Abre%20de%20Huffman	1
/apple-touch-icon-precomposed.png	1
/apple-touch-icon-120x120-precomposed.png	1
/apple-touch-icon-120x120.png	1
/apple-touch-icon.png	1
Total	52

Total	52

Fig. 21. The Occurred Web Server Errors.

VII. Conclusion

Nowadays, the Website is considered as the most used means by the internet visitors to collect desired and valuable information. Therefore, the usability and security of a web server resources are very important to provide a website more popular among its visitors.

In this paper, we proposed a methodology on predicting user behavior based on the web mining technique. Our target has accomplished by developing and applying our suggested tools that merge two separate techniques: the web usage mining technique and the cyber security approaches. Through these mechanisms, we can have a potential to create a suitable access patterns in order to detect and identify the system errors which occurred on the web server, as well as the identification of the user's behaviors and their important activities, the potential visitors, the navigation activities and a diagnostic study. Moreover, we can predict the pages that make the errors which can help us to detect the vulnerabilities on our web server, this can be done by controlling all the URL's, directories and web pages in order to provide a security mechanism with respect to the malicious users, specially the abnormal activities, by taking into consideration all the breaches which can be occurred on the faculty web server. Furthermore, our extracted results will be shown as summaries, tables, figures and charts that can be consider as a guide report which discuses each discovered pattern and its behavior, thus can be helpful for the System Administrators, Web Analysts and Website Maintainers to improve and enhance the usability and the security stability of the web server concerning their resources.

As perspective works, our aim is to divide the future work into two parts. The first part will enhance the detection of the visitor behavior, this can make our research more effective by tracking only the actions of interest activities which cause the errors, thus we will discover the accessed pattern in less period of time and with minimum memory system utilization.

On other hand, the second part will focus deeply on the cyber security approaches, it will be released by extending the data extraction period time in order to cover big amount of vulnerable data, this can help to provide a data set in order to develop an intelligent model using the machine learning algorithms that predict the abnormal visitors and the expected attacks.

References

[1] Zhang Huiying, Laing Wei "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining" Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 Hangzhou, P.R.China.

[2] Doru Tanasa et.al Advanced data preprocessing for inter sites Web Usage mining IEEEE computer society 2004.

[3] Kohavi "Mining E-Commerce Data: The Good, the Bad, and the Ugly" KDD 2001, Aug 26-29, San Francisco, CA. Copyright 2001 AC.

[4] Tsuyoshi Murata and Kota Saito "Extracting Users Interests from Web Log Data" Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006 IEEE.

[5] Show-Jane Yen, Yue-Shi Lee and Min-Chi Hsieh. "An Efficient Incremental Algorithm for Mining Web Traversal Patterns" Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05).

[6] A-Nikos Koutsoupias "Exploring Web Access Logs with Correspondence Analysis" 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece, Proceedings, Companion Volume, pp. 229-23.

[7] Meiyappan Nagappan, Malden A.vouk "Abstracting log lines to log event types for mining software system logs" 7th IEEE Working Conference on Mining Software Repositories (MSR 2010),2010.

[8] Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou and Jim Wiltshire "Measuring the Accuracy of Sessionizers for Web Usage Analysis" KDD'99 Workshop on Web Usage Analysis and User Pro_ling WEBKDD'99, San Diego, CA, Aug. 1999. ACM. Springer, LNCS series.

[9] Maristella Agosti Giorgio Maria Di Nunzio "Web Log Mining: A Study of User Sessions "UNIVERSITY OF PADUA Department of Information Engineering.10th DELOS Thematic Workshop on Personalized Access, Prole Management, and Context Awareness in Digital Libraries Corfu, Greece, 29-30 June 2007.

[10] Tsuyoshi Murata and Kota Saito "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining" International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), 2016.

[11] A. Deepa, P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.

[12] Ankita Kusmakar, Sadhna Mishra Web Usage Mining: A Survey on Pattern Extraction from Web Logs" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013 ISSN: 2277 128X,Page:-834-838.

[13] Nina, Shahnaz Parvin, Mahmudur Rahman, Khairul Islam Bhuiyan, and Khandakar Entenam Unayes Ahmed. "Pattern discovery of web usage mining." In Computer Technology and Development, 2009. ICCTD'09. International Conference on, vol. 1, pp. 499-503. IEEE, 2009.

[14] Dhawan, Sanjeev, and Swati Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs." American International Journal of Research in Science, Technology, Engineering & Mathematics (2013): 203-207.

[15] A. Saluja, B. Gour, L. Singh, "Web Usage Mining Approaches for User's Request Prediction: A Survey", IJCSIT-International Journal of Computer Science and Information Technologies, vol. 6, no. 3, 2015.

[16] Sudha Nagesh, "Roll of Data Mining in Cyber Security", Journal of Exclusive Management Science –May 2013-Vol 2 Issue 5 - ISSN 2277 – 5684.

[17] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai" Analysis of web logs and web user in web mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[18] G. K. Lekeas, "Data mining the web: the case of City University's Log Files," 2000.

[19] K. Suneetha and R. Krishnamoorthi, "Identifying user behavior by analyzing web server access log file," IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 327–332, 2009.

[20] K. Etminani, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method," IFSAEUSFLAT, pp. 396–401, 2009.

[21] Ratnesh Kumar Jain, Dr. R. S. Kasana1, Dr. Suresh Jain, "Efficient Web Log Mining using Doubly Linked Tree," International Journal of Computer Science and Information Security, IJCSIS, vol. 3,2009.

[22] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for," International Journal of Data Mining Techniques and Applications, vol. 02, no. 01, pp. 141–150, 2013.

Lightweight and Optimized Multi-Layer Data Hiding using Video Steganography Paper

Samar kamil[1], Masri Ayob Authors[2]
CAIT, Faculty of Information
Science and Technology,
Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia

Siti Norul Huda Sheikh Abdullah[3]
Cyber Security, Faculty of
Information Science and Technology
Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia

Zulkifli Ahmad[4]
School of Language Studies and
Linguistics, Faculty of
Social Sciences and Humanities
Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia

***Abstract*—The ever-escalating attacks on the internet network are due to rapid technological growth. In order to surmount such challenges, multi-layer security algorithms were developed by hybridizing cryptography and steganography techniques. Consequently, the overall memory size became enormous while hybridizing these techniques. On the other side, the least significant bit (LSB) and modified LSB replacing approaches could provide the variability as detected by steganalysis technique, most found to be susceptible to attack too due to numerous reasons. To overcome these issues, in this paper a lightweight and optimized data hiding algorithm is proposed which consume less memory, provide less variability, and robust against histogram attacks. The proposed steganography system was achieved in two stages. First, data was encrypted using lightweight BORON cipher that only consumed less memory as compared to conventional algorithm such as 3DES, AES. Second, the encrypted data was hidden in the complemented or non-complemented form to obtain minimal variability. The performance of the proposed technique was evaluated in terms of avalanche effect, visual quality, embedding capacity and peak signal to noise ratio (PSNR). The results revealed that the lightweight BORON cipher could produce approximate same avalanche effect as the AES algorithm produced. Furthermore, the value of PSNR had shown much improvement in comparison to optimization algorithm GA.**

Keywords—Video steganography; least significant bit technique; optimized data hiding; cloud computing; boron cipher

I. Introduction

Nowadays, the internet is the most used medium to access desired information. However, the internet misused activities or cybercrimes such as security breaches [1-19] are also increasing exponentially. In order to overcome these attacks, cryptography and steganography algorithms are used. In cryptography algorithms, secret data are scrambled but the encrypted output stream leaves clues to attackers. On the other hand, steganography algorithms conceal the visibility of secret messages by hiding them in cover media. Unfortunately, this technique could also be compromised by using statistical tests. Therefore, to overcome these issues and to provide multi-layer data security, hybridizing cryptography with steganography algorithms could strengthen data security.

In cryptography algorithm, secret data bits were altered in such a way that only trusted persons who have the key to the file can have access to the data. This technique is classified as private and public. In privateclass, same key is used for encryption and decryption purposes whereas in public class, different keys are used for encryption and decryption procedures [2].

In steganography algorithm, symmetric as well as asymmetric algorithms are gaining more popularity now. The most preferred encryption algorithm for steganography are 3DES, AES, Blowfish, RSA (Rivest, Adi Shamir and Leonard Adleman), and ECC (Elliptic Curve Cryptography) [3-7-20]. These algorithms provide more efficient security but they require large number of resources. Therefore, lightweight ciphers are studied such as BORON cipher [16] for encryption purpose as shown in Table 1. The table is show that the lightweight cipher consume minimum area for s-boxes as compared to other the conventional ciphers but increase the computation time for encryption due to the large number of rounds

TABLE I. Comparative of Encryption Algorithm

Algorithm	Block Size (bits)	Key Size (bits)	Number of Rounds	Number of S-boxes
Data Encryption Standard (3DES)	64	3Keys-56	16	4-64 entry S-box
Advanced Encryption Standard (AES)	128	128/192/256	10/12/14	256 Entry S-Box (Each position 8-bit long)
BLOWFISH	64	32-448	16	4-256 entry S-Box
BORON (Lightweight Cipher)	64	80/128	25	16 entry s-box (Each position 4-bit long)

TABLE II. MULTIMEDIA FILES AND THEIR CHARACTERISTICS

Multi-Media Files	Characteristics
Text	Line/Word Shifting Encoding
Protocol	Packet Payload and Packet Header
Audio	Phase Coding, Spread spectrum, Low-Bit Encoding
Image	Image Pixels
Video	Frame and Audio

In steganography, choosing multimedia file for data embedding has played an important role. The multimedia files include text, protocol, audio, image, and the video [3-18]. Various characteristics of these multimedia files have been used for data hiding as shown in Table 2. Images and videos are more popular in comparison to other media because they contain higher number of pixels information and could conceal secret information in a more organized manner. Further, the steganography embedding domain has also an important parameter for data embedding. The embedding domain is classified into spatial and transform domain. In spatial domain, cover media data/pixels are used to conceal information. In transform domain, cover media is changed into other signal/form for data hiding procedure. In this paper, video files and the spatial domain are used for data embedding.

In steganography, LSB and its variant are the most techniques used for data hiding due to its implementation simplicity, low complexity, and high embedding capacity [4-5-17]. The replacement of cover media LSB bits with secret data bits has resolved the visual quality problem. For reduction of variability in cover pixels, a number of techniques such as LSB matching [6-7], optimization technique GA, PSO [8-10-21-22], and optimal position to match with secret data bits [11] have been used in the past

When suitable match is not found, the algorithm hides the secret data in LSB bits. Hence, the visual quality is maintained.

In this paper, lightweight and optimized multi-layer data hiding technique is designed for video steganography. In videos, number of frames are available which improve data hiding capacity and security. The proposed technique is improve visual quality, enhanced security by hiding secret data bits in random frames in less execution time. In this proposed technique, BORON cipher is used for secret data encryption procedure. The encrypted secret data bits incomplemented or non-complemented form are determined and matched with cover frame LSB bits. The matched combination index is determined and based on index information where the encrypted data is decrypted at the receiver side. Next, the proposed technique analysis is conducted to show the security level of the lightweight cipher and the visual quality of frames after the data hidingprocedure. Finally, a counter measure technique is proposed to resolve any issues pertaining to the result of the analysis as defined in table 3.

TABLE III. ISSUES AND COUNTER MEASURE

Issues	Counter Measure
Multi-layer security algorithms are increased overall area.	Preferred Lightweight Cipher which consume less area
Steganography LSB replacing techniques is provide maximum variability and easy to break.	Optimized data hiding is done in complemented or non-complement form.

The rest of the paper is organized as follow: Section II defined related works regarding secret data encryption and optimized data hiding techniques. Section III illustrates the proposed technique in details. Section IV presents the experimental results and Section V states the conclusion of the work.

II. RELATED WORK

In this section, the cryptography algorithms, spatial domain and optimal match data hiding techniques are studied.

Yadav, et al. [19] used a key named XOR operation for data encryption. Next, they used sequential encoding and LSB techniques for data embedding and data hiding subsequently. Even though the encryption process required less time as compared to 3DES, AES and Blowfish but it is still considered relatively easy to break. **Apau, *et al.* [6],** designed multilayer security in spatial domain using RSA, Huffman coding, and LSB technique. In their work, secret data are encrypted using asymmetric algorithm namely RSA. Firstly, the data are compressed using lossless technique specifically Huffman code. Next, Huffman code compression lossless technique provides cover data size reduction without causing any data loss. Finally, the secret data are then hidden using LSB technique. Although RSA algorithm provides effective encryption as well as authentication, its security are dependent on large key size thus and consumes huge amount of time for encryption process. **Ramakrishna Hedge and Jagadeesha S [7],** employed ECC and optimization in their work. Their method uses data encryption via ECC (Elliptic Curve Cryptography) algorithm and its data embedment is in the form of H.264 videos. They deployed artificial bee colony (ABC) algorithm to reduce variability and to find the best position in the data embedding procedure. They also took advantage of ECC technique due to its smaller key size and less storage requirement. Hence, the technique succeeds in improving overall processing speed for data encryption. **Mstafa, et al. [5]** improve embedding capacity and robustness of security system against attacks by hybridizing spatial domain of various LSB techniques (1-bit, 2-bit, 3-bit, 4-bit LSB) with hamming codes (15, 11). In their experimental setup, they have four stages. In the first stage, secret messages are pre-processed using hamming algorithm. In the second stage, Region-of-Interest (ROI) is detected from the cover videos for data embedding procedure. In the third stage, data embedding is performed using various LSB techniques. In the fourth stage, data extraction from stego video is achieved. Their result shows that they have high embedding capacity and enhanced visual quality.

In order to reduce variability, Mielikanien [6] used LSB matching technique for data hiding. In LSB matching technique, if secret bits do not match with the cover bits, then ±1 is randomly added in the corresponding pixels. The LSB match technique provides better visual quality and similar embedding capacity as compared to LSB replacement technique. However, the LSB matching technique deals with given pixel/pixel pair without considering the difference between the pixel and its neighbour [7]. Dasgupta et al. [8] split the data bits into 3:3:2 ratio to improve embedding capacity and create less variability as compared to other ratio. They used genetic algorithm to search optimal position in video steganography. Moreover, they used optimization algorithms such as GA [9] and PSO [10] to find optimal matches for data embedding procedure. However, these algorithms search optimal matches in cover pixels using a number of iterations that cause an increase in computational time. Next, [11], used a technique which search cover pixel bits 0 to 7 to find optimal matches for secret data bits. When optimal matches are found, the index position will be determined. Otherwise, the data bits will be hidden in the LSB position. The secret data bits and index information are hidden in the stego media. The index position varies 0 to 7 and 3 bits of pixel is used to hide the index information and to provide high variability (maximum variability $2^3=8$). In order to reduce variability, in which data is embedding in complemented or non-complemented form provide zero variability for embedding secret data. Further, hide to their index only 2 bits are required which provide maximum variability $2^2=4$.

Their contribution for data encryption preferred 3DES, AES, Blowfish, RSA, and ECC algorithms. Even though, these algorithms provide efficient security, they consume large number of resources in terms of memory space. In order to overcome these issues, lightweight ciphers are introduced to provide the same level of security and consume less memory space. On the other hand, techniques that use optimal match search do provide better visual quality but they increase the computational time, index information bits, and unable to distinguish the difference between pixels and its neighbours [6, 8-10,18].

Therefore, in this paper after considering all these parameters and issues, proposed complemented or non-complemented form technique for data hiding procedure. This data hiding procedure embeds the secret data in the LSB bits of the pixel either in complemented or non-complemented form. It provides zero variability, high embedding capacity and enhanced security level.

III. Proposed Technique

In this section, the proposed technique block diagrams and its components are explained. The block diagram of data hiding technique is shown in Fig 1.

The secret data is encrypted using BORON cipher. Next, the video frames, encrypted secret data, and Look-up Table are input parameters for data hiding procedure. The Look-up Table shows the number of times taken to determine complemented or non-complemented secret data to find complete matches with cover pixel bits. In this proposed technique, about half of the frame contains encrypted secret data and another half contained index information which counts the number of times to achieve the complemented form of data bits.

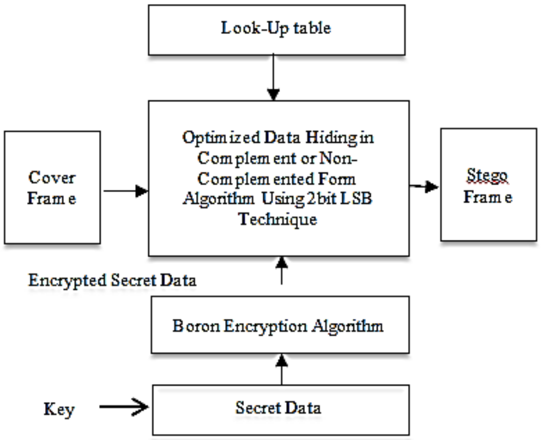

Look-Up table
Cover Frame
Optimized Data Hiding in Complement or Non-Complemented Form Algorithm Using 2bit LSB Technique
Stego Frame
Encrypted Secret Data
Boron Encryption Algorithm
Key
Secret Data

Fig. 1. Block Diagram for Data Hiding Technique.

The extraction process block diagram is in Fig. 2. In the extraction process, the Stego Frame and Look-up Table (index information regarding complemented or non- complemented data bits to the cover bits) serve as inputs to the Data Extraction algorithm. The Data Extraction algorithm will then extract original Secret Data. The detail description of block diagram components is explained below.

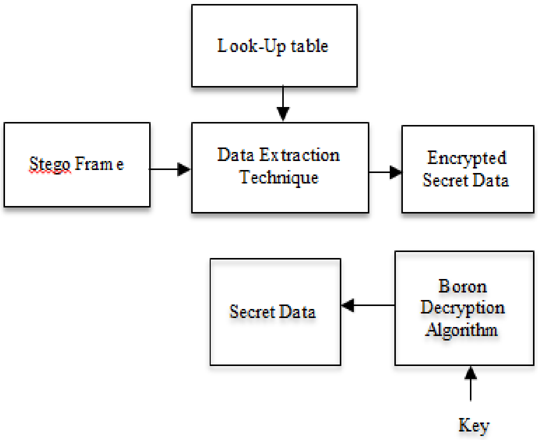


Fig. 2. Block Diagram for Data Extraction.

A. *Secret Message Encryption using Lightweight Boron Cipher*

The BORON cipher is a lightweight block cipher. It is based on substitution-permutation network. This cipher has 64-bit block size, two key variant of size 80 and 128 bit, and total 25 rounds [16]. The lightweight algorithm has block size 64 bit and process in 4-bit chunk. Therefore, 24 = 16 is the combination required in the s-box. On the other side, In the conventional technique, the block size is 128 bit and process the data in 8-bit and 28= 256 combination is required in the s-box implementation which increase memory requirement. The encryption pseudo code for the BORON cipher is explained in Figure 3.

```
Encryption Algorithm
For m is 0 to 24 rounds
{XOR Operation(State, Key)
S-Box Layer(State)
Block Shuffle(State)
Left Circular Shift (State)
XOR Operation (state)
Key Scheduling (Key)}
XOR Operation (State, Key)
Key Scheduling Algorithm
For 80-bit Key
Left Circular shift(Key, 13)
S-Box Layer($Key_{0-3}$)
XOR Operation ($Key_{63-59}$, $Round\_Counter$)
For 128-bit Key
Left Circular shift(Key, 13)
S-Box Layer($Key_{0-7}$)
XOR Operation ($Key_{63-59}$, $Round\_Counter$)
```

Fig. 3. Pseudo-Code for BORON Cipher.

Furthermore, the lightweight Boron cipher has large number of rounds such as 25 as compared to AES which has 10 rounds for encryption. Therefore, the time required to encrypt data using BORON cipher is higher than AES. To improve the computational time of encryption to the algorithm, optimization algorithms such as loop-unrolling is preferred [16].

B. Encrypted Secret Message Complemented or Non-Complemented Form

The encrypted secret message pixels are broken into 2-bit using logical operations. The encrypted secret bits complemented or non-complemented matrix is formed as shown in Table 4. Table 4 shows that indexes are defined according to the number of times taken to form complements. For example, the 0th index shows no complement is taken, 1st index shows first bit of the encrypted secret bit is complemented. In the 2nd index, the second bit of the encrypted secret bit is complemented. In the 3rd index, both of the encrypted secret bits are complemented.

TABLE IV. LOOKUP TABLE

0th Index	1st Index	2nd Index	3rd Index
00	01	10	11
01	00	11	10
10	11	00	01
11	10	01	00

TABLE V. DATA HIDING TECHNIQUE

Cover Pixels	10010100	11110000	11001100	10010001
Stego Pixels	100101**00**	111100**00**	110011**00**	100100**01**
Index	2	1	2	1

C. Optimized Data Hiding of Encrypted Data

For the optimized data hiding, LSB 2 bits of the cover pixel are extracted and compared with the encrypted secret bits. The index information determines where the encrypted secret bits matches with cover bits are shown in Table 5

For example

The encrypted data bits: 10 01 10 00

The 10 pixels of the encrypted secret bit is compared with cover pixel bit 00. Based on the matches bits in the Look-up Table, index 2 is stored and zero variability in stego pixels.

In figure 4 the algorithm for the proposed technique is given

Transmitter Side
1.Read the video and extract the frames.
2. Read the secret data, key and encryption using BORON Cipher.
3. Read the encrypted secret messages and divide them into chunks. Each chunk size is 2bit.
4. For data embedding, the cover 2 LSB bits are compared to the encrypted secret 2 bits which are based on the Look-up Table the matches are found an index values are determined.
5. The indexes are hidden in the cover frame using 2 bit LSB technique.
6. The performance analysis is conducted using various parameters such as embedding capacity, PSNR, SM, BER.
7. The stego frames are combined together to form stego video.
Receiver Side
1. Read the stego video and extract the frames.
2. Determine pixels in which indexed bits are hidden.
3. The index information is compared with stego 2 LSB bits and is based on the Look-up Table matches to determine encrypted secret messages.
4. The encrypted secret messages and key input to BORON decryption module enable the retrievement of the secret messages.

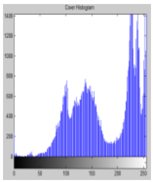
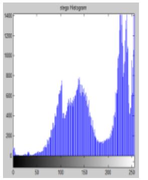
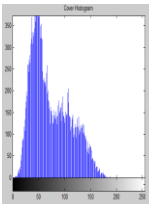
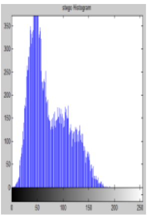
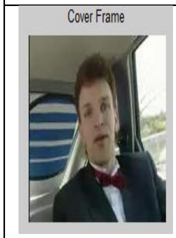
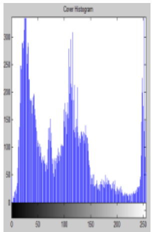
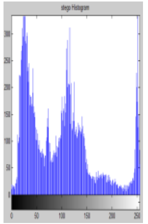
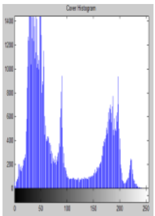
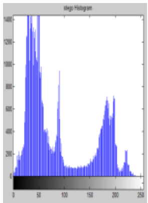
Fig. 4. Algorithm for the Proposed Technique.

IV. EXPERIMENTAL RESULTS

In the experimental setup, various video sequences which includes: Foreman (352×288), Salesman, Car Phone 176×144) and an akiyo video (352×264) are taken [12]. The MATLAB version 8.1.0.604 (R2013a) is used for coding purposes. In Table 6, first frames of all the videos, visual effect between the cover and stego frames are shown after performing the data hiding process. The results show that the histogram looked similar between the cover and stego frames

The performance analysis of the proposed technique is done based on parameters avalanche effect, invisibility, embedding capacity, and robustness against attacks. The parameters are explained below:

TABLE VI. VISUAL COMPARATIVE ANALYSIS BETWEEN COVER AND STEGO FRAME

Cover Frame	Cover Histogram	Stego Frame	Stego Histogram
			
			
			
			

1) Avalanche effect: This parameter is defined the strength of the encryption algorithm. In the ideal scenario, 50% change in the ciphertext required while changing in the one bit of the key. For the BORON cipher, on standard dataset test vector avalanche effect is determined using equation (1) and found that it is provides 50% as shown in table 7.

$$Avalanche\ Effect = \frac{Number\ of\ bits\ Changed}{Block\ Size} \qquad (1)$$

TABLE VII. AVALANCHE EFFECT

Plaintext (In Hex Value)	Key	Number of bits changed
0000 0000 0000 0000	0000 0000 0000 0000 0000	32
	0001 0000 0000 00 0	

2) Embedding capacity: The embedding capacity depends on how much information bits are embedded in the cover frame [13]. It is calculated using equation (2)

$$EC = \frac{Total\ Number\ of\ Bits\ Embedded}{Size\ of\ the\ Cover\ Frame}(\text{bpp}) \qquad (2)$$

Here, bpp represents bits per pixel

A. Peak Signal to Noise Ratio (PSNR)

The invisibility of the proposed technique is measured based on the visual quality. The PSNR parameter is used to measure the distortion in stego frame after data embedding [14]. The decibel unit is used to measure PSNR. It is calculated by using equation (3-4).

$$PSNR = 10 \times \log_{10} \frac{255 \times 255}{MSE} \qquad (3)$$

Here,

$$MSE = \frac{1}{J \times K} \sum_{m=1}^{J} \sum_{n=1}^{K} (X_{m,n} - Y_{m,n})^2 \qquad (4)$$

Here, J, and K defines as the row and column of the frame. The X and Y represent the cover and stego frame.

B. Normalized Cross-Correlation (NCC)

The closeness or similarity between cover and stego frame is determined using this parameter. The value lies between -1 to 1. In the ideal case, NCC value 1 is required. It is measured using equation (5)

$$NCC = \frac{\sum_{J=1}^{M} \sum_{K=1}^{N} C(J,K) \times S(J,K)}{\sum_{J=1}^{M} \sum_{K=1}^{N} C(J,K)^2} \qquad (5)$$

3) Average difference: This parameter is given an actual difference between cover and stego frames as mentioned in equation (6)

$$verage\ Difference = \sum_{m=1}^{J} \sum_{n=1}^{K} \frac{(C(i,j) - S(i,j))}{JXK} \qquad (6)$$

Here, J and K represent the row and column of the frame. The average difference varies in the interval from -255 to 255.

4) Maximum difference (MD): The maximum difference parameter measures the magnitude difference between cover and stego frame. Its value varies from 0 to 255 and determined using equation (7).

$$MD = Max|(C(J,K) - S(J,K))| \qquad (7)$$

Here, C and S represented the cover and stego frame. J and K total number of rows and column.

5) Normalized absolute error (NAE): This parameter is measured by the absolute error between cover and stego frames. It is determined using equation (8).

$$NAE = \frac{\sum_{J=1}^{M} \sum_{K=1}^{N} |C(J,K) - S(J,K)|}{\sum_{J=1}^{M} \sum_{K=1}^{N} |S(J,K)|} \qquad (8)$$

The performance analysis for the propsed technique is show in table 8

6) Similarity index measure (SIM): The similarity index measure (SIM) parameter is used to evaluate the performance of the proposed technique by determining how much information has been extracted after attacking process [15] as mentioned in equation (9).

$$IM = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} J(i,j) K(i,j)}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} J^2(i,j)} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} K^2(i,j)}} \qquad (9)$$

TABLE VIII. PERFORMANCE ANALYSIS FOR THE PROPOSED TECHNIQUE

Parameters	**Videos**			
	Foreman	Salesman	Carphone	Akiyo
Embedding Capacity	1	1	1	1
$PSNR_{Average}$(dB)	50.18	50.19	50.24	50.22
Normalized Cross Correlation	0.999	0.999	0.996	1
Average Difference	0.31	0.31	0.30	0.31
Maximum Difference	3	3	3	3
Normalized Absolute Error	$1.78X10^{-3}$	$4.24X10^{-3}$	$3.15X10^{-3}$	$3.54X10^{-3}$

7) The bit error rate (BER): BER is used to measure how much bits are changing between original and extracted secret message after attacking process. Below is the respective equation (10),

$$BER = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} J(i,j) \oplus K(i,j)}{M \times N} X100\% \qquad (10)$$

Here, J and K represent the original and extracted secret message and M and N represent the size of the secret message.

In the ideal scenario, SIM 1 and BER 0 are required. The various attacks on the proposed technique for video 1 is shown in Table 9. The table shows that SIM and BER are acceptable for salt & pepper attack and highly influenced for other attacks.

TABLE IX. VARIOUS ATTACK ON THE PROPOSED TECHNIQUE

Attacks	**SIM**	**BER%**
Salt & Pepper attack (D=0.1)	0.954	76
Median Filtering	0.834	341
Sharpening	0.816	395
Histogram Equalization	0.78	392
Gaussian attack (mean=0, Variance=0.1)	0.82	402

V. COMPARATIVE ANALYSIS WITH EXISTING TECHNIQUE

The video frames are basically a collection of the images. Also, in the literature number of optimization algorithm results available for the images. Therefore, the proposed technique is run on standard dataset images and compared with existing technique which prefer GA algorithm for optimized datahiding [11]. The most preferred performance parameters such as PSNR and embedding capacity are determined and compared in table (10-12). The proposed technique has better embedding capacity and approximate same PSNR as compared to existing techniques. Because, the index in the proposed technique for searching optimal combination varies from 0 to 4(00 to 11) where in the existing techniques such as GA the index position varies from 0 to 7 for searching optimal position. If matches are not found then the data will be hidden in the LSB.

TABLE X. COMPARISON OF THE PROPOSED TECHNIQUE WITH EXISTING TECHNIQUE BASED ON PSNR VALUE

Cover Image (.jpg)	**Shah, P.D. and Bichkar [11] PSNR (dB)**	**Proposed Technique PSNR (dB)**
Baboon	54.43	50.21
Lena	52.33	50.62
Barbara	53.80	50.25
Cameraman	52.36	50.01

TABLE XI. COMPARISON OF THE PROPOSED TECHNIQUE WITH EXISTING TECHNIQUE BASED ON EMBEDDING CAPACITY

Parameter	**Shah, P.D. and Bichkar [11]**	**Proposed Technique**
Embedding Capacity for Secret Message	Quarter part of the cover image	half part of the cover image

TABLE XII. T-TEST: PAIRED TWO SAMPLE FOR MEANS FOR THE PROPOSED AND SHAH & BICHKAR [11] TECHNIQUES

Parameters	**Shah & Bichkar[11]**	**Proposed Method**
Mean	53.23	50.27
Variance	1.11	0.065
P(T<=t) one-tail	0.001	
P(T<=t) two-tail	0.013	

Referring to Table 12, we assume that null hypothesis is no significant difference between the proposed and Shah & Bichkar [11] methods. Since the p – value of both one (0.001) and two tail (0.13) are less than (p<0.05), we reject the null hypothesis and conclude that there is a significant difference between the proposed and Shah & Bichkar [11] methods

VI. CONCLUSION

In this paper, a multi-layer data hiding technique is proposed for video steganography. The videos have large number of frames which improve embedding capacity and security. In the multi-layer data hiding techniques, conventional encryption algorithms are used for data encryption which consumes large memory for data encryption. To overcome this issue, lightweight algorithms are studied which consume less memory but influence on the computation time. To reduce computation time software optimization technique loop-unrolling is used. Furthermore, optimal match techniques increase computational time and provide small variability when optimal matches are not found. In order to overcome this issue complemented or non-complemented technique is proposed which provide less computational time and zero variability for secret data. The proposed technique is applied on standard dataset videos and is found to perform better in terms of avalanche effect, PSNR, Embedding Capacity, Normalized Cross Correlation, Average Difference, Maximum Difference, Normalized Absolute Error. In addition, from the analysis of SIM and BER, the spatial domain is found to be highly influenced by noise. Hence, in the future, the proposed technique is hybridized with error correction code. Furthermore, the comparative analysis shows that the proposed technique is much better in comparison to existing techniques [11]. In the future, to improve the robustness of proposed technique, the technique is hybrid with error correction codes.

ACKNOWLEDGMENT

This work was supported by UniversitiKebangsaan Malaysia grant Dana Impak Perdana (DIP-2014-039) and AP 2017-005/2 for supporting this project.

REFERENCES

[1] Sebastian Neuner, Artemios G. Voyiatzis, Martin Schmiedecker, Stefan Brunthaler, Stefan Katzenbeisser, and Edgar R. Weippl, “Time is on my side: Steganography in filesystem media,” *Digital Investigation*, vol. 18, pp.576-586, 2016.

[2] Mamta Jain, Saroj Kumar Lenka, Sunil Kumar Vsistha, “Adaptive circular queue image steganography with RSA crypto-system,” *Perspective in Science*, vol. 8, pp. 417-420, 2016.

[3] Mehdi Hussain, Ainuddin Wahid Abdul Wahab, Yamani Idna Bin Idris, “Image Steganography in spatial domain: A survey,” *Signal Processing: Image Communication*, vol. 65, pp. 46-66, 2018.

[4] Khan Muhammad, Jamil Ahmad, Seungmin Rho, Sung Wook Baik, “Image Steganography for authenticity of visual contents in social networks,” *Multimedia Tools and Application*, vol. 76, pp. 18985-19004, 2017.

[5] Mstafa, Ramadhan J., and Khaled M. Elleithy “A video steganography algorithm based on Kanade Lucas Tomasi Tracking algorithm and error correcting codes,” *Multimedia Tools and Application*, vol. 75, pp. 10311-10333, 2016.

[6] J. Mielikainen, “LSB matching revisited,” *IEEE signal processing letter*, vol. 13, issue 5, 2006.

[7] Luo, W., Huang, F. and Huang, J., “Edge adaptive image steganography based on LSB matching revisited,” *IEEE transactions on information forensics and security*, vol.5, issue 2, pp.201-214, 2010.

[8] Dasgupta, K., Mondal, J.K. and Dutta, P., “Optimized video steganography using genetic algorithm (GA)”. *Procedia Technology*, vol. 10, pp.131-137, 2013.

[9] Wang, S., Yang, B. and Niu, X., “A secure steganography method based on genetic algorithm,” *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, issue 1, pp.28-35, 2010.

[10] Li, X. and Wang, J.,“A steganographic method based upon JPEG and particle swarm optimization algorithm,” *Information Sciences*, vol. 177, issue 15, pp.3099-3109, 2007.

[11] Shah, P.D. and Bichkar, R.S., “A Secure Spatial Domain Image Steganography Using Genetic Algorithm and Linear Congruential Generator” *In International Conference on Intelligent Computing and Applications*, pp. 119-129. Springer, Singapore, 2018.

[12] http://trace.eas.asu.edu/yuv/index.html

[13] Sadek, M.M., Khalifa, A.S. and Mostafa, M.G., “Robust video steganography algorithm using adaptive skin-tone detection,” *Multimedia Tools and Applications*, vol. 76, issue 2, pp.3065-3085, 2017.

[14] Kumar, V. and Kumar, D., “A modified DWT-based image steganography techniques”, *Multimedia Tools and Applications*, vol. 77, issue 11, pp.130279-130308, 2017.

[15] He Yingliang, Yang Gaobo, Zhu Ningbo, “A real-time dual watermarking algorithm of H.264/AVC video stream for Video-on-Demand service,” *International Journal of Electronics and Communication*, vol. 66, pp. 305-312, 2012.

[16] Bansod, Gaurav, Narayan Pisharoty, and Abhijit Patil, ‘BORON : An Ultra-Lightweight and Low Power Encryption Design for Pervasive Computing’, Frontier of Information Technology & Electronic Engineering, 18 (2017), 317–31

[17] Majeed, Mohammed Abdul, and Rossilawati Sulaiman, "An Improved Lsb Image Steganography Technique Using Bit-Inverse In 24 Bit Colour," Journal of Theortical and Applied Information Technology, 2016

[18] Ali, Ahmed Hussain, Mohd Rosmadi Mokhtar, And Loay E George, "Enhancing The Hiding Capacity Of Audio Steganography Based On Block Mapping Enhancing The Hiding Capacity Of Audio," Journal of Theoretical & Applied Information Technology, vol.95, no.7, 2017.

[19] Ali, Ahmed Hussain, And Loayedwar George, "A Review On Audio Steganography Techniques," Research Journal of Applied Sciences, Engineering and Technology, vol.12, No. 2, pp. 154-162. 2016

[20] Othman, I O R, And Tructure In, "Key Exchange In Elliptic Curve Cryptography Based On The Decomposition Problem," Sains Malaysiana, vol. 41, pp. 907–10, 2012.

[21] Hussein, Wasim Abdulqawi, And Shahnorbanun Sahran, ‘An Improved Bees Algorithm For Real Parameter Optimization’, Int J Adv Comput Sci Appl, vol. 6, pp. 23-39, 2015.

[22] Abdul, Rafidah, Aziz Masri, Ayob Zalinda, Othman Zulkifli, And Nasser R Sabar, ‘An Adaptive Guided Variable Neighborhood Search Based On Honey-Bee Mating Optimization Algorithm For The Course Timetabling Problem’, Soft Computing, 2016.

Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction

Ananto Setyo Wicaksono[1], Ahmad Afif Supianto[2]
Department of Informatics
Faculty of Computer Science, Brawijaya University
Malang, Indonesia

***Abstract*—Online news is a media for people to get new information. There are a lot of online news media out there and a many people will only read news that is interesting for them. This kind of news tends to be popular and will bring profit to the media owner. That's why, it is necessary to predict whether a news is popular or not by using the prediction methods. Machine learning is one of the popular prediction methods we can use. In order to make a higher accuracy of prediction, the best hyper parameter of machine learning methods need to be determined. Determining the hyper parameter can be time consuming if we use grid search method because grid search is a method which tries all possible combination of hyper parameter. This is a problem because we need a quicker time to make a prediction of online news popularity. Hence, genetic algorithm is proposed as the alternative solution because genetic algorithm can get optimal hypermeter with reasonable time. The result of implementation shows that genetic algorithm can get the hyper parameter with almost the same result with grid search with faster computational time. The reduction in computational time is as follows: Support Vector Machine is 425.06%, Random forest is 17%, Adaptive Boosting is 651.06%, and lastly K - Nearest Neighbour is 396.72%.**

***Keywords*—Hyper parameter; genetic algorithm; online news; popularity; machine learning**

I. Introduction

The news is information about what is happening in the world. This information is useful to people, for example as the topic of conversation, decision making, etc. News can be obtained by several ways, such as printed media, television, and online. Nowadays, online media is one of the most accessed media to obtain news [1].

Online news is a popular media to get information because a lot of people is using internet on their gadget [2]. The ability to comment and share the news is one of the appeals of the online news. Usually, people tend to share news they think interesting. The number of shares is one way to measure the popularity [3]. After reading the news, people don't read the same news contained in media news. This makes unshared online news not read by people [4]. If the news is not read by many people, an advertiser will not insert their advertisement and without advertisement, the online news media will lose its source of income. To solve it, it is necessary to know whether a news will be popular or not before its' publication [2].

Online news popularity prediction can be solved by one of the data mining techniques that are classification [5]. Classification is a way to discover the class of a certain input data using a certain rule. Support Vector Machine (SVM) [2], Random Forest (RF) [2] [6], Adaptive Boosting (AdaBoost) [2], K-Nearest Neighbour (KNN) [2], Naive Bayes (NB) [2] [7] was used to make a classification of online news popularity. The result was most of them can only get a result in 60% accuracy. To be able to get a more reliable result in online news popularity prediction, it is needed to produce a method that can make a result better than the previous research. This unsatisfactory prediction performance may occur because of the characteristic of the online news data.

In online news, there are a lot of features that can influence the amount of the popularity. That redundant feature usually makes the result of the prediction worse [8]. That is why feature reduction is one way to raise the result. Feature reduction is to use reduced feature instead of a full feature of the dataset. Feature reduction can be achieved by using feature selection and feature extraction. The two of them are methods to reduce the feature used, but the difference is, in feature selection, we select a number of features from the entire feature to be used in classification process, while in feature extraction, all of the features are converted into a new number of features that is less than the original feature. In [5], feature selection method is used to increase the result of the online news. Information gain, correlation-based, gain ratio, learner-based selection method was proposed and between them, the learner based achieved the best accuracy that is 63%. Even if it is able to get that result, when it comes to very high feature data, feature selection can have a lower performance [9]. For feature extraction, Principal Component Analysis (PCA) was used in [10]. After using PCA, the reduced feature then used as the feature of machine learning methods to do classification task for online news popularity prediction. The result is the SVM can achieve up to 69% accuracy. But, the weakness of using PCA is the feature became uninterpretable and it is necessary to define the optimal number of the principal component to get the optimum result [11]. These methods were methods to increase the performance of the machine learning methods from the attribute used.

Even without feature reduction, machine learning methods' performance can be increased, and one of the methods we can use is a hybrid algorithm. The hybrid algorithm is combining algorithm with another algorithm to get a better result [12]. In [13] hybrid SVM-RF was proposed to increase the performance of the online news popularity prediction. This implementation indeed gives a better performance by achieving up to 73% accuracy. While it can give a better performance, the downside of using hybrid algorithm is the difficulty in implementation and choosing the combination of the algorithm. If the combination of the algorithm is not right, it will give worse performance than the original algorithm.

Actually, there is a simpler method that can be used to enhance the performance of machine learning methods that is by tuning its hyper parameter. Hyper parameter is parameters inside machine learning methods. Each machine learning method has a different set of hyper parameter. In order to change this hyper parameter, we can just manually change the value until the result is satisfactory. However, manually tuning hyper parameter is a tedious task because there is a lot of possible combination that can be used. That's why automatic hyper parameter tuning is used to find the optimal value.

Grid Search can be used to automatically find the optimal hyper parameter. Grid Search is an optimization algorithm that searches all possible combination in the search space. In [2], Grid Search was used to optimize the hyper parameter of the machine learning methods. It can make the performance of machine learning algorithm enhanced and the best result out of all the machine learning method is RF with 67% accuracy. Grid Search also comes with a downside that is, the computational time is slow [14]. To overcome this problem, another optimization method can be used to replace this method. One of the methods is Genetic Algorithm.

Genetic Algorithm is an optimization algorithm that can get near the optimal value of a function [15]. The time needed to get the optimal value is also plausible. In [14], SVM with genetic algorithm hyper parameter optimization was used to try several free dataset and the result is SVM with a genetic algorithm can get the near optimal hyper parameter using a lot faster computational time.

Hence, a genetic algorithm is proposed in this research to replace the grid search method. This is the novelty of this research because genetic algorithm hasn’t been used to determine hyper parameter in online news data. The purpose of using genetic algorithm is to get a faster computational time for determining the popularity of online news because a lot of news are created every day and speed is necessary in online news media.

II. Material and Methods

Several steps are needed in order to make a popularity estimation of online news. The online news texts need to be translated into several attributes. Then, in order to enhance the result of the prediction, the hyper parameter is tuned using a genetic algorithm. After that, machine learning methods with optimal hyper parameter are used to classify whether online news is popular or not.

A. Data for Online News Popularity Prediction

The data used in this research was online news popularity dataset downloaded from UCI machine learning website[1]. The dataset consists of 39797 instances and 61 attributes. It consists of many online news articles from Mashable news service [2]. These articles are articles published from January 2013 until January 2015. This many attribute and instances make it a challenging dataset to use because it is a lot of data and the higher the amount of data, the longer the computational time needed. This dataset is the summary information of the news that is needed in order to know the popularity. For example, the number of positive words, negative words, the number of images, the number of shares, etc.

The target of this dataset is the number of shares. If the number of shares reaches a certain threshold, then the news is considered as popular, otherwise, it is not popular. The number of shares can be a way to determine popularity because people won't share an article they don't like to other people. The number of shares is used as popularity measurement. In this research, the minimum number of share of news considered as popular is 1400 shares [2].

B. Machine Learning Methods

Machine learning is a method to make a computer to have the ability to learn. Statistical methods are often used to achieve it. There are a lot of things that can be learned by using machine learning methods, one of them is classification. Until now, several machine learning methods are already developed to get a better result on learning. Machine learning methods can be used to solve a classification problem, clustering problem, and forecasting problem.

1) Support Vector Machine (SVM): SVM is a machine learning method for regression and classification [14]. SVM as classifier will make a hyper plane that separates the data into several classes. SVM has C, gamma, and kernel as hyper parameter. SVM is a popular machine learning methods used in many classification problems. SVM can be used to classify many things like in [16] as tumour classifier. In [17] SVM is used to classify poetry. [18] Used SVM in bank direct marketing problem.

2) Random Forest (RF): Random Forest is a machine learning method using several decision trees. It can be used as regressor or classifier. Random forest initializes a number of the tree via randomization technique [6]. The hyper parameter of Random Forest is the number of trees, the depth of the tree, etc. If it is tuned, it can achieve higher result than using a default hyper parameter.

3) K-Nearest Neighbor (KNN): KNN is a machine learning method that uses neighbouring data to get the result. KNN can be used as regressor or classifier. As a classifier, KNN will assign new data to a class that is nearest to its K neighbour. K is the number of the neighbour.

[1] https://archive.ics.uci.edu/ml/datasets/online+news+popularity

In KNN, firstly, the distance of new data will be measured against all of the data in the dataset. There are various algorithms to measure the distance of the data, such as Euclidean distance and Manhattan distance. Then, after the distance of the new data and all of the data is known, several data with the shortest distance with the new data will determine what the label of the data is. The new data will be assigned to a label with n data with the shortest distance. The number of n is equal to the number of K.

The number of K will affect the result of KNN. KNN is a simple yet effective classification method used in many classification problems [16] [19].

4) Adaptive Boosting (AdaBoost): AdaBoost is a classification method that combines several weak classifier methods to achieve better result [2]. The weak classifier will be trained into the dataset, and then the weight is assigned to them until all classifier has optimal weight and best prediction result. AdaBoost uses only a significant feature in the dataset as training. It makes AdaBoost's accuracy result have a higher accuracy Adaptive boosting tend to have an over fitting problem and sensitive to noisy data and outliers.

5) Grid Search: Grid Search is a widely used method for hyper parameter optimization [16]. Grid Search is a hyper parameter optimization method that uses brute force in order to find the best hyper parameter. It is more guaranteed to find the optimal hyper parameter because Grid Search will try all possible combination within a set of parameters. [20] and [21] is the usage of grid search for hyper parameter optimization.

6) Genetic algorithm: The genetic algorithm is optimization algorithm inspired by the process of evolution [17]. A chromosome is the representation of solution in the genetic algorithm. Then, it uses crossover and mutation to generate the new solution. Crossover is a mechanism of combining two chromosomes into one chromosome. The mutation is a process to get a new solution by changing one chromosome. The solution will then be evaluated with the objective function and the solution that didn't fit the criteria will be dropped. The solution kept will continue the process of crossover and mutation and evaluation until the stop condition is met. The stop condition can be a number of iteration and a certain time limit.

The genetic algorithm is a versatile algorithm which can be used in several problems. In [22] the authors introduce an approach to multilingual single-document extractive summarization where summarization is considered as an optimization or a search problem which is solved by using genetic algorithms. In [23] genetic algorithm has been applied for effective personalize web search-based on clustered query sessions. The genetic algorithm also used for term-weighting learning in term of text classification [24].

The genetic algorithm can be used as a method to find an optimal hyperparameter of machine learning methods. For example in [25], Genetic Algorithm is used as hyperparameter tuning of a fuzzy rule to classify facial expression. The purpose is to make the fuzzy method to get a better classification result.

C. Hyperparameter Optimization Methods

Hyper parameter is a parameter that is necessary for machine learning methods to make a classification. Each machine learning methods have different hyper parameters. Choosing the right parameters can make a significant difference in prediction results. That's why it is important to make a tuning in hyper parameter instead using the default parameters of machine learning. Determining these hyper parameters can be done manually by trying all of the possible value. But, doing that is time-consuming because the number of possible combination is very large. That's why optimization algorithm to automatically find the optimal hyper parameter such as Grid Search is often used.

D. Proposed Method

The implementation of this research is shown in Fig. 1. Firstly the dataset needs to be downloaded from UCI Machine Learning website. The format file of the dataset is Comma Separated Values (CSV). This dataset don’t have any missing value and all of the data is already in number, so a preprocessing method to fill the missing value and data conversion is not necessary. Almost all of the attributes will be used in the experiment which is 58 attributes. These attributes are the attributes that influence the popularity of the online news.

This dataset needs to be labelled first in order to do a classification because this dataset didn't have a label determining whether it is popular or not. The data in the dataset that can be used as popularity measurement is the number of the shares. If the number of shares reaches 1400 or more, then it is considered as popular, otherwise, it is not popular. This label is important because, in the methods that will be used, the data need to be a categorical data or in this research a binary data because there are only two labels for the data. The new data will be classified into these labels.

After the dataset was labeled, the next step is to split the dataset into two parts. The first part is the training dataset and the second part is testing dataset. The splitting ratio used in this research is 70% for training and 30% for testing. This was done using the Scikit Learn library in python [26]. Scikit Learn is a collection of library specialized at handling machine learning problem. From here on, until the evaluation, Scikit Learn library was used in the implementation.

Hyperparameter optimization comes after that. It is to determine the value of hyperparameter in machine learning automatically. Here, the genetic algorithm was proposed to replace the grid search method. The genetic algorithm will be used before the classification process to make the classification result of machine learning better. It is necessary to define several things before we can use a genetic algorithm, such as the chromosome, crossover rate, mutation rate, the number of iteration, etc. The chromosome of this algorithm is the hyperparameter of the machine learning. Each of the machine learning has different hyperparameter to optimize. In SVM, the chromosome of genetic algorithm will be Gamma, C, and kernel, in AdaBoost, they are number of estimator and learning rate, then in Random Forest, the parameters are number of decision tree used, minimum sample leaf, and minimum weight fraction of leaf , and finally the number of K is the hyperparameter to optimize in KNN method. The crossover rate used is 0.5 and the mutation rate is 0.1. The tournament will be used as evaluation method to determine the next generation of the population. Lastly, the number of iteration used is 10 iteration. To implement this method, we use evolutionary algorithm search CV in Scikit-learn with population size 50 and generation number 1000.

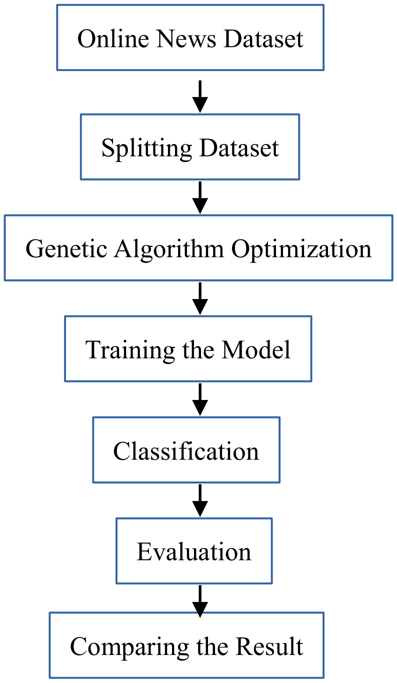


Fig 1. Diagram of the Proposed Method.

The next part is training the machine learning methods. The result of the genetic algorithm is an optimized hyper parameter of machine learning. The machine learning method using optimized hyper parameter will be trained using training data. We use Scikit-learn library for this. After that, the classification will be done using testing data in order to predict the popularity of online news. The result will be evaluated using several methods. They are accuracy, and the time needed to find the hyper parameter. The accuracy is obtained by using accuracy score library in Scikit-learn. The most important evaluation is the time because the proposed method is alternative hyper parameter tuning algorithm that has faster computational time.

III. Result and Discussion

The experiment in this research is comparing the accuracy and computation time of online news popularity prediction by using grid search and genetic algorithm. The result of the experiment is in Table 1 for implementation using Grid Search and Table 2 is Genetic Algorithm implementation. The time in this experiment is in second. Fig. 2 is the chart of the evaluation of prediction using optimized hyper parameter. The evaluation method that is used in this research is the accuracy of machine learning methods using optimized hyper parameter. The methods used these experiments are SVM, RF, AdaBoost, and KNN. Fig. 3 is the chart of the computational time needed for Grid Search and Genetic Algorithm to solve it. The time measurement is in second.

From the result of the experiment, the accuracy, of Grid Search and Genetic Algorithm is almost the same. Genetic Algorithm generates several solutions in one iteration. To get a new solution, Genetic Algorithm uses crossover and mutation method. Crossover is a combination of two solutions while mutation is a modification of one solution. This way of getting a new solution is derived from a process of evolution. The child is usually better than the parent, that's why it can be assumed that a combination or modification of solution can make a better solution result.

When it comes to computational time, Genetic Algorithm can achieve a much better result. From Fig. 3, the difference of computational time can be clearly seen. This make the prediction of online news popularity become faster. In Support Vector Machine, the time improvement to obtain the optimal hyper parameter is 7481 seconds. In the Random forest, 112 seconds is the time improvement result. Adaptive Boosting has 6758 seconds improvement, and lastly, K - Nearest Neighbour is improved by 242 seconds. The entire machine learning methods has significant time improvement. This happened because genetic algorithm did not search all possible hyper parameter.

TABLE I. Hyper Parameter Grid Search Result

Methods	**Grid Search**	
	Accuracy	**Time**
SVM	0.53	9241
RF	0.67	771
AdaBoost	0.66	7796
KNN	0.58	303

TABLE II. Hyper Parameter Genetic Algorithm Result

Methods	**Genetic Algorithm**	
	Accuracy	**Time**
SVM	0.54	1760
RF	0.67	659
AdaBoost	0.66	1038
KNN	0.58	61

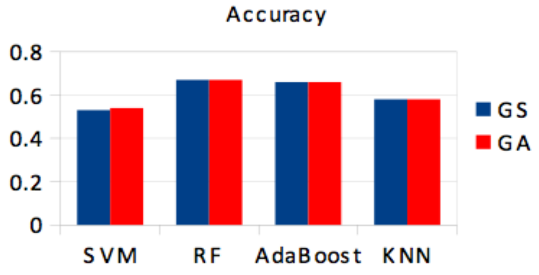


Fig 2. Accuracy of Prediction.

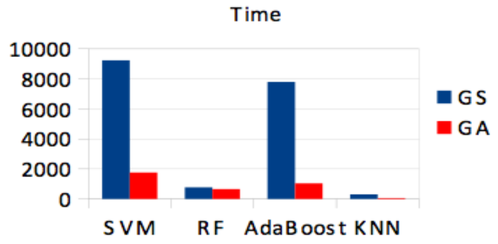


Fig 3. Computational Time.

Genetic Algorithm uses crossover and mutation to get a new solution in each iteration. In each iteration, Genetic Algorithm will get a better solution. The iteration in Genetic Algorithm will be executed until a stop condition is met, such as the execution time and the number of iteration. This method makes Genetic Algorithm can get better hyper parameter when the iteration ends without trying all possible combination to get the best result. This makes Genetic Algorithm can have a faster computational time. For the next work, we can try deep learning to make an online news popularity prediction and use genetic algorithm to get the hyper parameter of deep learning methods for a better accuracy.

IV. Conclusions

The rapid usage of the internet makes online news become a popular source to obtain information. It is important to measure the popularity of online news prior to its publication. To solve this problem, we can use machine learning methods such as SVM, Random Forest, etc. The accuracy of machine learning methods' classification can be increased by tuning its hyper parameter. In this research, a genetic algorithm is proposed as hyper parameter tuning. The experiment is implemented using Scikit Learn library and the data used is the online news dataset downloaded from UCI machine learning site. This dataset has 39797 instances and 61 attributes.

Based on the experiment, it can be concluded that genetic algorithm can produce an optimal hyper parameter for machine learning with a reasonable amount of time. This happen because genetic algorithm can search for a better solution without trying all possible solution. It makes Genetic Algorithm a better replacement for Grid Search when the dataset that needs to be processed is very large.

References

[1] L. H. Cheeks and A. Gaffar, “A Social Influence Model for Exploring Double Subjectivity through News Frames in Online News,” p. 11, 2017.

[2] K. Fernandes, P. Vinagre, and P. Cortez, “A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News,” in Progress in Artificial Intelligence, vol. 9273, F. Pereira, P. Machado, E. Costa, and A. Cardoso, Eds. Cham: Springer International Publishing, 2015, pp. 535–546.

[3] Q. Hu, G. Wang, and P. S. Yu, “Public Information Sharing Behaviors Analysis over Different Social Media,” 2015, pp. 62–69.

[4] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, “A survey on predicting the popularity of web content,” Journal of Internet Services and Applications, vol. 5, no. 1, Dec. 2014.

[5] W. Alswiti and A. Rodan, “Features Selection Effect on Predicting the Popularity of Online News,” Proceedings of the New Trends in Information Technology (NTIT-2017), p. 5, Apr. 2017.

[6] R. Shreyas, D. Akshata, B. Mahanand, B. Shagun, and C. Abhishek, "Predicting popularity of online articles using Random Forest regression," 2016, pp. 1–5.

[7] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. Nawi, “Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses.,” International Journal on Advanced Science, Engineering and Information Technology, vol. 7, no. 4, p. 1419, Aug. 2017.

[8] M. Bahrololum, E. Salahi, and M. Khaleghi, “Machine Learning Techniques for Feature Reduction in Intrusion Detection Systems: A Comparison,” 2009, pp. 1091–1095.

[9] S. Srivastava, N. Joshi, and M. Gaur, “A Review Paper on Feature Selection Methodologies and Their Applications,” p. 5.

[10] H. Ren and Q. Yang, “Predicting and Evaluating the Popularity of Online News,” p. 5.

[11] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” 2014, pp. 372–378.

[12] F. Mar’i and A. A. Supianto, "Clustering Credit Card Holder Based on Billing Payment using Improved K-Means with Particle Swarm Optimization," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 5, no. 6, pp.737-744, Nov 2018.

[13] A. Kathal and M. Namdev, “Correlation Enhanced Machine Learning Approach based Online News Popularity Prediction,” p. 6.

[14] I. Syarif, A. Prugel-Bennett, and G. Wills, “SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance,” TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 14, no. 4, p. 1502, Dec. 2016.

[15] M. Z. Sarwani, A. Rahmi, and W. F. Mahmudy, “An Adaptive Genetic Algorithm for Cost Optimization of Multi-Stage Supply Chain,” vol. 9, no. 2, p. 6.

[16] W. Li, X. Xing, F. Liu, and Y. Zhang, “Application of Improved Grid Search Algorithm on SVM for Classification of Tumor Gene,” International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 11, pp. 181–188, Nov. 2014.

[17] Rang, “Poetry Classification Using Support Vector Machines,” Journal of Computer Science, vol. 8, no. 9, pp. 1441–1446, Sep. 2012.

[18] I. Oktanisa and A. A. Supianto, “A Comparison of Classification Techniques in Data Mining for Bank Direct Marketing,” Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 5, no. 5, pp. 567-576, Oct 2018.

[19] A.A. Soebroto and A.A. Supianto, “Development of Decision Support Systems for Selection of Bali Cow Superior Seed Using the K-Nearest Neighbor Method,” Journal of Environmental Engineering and Sustainable Technology, vol. 2, no. 1, pp.49-57, Jul 2015.

[20] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov, “SVM Parameter Tuning with Grid Search and Its Impact on Reduction of Model Over-fitting,” in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, vol. 9437, Y. Yao, Q. Hu, H. Yu, and J. W. Grzymala-Busse, Eds. Cham: Springer International Publishing, 2015, pp. 464–474.

[21] L. Lin, Z. Xiaolong, Z. Kai, and L. Jun, “Bilinear Grid Search Strategy Based Support Vector Machines Learning Method,” p. 8.

[22] M. Litvak, M. Last, and M. Friedman, “A New Approach to Improving Multilingual Summarization Using a Genetic Algorithm,” in: ACL, The Association for Computer Linguistics, pp. 927–936, 2010.

[23] S. Chawla, “Application of genetic algorithm and backpropagation neural network for effective personalize web search-based on clustered query sessions,” International Journal of Applied Evolutionary Computation (IJAEC), vol. 7, no. 1, pp. 33-49, 2016.

[24] H. J. Escalante, M. A. García-Limón, A. Morales-Reyes, N. Graff, M. Montes-y-Gómez,E. F. Morales, and J. Martínez-Carranza, “Term-weighting learning via genetic programming for text classification,” Knowledge-Based Systems, 83, pp.176-189, 2015.

[25] A. Jamshidnezhad, “A Classifier Model based on the Features Quantitative Analysis for Facial Expression Recognition,” p. 4.

[26] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Machine Learning in Python, p. 6.

Classification based on Clustering Model for Predicting Main Outcomes of Breast Cancer using Hyper-Parameters Optimization

Ahmed Attia Said[1]
Information Systems Dept.,
Faculty of Computers and Information,
Helwan University

Laila A.Abd-Elmegid[2]
Assistant Professor, Information Systems Dept.,
Faculty of Computers and Information,
Helwan University

Sherif Kholeif[3]
Associated Professor, Information Systems Dept.,
Faculty of Computers and Information,
Helwan University

Ayman Abdelsamie Gaber[4]
Associated Professor of Medical Oncology, National
Cancer Institute,
Cairo University

Abstract—**Breast cancer is a deadly disease in women. Predicting the breast cancer outcomes is very useful in determining the efficient treatment plan for the new breast cancer patients. Predicting the breast cancer outcomes (also called Prognosis) are done based on the previous patient's data, which show the patient's characteristics and how the doctors treated the patient. In this paper we propose a new efficient model for predicting the main outcomes; Survival Rate, Disease Free Survival, and Recurrence detection; of breast cancer. The proposed model utilizes two techniques to increase the accuracy of the predictive results. The first technique is applying the classification model on various data clusters rather than the full dataset. In such steps, the data is grouped in different clusters according to the similarity of the main characteristics, then the classification model is applied on these clusters. The second technique is using the Hyper-Parameters Optimization (also called Hyper-Parameters Tuning) to increase the accuracy of the classification model. In this step, the proposed model uses Hyper-Parameters Optimization to find a tuple of hyper-parameters that yields on the optimal model which minimizes a predefined loss function on given dataset. The experimental study shows in detail how utilizing such two techniques results in an efficient prediction model producing accurate results.**

***Keywords*—Breast cancer; Survival Rate (SR); Disease Free Survival (DFS); recurrence detection; egy; prediction; data mining; classification; clustering; hyper-parameters optimization**

I. Introduction

Breast cancer is the famous type of cancer which infects women; it is considered as one of the highest deadly diseases at these times [1]. Diagnosis of such disease at an early stage results in better opportunities for good outcomes. The main breast cancer prognostication types are: Five-year Survival rate, Disease-Free Survival and the Recurrence Detection [2]. Survival rate indicates how many patients who were diagnosed as breast cancer patients' have survived for a time, this time can be 5 , 7 , 10 years according to the used time interval. Survival rates help to give a good knowledge about how the treatment was successful in these cases. Recurrence of the breast cancer is a breast cancer's case in which the cancer is back in the same breast or the other breast or at chest after a time. Disease Free Survival (DFS) is a measure for the time of period. This period measures the time interval between starting treatment plan and the time in which the patient survives without any symptoms of that cancer. Disease-free survival measuring is a way to see how the treatment works. Also called DFS, relapse-free survival and RFS. Breast cancer in Egypt is represented in incidence rate (48.8/100,000) and mortality rate (19.2/100,000) [3]. According to WHO, cancer country profiles in 2014 was that 21.6% of women cancer deaths in Egypt happen because of the breast cancer.

Data mining has a big importance in healthcare field, especially in building detection models, diagnosis and prognosis data [4]. In this paper, we propose a classification model based on data clustering and Hyper-Parameters Optimization techniques to predict the main outcomes of breast cancer prognosis with the highest possible accuracy.

II. Related Work

Rohit J.Kate et al. [5] built a predictive model for stages survival rate with the following classifiers: decision tree, naïve Bayes and the logistic regression at the SEER data set. The work proved that predicting the stage survivability achieves high accuracy than predicting the full data set survivability.

Houriyeh Ehtemam et al. [6] made a comparison between 64 classifiers' techniques for breast cancer early diagnosis and prognoses. The research was done on 208 Iranian records and consists of 10 attributes collected between 2014 and 2015. Bayesian network achieved accuracy of 95.7% as most accurate classifier.

M. Mehdi Owrang O. [7] was predicting the breast cancer survival rate using the association rules and naive Bayes with the SEER dataset. Results show that the two techniques were very similar in most of the cases.

Hadi Lotfnezhad Afshar et al. [8] predicted the breast cancer survivability using the SEER dataset and three classifiers: SVM, Bayes net and CHAID. The results showed that the SVM classifier was achieved the highest accuracy with 97.7 % and the two other classifiers were 81.8% and 82.2% sequentially.

Woojae Kim et al. [9] built a recurrence prediction model using SVM classifer .It uses a data set of 733 records and 7 attributes. The study conducted a comparison between the SVM and ANN techniques. Support Vector Machine scored 84.58% accuracy better than the Artificial Neural Network.

Hamid Karim Khani Zand [10] predicts survivability of breast cancer by making a comparison between many classifiers on the SEER dataset. Results showed that C4.5 classifier was more accurate with 86.7% accuracy than the Artificial Neural Network 86.5% and Naïve Bayes 84.5%.

Mamour Gueye et al. [11] works on the inflammatory breast cancer patient's outcomes .cases were treated in Dakar. The mean time to recurrence is 11.2 months and was found in 45.5% of cases. Survival rate was 31.8%.The median overall survival was 13.3 months.

Lone Winther Lietzen et al. [12] this study observed that the 1-year survival increased from 90.9% to 94.4%. 5-years survival increased from 70.0% to 74.7%.

Omar Farouk et al. [13] studied Egyptian women <=35 breast cancer characteristics .Results say that the breast cancer has more biological aggressive behavior at advanced stages for the young women in Egypt. The 3-year OS was 88% and 5-year OS 68%.The disease free survival median was 61 months, the 3-year disease-free survival was 58% and 5-year disease-free survival was 50%.

S.Kharya [14] predicting breast cancer prognosis on the SEER data set using different classifiers. Decision tree was achieved the highest accuracy with 93.62%.

Si Chen et al. [15] in this study they introduced a new post-labelling algorithm, creates partitions on the data sample, next they analyze the clusters results. Unlabeled data with high labeling confidence are selected to label and added into the labeled training set. Results show that the average result of over all the experimental data sets is highest .The algorithm outperforms self-training on all the data sets.

M.I. López et al. [16] used cluster for classification process in a meta-classifier. Clustering process is executed using the training set then training the classifier to classify using the training set to classifying the unseen data in the test set. The results show that the EM clustering algorithm produce results similar to the best classification algorithms, especially when using only a group of selected attributes.

Ritu Yadav et al. [17] proposed the Nottingham Prognostic Index (NPI) that uses lymph node status, tumor size and histological grade to define three types of patients with different probabilities of dying from breast cancer; good, moderate , and poor prognosis groups. Increase in numerical value of NPI is related with poor prognosis.

Gordon C Wishart et al. [18] proposed the PREDICT model which is a prognostic model developed for patient's whom diagnosed in the early breast cancer based on United Kingdom cancer registry data. This model is predicting breast cancer survival after surgery for invasive breast cancer and includes mode of detection for the first time.

Mohammad R. Mohebian et al. [19] used the PSO (particle swarm optimization) algorithm and BDT (bagged decision tree) to achieve highly accuracy in breast cancer recurrence prediction model .Three classifiers (SVM, DT, and multilayer perceptron NN) were used for comparison. The results show that the HPBCR achieve the highest accuracy with 89.2% while other three classifiers achieved SVM 77.6 %, Decision tree 77.1% and MLP 75%.

Chi-Hyuck Jun et al. [20] used PSO to improve tree-based classification rules. The study used the CART classification algorithm through three stages – tree building, threshold optimization, and simplification of the rules.

To conclude, the previous related work for predicting the breast cancer outcomes suffers from the following problems; work with small data set as in [6], work with an un-updated data set as in [12], predicting partial breast cancer outcomes as in [9, 13], from our point of view working with local data set as SEER and Breast Cancer Wisconsin Data Set doesn't reflect the main characteristics of other nation as in [5, 7, 8, 10, 14].

In our Breast Cancer Outcome Accurate Predictor proposed model (BCOAP), we work with an updated patient's data between 2010 and 2012.Also work with a national data set.

III. Contribution

A. Materials and Methods

1) Dataset: The first branch of the Egyptian NCI "National Cancer Institute" for breast tumors is "NCI First Settlement Hospital". Our study automated for the first time the manual cases' files of this hospital and presents the required information into a new developed dataset for Egyptian patients. The proposed mining model of the study developed data set to predicting the main outcomes of breast cancer for Egyptian patients, table1 shown the selected data set attributes. The developed data set captures the important attributes related to the prognosis process. About more than 40 attributes are used to specify the main characteristics of patients selected from the manual records and about 30 attributes of the overall attributes are used in the prediction model.

The total number of processed cases is 1692 case. These cases are diagnosed as breast cancer patients at the period from 2010 to 2012. After preprocessing the data and excluding the missing data records, 1471 records have been selected form the whole sample to be used in the experimental study. The selection criteria of the cases are:

TABLE I. DATA SET ATTRIBUTES

#	**Attributes**
1	Year of Diagnosis
2	Date of Birth
3	Death/Last Follow Year
4	T (TNM)
5	N (TNM)
6	M (TNM)
7	Tumor Size (Category)
8	Tumor Size (Cm)
9	Stage
10	Stage Subtype
11	Histological Type
12	Histological Grade
13	LN Status
14	No. of Positive Nodes
15	ER Status
16	PR Status
17	HER-2 Status
18	HER-2 Score
19	Surgery
20	Surgery Location
21	Surgery Year
22	Therapy 1
23	Therapy 2
24	Therapy 3
25	Recurrence Type
26	Recurrence Year
27	Metastatic Location
28	Status
29	Disease Free Survival
30	Disease Free Survival – Year

a) Being a female patient.

b) The case has been diagnosed since five years or more.

c) A complete data record of the main required data.

B. Proposed Model

In this study, we propose a model for predicting the main breast cancer outcomes, using the classification based on clustering and Hyper-Parameters Optimization to achieve the highest possible accuracy. The model is tested on data set of Egyptian patients developed through the study. The (**BCOAP)** model consists of **four** phases, Clustering phase, Features Selection phase, Classification phase, and Hyper-Parameters optimization phase. The following sections explain in details each of these phases.

1) Phase 1: Clustering*:* Clustering is a method of unsupervised learning, it is a ML (Machine Learning) technique that is grouping of data points. A set of data points are given, the clustering algorithms are used to classify each data point into a group. Each group should have similar properties and/or features. The data points in different groups should have dissimilar properties and/or features. Clustering methods can used to understand the relations between the data point in every data group.

The **BCOAP** model is used the TwoStep-AS Cluster algorithm which built in the IBM SPSS Modeler tool. TwoStep-AS Cluster is a tool designed to clustering a data set. This algorithm has several desirable features that differentiate it from traditional clustering techniques [a]:

- Handles two variables types categorical and continuous.
- Automatic selection of clusters number.
- **S**calability.
- TwoStep can analyze large data files.

2) Phase 2: Features selection: Features selection is an important ML technique which is used in classification models creates. It is used to decrease number of the features; it results in better classification performance. The feature selection's main purpose is to determining the important features from inputs to create the classification model with the highest accuracy. We can call a feature as a good one when it is relevant to the other features but not redundant. Feature selection algorithms care about limit the features to only features which would improve a task performance.

3) Phase 3: Classification: Decision tree is one of the most important and popular ML algorithms .Decision trees are supervised learning algorithms. It is very easy to understand. Decision trees are use the training data to build a predictive model which in a tree structure. The objective is to achieve high accuracy classification with low number of decisions. The decision tree consists of three nodes types: Decision nodes, Chance nodes and End nodes. Decision trees are drawn from top to bottom with its root at the top.

TABLE II. SELECTED FEATURES

#	**Attributes**
1	Age at Diagnosis
2	T (TNM)
3	N (TNM)
4	M (TNM)
5	No. of Positive Nodes
6	Surgery
7	Stage Subtype
8	ER Status
9	PR Status
10	HER-2 Score
11	Tumor Size (Cm)
12	LN Status
13	Histological Type
14	Histological Grade
15	Recurrence Type
16	Metastatic Location
17	Therapy 1
18	Therapy 2
19	Therapy 3
20	Status
21	Disease Free Survival

[a] IBM SPSS Modeler Documentation, https://www.ibm.com/eg-en/marketplace/spss-modeler.

4) Phase 4: Hyper-Parameters optimization: Hyper-Parameters Optimization (or tuning) is the solution of the problem of choosing a set of optimal hyper-parameters for a learning algorithm. Machine learning model may require different parameters to generalize different data patterns. The measures required are called hyper-parameters and must to be tuned to make the model optimally find a solution of the machine learning problem. Hyper-Parameters Optimization has many techniques to achieve the purpose of parameters optimization.

According to such feature, different algorithms can be applied at each of its phases which offer more flexibility. In our research we implemented the BCOAP model with the following algorithms; TwoStep-AS cluster in the clustering phase, Decision Jungle in the classification phase, and Hyper-Parameters Optimization at the final phase. The structure of the BCOAP model is presented in Fig. 1.

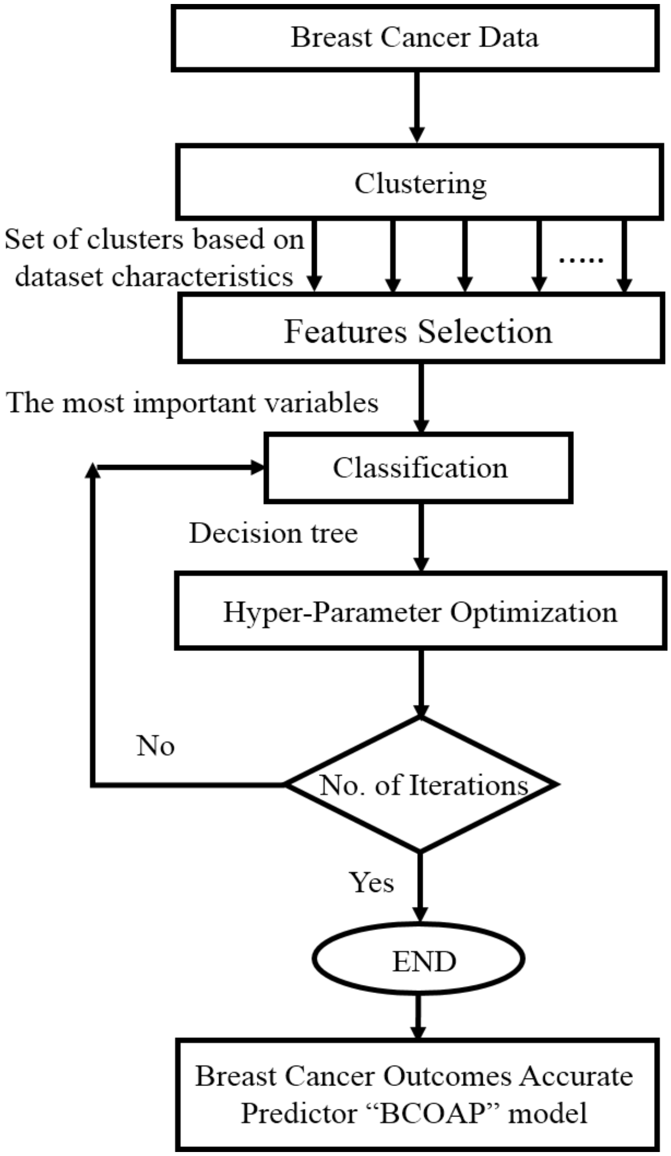

Breast Cancer Data
Clustering
Set of clusters based on dataset characteristics
.....
Features Selection
The most important variables
Classification
Decision tree
Hyper-Parameter Optimization
No
No. of Iterations
Yes
END
Breast Cancer Outcomes Accurate Predictor "BCOAP" model

Fig 1. The Proposed (BCOAP) Model.

IV. Experimental Study

The aim of this study is to prove the efficiency of the proposed model in predicting the breast cancer main outcomes. The first step as described in Section 3.2, we use the Two Step-AS clustering technique to assign each patient's file in a data group "cluster". Two Step-As algorithm recommended ten clusters as the optimal suitable number of clusters to the data set. Next, Features Selection phase is applied to determine the most important features in the selected dataset as an input to the classification phase. The list of selected features is shown in Table 2. For the prediction phase, Decision Jungle algorithm is used to create a machine learning model that is based on a supervised ensemble learning algorithm and train this model to predict the main outcomes of breast cancer in a decision tree form.

The first set of experiments predicts the breast cancer main outcomes at the level of the full dataset and at the level of clusters to highlight how clustering technique results in more accurate prediction.

The second set of experiments applies Hyper-Parameters Optimization on the classifier to test how it can improve the final prediction model. The model is configured to use a loop of 70 iterations to find the optimal classification model. The dataset we use is collected form the patient's files in the NCI First Settlement Hospital. IBM SPSS Modeler Subscription tool is used. Machine configurations are: Processor: Intel Core i3-3110M CPU @ 2.40GHz, Installed Memory (RAM): 4.00 GB and System Type: Windows Professional 64-bit.

V. Results

The results show the efficiency of the **proposed BCOAP** model in predicting the main outcomes of the breast cancer. The model achieved the highest prediction accuracy for the three main breast cancer outcomes; 5-Years survival rate (SR), breast cancer recurrence and disease free survival (DFS). The following section shows in details the results of the breast cancer main outcomes prediction.

A. *Predicting the Breast Cancer Main Outcomes using the Classification based on the Clustering Technique*

TABLE III. Predicting the Breast Cancer Outcomes

Data Set	**5-Years SR**	**Recurrence Detection**	**DFS**
Full Dataset	86.00	85.40	96.50
Cluster 1	90.00	96.40	80.60
Cluster 2	93.80	81.30	75.00
Cluster 3	91.40	87.70	64.90
Cluster 4	87.80	95.10	65.90
Cluster 5	88.70	97.90	85.10
Cluster 6	88.90	96.30	92.60
Cluster 7	71.10	97.70	91.10
Cluster 8	85.90	94.90	88.50
Cluster 9	63.20	94.70	94.70
Cluster 10	84.20	66.70	78.90

Table 3 shows the accuracy percentage when applying the classification model on the full dataset level and on the individual data clusters level. The results we got in the accuracy comparison between the full dataset and the data clusters showed that the accuracy was improved in almost of the dataset clusters for all the breast cancer outcomes we have predict in our BCOAP model. On other hand some clusters showed less accuracy than the full dataset. The reasons for this accuracy's decreasing are: some clusters has one label set in this records to predict, some records have missing values that has effect on the predicting process. In 5-years survival rate prediction, clusters from 1 to 6 achieved higher accuracy than the full dataset prediction is achieved, clusters from 7 to 10 achieved less than the full dataset achieved because of some missing data effect on the 5-years survival rate prediction. In the recurrence detection prediction, clusters 1 and from 3 to 9 have achieved higher accuracy than the full dataset achieved. Disease free survival prediction clusters have achieved no improvements in the accuracy predication after using the classification based on clustering technique because it has one value in the most of the data set records. This problem is found in the most of the classification algorithm.

B. *Predicting the Breast Cancer Main Outcomes using the Hyper-Parameters Optimization:*

In this section we present the results of the second set of experiments in which we apply the hyper-parameters optimization after using the classification based on clustering technique to predict the breast cancer main outcomes. The results we discuss in this section shows how is the using of hyper-parameters optimization is very helpful and very effective in achieving our contribution purpose in which we seek to improve the accuracy of predicting the breast cancer main outcomes.

1) Predicting the 5-years survival rate: As shown in Fig. 2, Cluster 1 has no change in the prediction accuracy after using the hyper-parameters optimization with 90%, clusters 2,6,10 accuracy's have decreased from (93.80%, 88.90% and 84.20%) before using the hyper-parameters optimization to (91.17% , 87.27% and 82.05%) after using it. Clusters 3,4,5,7,8,9 accuracy's have increased from (91.40% , 87.80% , 88.70%, 71.10% , 84.90% and 63.20%) to (93.10% , 87.95% , 90.84% , 79.12% , 86.62% and 79.48%) after using the hyper-parameters optimization.

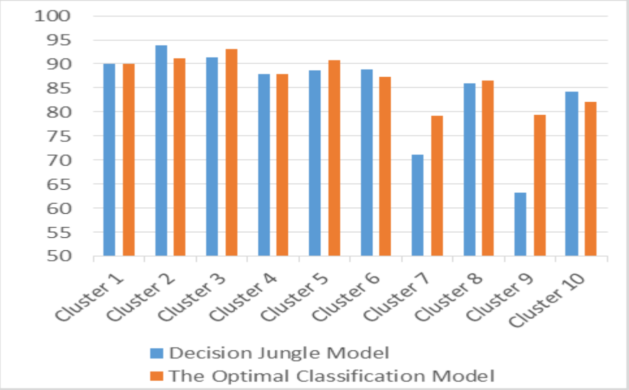


Fig 2. Predicting the Breast Cancer 5-Years Survival Rate.

2) Predicting the breast cancer recurrence detection: In predicting the breast cancer recurrence detection, we predict the recurrence of the breast cancer and also we predict the type of this recurrence as a local recurrence or metastatic recurrence. As results we note that the hyper-parameters optimization has a significant improving in the prediction accuracy. In Fig. 3, Cluster 8 has no change in the prediction accuracy after using the hyper-parameters optimization with 94.90%, clusters 1,4,5,6,7,9 accuracy's have decreased from (96.40% , 95.10% , 97.90% , 96.39% , 97.70% and 94.70%) before using the hyper-parameters optimization to (95.30% , 93.97% , 97.18% , 90.90% , 93.40% and 92.30%) after using it. Clusters 2,3,10 accuracy's have increased from (81.30%, 87.70%, and 66.70 %,) to (100 %, 99.13% and 94.87%) after using the hyper-parameters optimization.

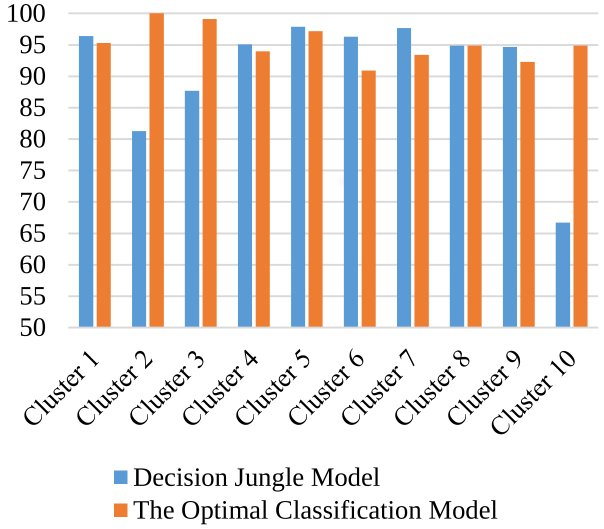


Fig 3. Predicting the Breast Cancer Recurrence Type.

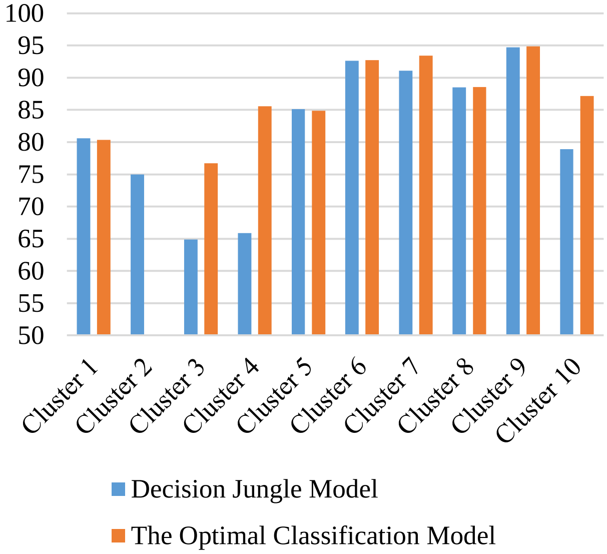


Fig 4. Predicting the Breast Cancer Disease Free Survival.

3) Predicting the Disease Free Survival (DFS): Fig. 4 shows that hyper-parameters optimization has a very good performance in the disease free survival prediction. Cluster 1, 5 accuracy's have decreased from (80.60% and 85.10%) before using the hyper-parameters optimization to (80.35% and 84.85%) after using it. Clusters 2,3,4,6,7,8,9,10 accuracy's have increased from (75% , 64.90% , 65.90% , 92.60% , 91.10% , 88.50% , 94.70% and 78.90%) to (85.29% , 76.72% , 85.54% , 92.72% , 93.40% , 88.53% , 94.87% and 87.17%) after using the hyper-parameters optimization.

VI. Conclusions

In this paper we introduced an optimal classification model for predicting the breast cancer main outcomes. In the proposed **BCOAP** model we used two techniques to increase the prediction accuracy. First technique is the classification based on clustering, in which we used the Two Step-AS clustering technique to group the data in clusters. After clustering process, the features selection is applied to select the most important features variables as an input to the third phase; the classification phase. The Decision Jungle algorithm is used as a classification model. The results of the experiments show that the classification of each single cluster is more accurate than the classification of the full dataset in the most of the clusters. In the fourth phase the hyper-parameters optimization is used to increase the accuracy by tuning the model parameters to find the optimal classification model. The experimental study proved the efficiency of the proposed **BCOAP** model and shows how it increases the accuracy of predicting the main outcomes of breast cancer significantly.

Future Work

Future work of this research includes the following main points:

- Utilizing different algorithms of clustering and decision trees to increase the efficiency of the proposed model.
- Considering the different therapy methods in the data set and the prediction outcomes.
- Extending the proposed model to highlight the outliers of the patient's characteristics.

References

[1] Belgian Cancer Registry "Cancer Incidence in Belgium 2010", Brussels, Feb 2013; 137 pages.

[2] American Cancer Society" Understanding a Breast Cancer Diagnosis", Cancer.org, 2016; 37 pages.

[3] Amal S. Ibrahim, Nabiel N. Mikhail , Hossam Darwesh, Tarek Heikel ,"Egypt National Cancer Registry Damietta Profile – 2009" , National Cancer Registry Program of Egypt, Jul 2010 ; 69 pages.

[4] Anupama Y.K, Amutha .S, Ramesh Babu.D.R, "Survey on Data Mining Techniques for Diagnosis and Prognosis of Breast Cancer", International Journal on Recent and Innovation Trends in Computing and Communication, Vol.5 Issue: 2, Feb 2017; pp. 33-37.

[5] Rohit J.Katea, Ramya Nadigb "Stage-Specific Predictive Models for Breast Cancer Survivability", International Journal of Medical Informatics, Vol. 97, Jan 2017; pp. 304-311.

[6] Houriyeh Ehtemam, Mitra Montazeri, Reza Khajouei , Raziyeh Hosseini, Ali NEMATI, Vahid Maazed "Prognosis and Early Diagnosis of Ductal and Lobular Type in Breast Cancer Patient" ,Iran J Public Health, Vol. 46, No.11, Nov 2017; pp. 1563-1571.

[7] M. Mehdi Owrang O. "Application of Data Mining Techniques for Breast Cancer Prognosis", IGI Global, 2015, pp. 1654-1665.

[8] Hadi Lotfnezhad Afshar, Maryam Ahmadi, Masoud Roudbari, Farahnaz Sadoughi "Prediction of Breast Cancer Survival through Knowledge Discovery in Databases", Global Journal of Health Science; Vol. 7, No. 4, 2015; pp. 392-398.

[9] Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, ParkMY, Park RW "Development of novel breast cancer recurrence prediction model using support vector machine", J Breast Cancer, Jun 2012; pp. 230-238.

[10] Hamid Karim Khani Zand "A Comparative Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction", Indian Journal of Fundamental and Applied Life Sciences 2015 Vol. 5, 2015; pp. 4330-4339.

[11] Mamour Gueye, Serigne Modou Kane-Gueye, Mame Diarra Ndiaye-Gueye, "Inflammatory breast cancer: features and outcomes in a breast unit in Dakar, Senegal" International Journal of Reproduction, Contraception, Obstetrics and Gynecology, Feb 2016; pp. 361-366.

[12] Lone Winther Lietzen, Gitte Vrelits Sørensen, Anne Gulbech Ordingd, "Survival of women with breast cancer in central and northern Denmark, 1998–2009", Dovepress, Clinical Epidemiology. 2011; pp. 35-40.

[13] Omar Farouk, Mohamed, AEbrahim, Ahmad Senbel, "Breast cancer characteristics in very young Egyptian women <=35 years", Dovepress, 2016; pp. 53-58.

[14] S.Kharya," Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, No.2, Apr 2012; pp. 55-66.

[15] Si Chen, Gongde Guo, Lifei Chen,"Semi-supervised Classification Based on Clustering Ensembles", International Conference on Artificial Intelligence and Computational Intelligence AICI 2009: Artificial Intelligence and Computational Intelligence, 2009; pp. 629-638.

[16] M.I. López, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums", Proceedings of the 5th International Conference on Educational Data Mining, Jan 2012; pp. 148-151.

[17] Ritu Yadav, Rajeev Sen, Preeti Chauhan, "Calculation of NPI Score: Prognosis of Breast Cancer" Indian Journal of Public Health Research & Development, Vol. 6 Issue 2,Springer, Apr-Jun 2015; pp. 199-202.

[18] Gordon C Wishart, Elizabeth M Azzato, David C Greenberg, Jem Rashbass, Olive Kearins, Gill Lawrence, Carlos Caldas, Paul DP Pharoah, "PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer", Breast Cancer Research, Jan 2010; 10 pages.

[19] Mohammad R. Mohebian, Hamid R. Marateb, Marjan Mansourian, Miguel Angel Mañanas, Fariborz Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning", Computational and Structural Biotechnology Journal ,2017; pp. 75–85.

[20] Chi-Hyuck Jun, Yun-Ju Cho, Hyeseon Lee "Improving Tree-Based Classification Rules Using a Particle Swarm Optimization", IFIP International Conference on Advances in Production Management Systems APMS 2012: Advances in Production Management Systems. Competitive Manufacturing for Innovative Products and Services, Springer, 2012; pp. 9-16.

Morphological Features Analysis for Erythrocyte Classification in IDA and Thalassemia

Izyani Ahmad[1], Siti Norul Huda Sheikh Abdullah[2]
Centre for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia Bangi, Malaysia

Raja Zahratul Azma Raja Sabudin[3]
Pathology Department
Hospital Canselori Tuanku
Mukhriz Cheras, Malaysia

***Abstract*—Iron Deficiency Anemia (IDA) and Thalassemia is a common disease in the world population. In hospital routine, those diseases are being recognized based on level of hemoglobin in Complete Blood Count (CBC) result. Then, visual experts will conduct examination under the light microscope which is subjected to human error. In this research, we suggested a methodology via machine learning to classify and characterize erythrocyte related with IDA and Thalassemia. We employ some image pre-processing techniques on the blood smear images to enhance edges and reduce image noise such as gamma correction and morphological processing. Then, every single erythrocyte image will segment the background and foreground by using Otsu's threshold method. Here, we have considered nine types of erythrocyte such as teardrop, echinocyte, elliptocyte, microcytic, hypochromic, target cell, acanthocyte, sickle cell and normal cell to be classified and portray based on their morphological features. Later, these 24 and 31 features from Hue's moment, Zernike moment, Fourier descriptor and geometrical features are confirmed as potential features for each condition by calculating one-way ANOVA. Next, the rank of subset features is done based on their information gain value from maximum to minimum. Each of subset is separated by incremental of five features. Here, we compare the performance for each subset with five selected classifiers namely logistic regression, radial basis function network, multilayer perceptron, Naïve Bayes Classifier and Classification and Regression Tree. The best subsets from 31 features provide the highest result of classification with 83.5% accuracy, 83.5% sensitivity and 83.3% positive predictive value respectively via logistic regression compared to other classifiers. This study could be extended by using image dataset from other blood based disease for future work.**

Keywords—(Iron Deficiency Anemia) IDA; Thalassemia; erythrocyte; morphological features; classifier; information gain; logistic regression

I. Introduction

Red blood cells (RBC) also termed as erythrocyte, forms the main part of human body system. They deliver oxygen to the different body tissues via blood flow through the circulatory system. It also contains haemoglobin which provide red colour and protein lipids to maintain their stability and deformability.

In hospital routine, pathologist detects erythrocyte abnormality in blood smear images under the light microscope. This process is very subjective evaluation that lead to tedious, time consuming and error prone work [1], [2] The pathologist accomplishes judgement based on his clinicopathological understanding in disease diagnostics. Thus, expose it to a higher variability of intra-observer due to small disparity in morphological feature.

Amongst various microscopic of blood smear test, most common disease detection including anaemia, thalassemia, leukaemia, red cell haemolysis, etc. In specific, anaemia and thalassemia frequently happens in our population. According to [3], approximately 1 in 10 Malaysian is a carrier for Thalassemia. Currently, there are 6753 Thalassemia patients that require medical attention.

The detection of certain type of anaemia and thalassemia can be based on microscope evaluation of erythrocyte in terms of their shape and size. So, the classification of abnormal and normal erythrocytes from microscopic images of blood smears has a large influence in developing a rapid and efficient tool. Hence, feature descriptors play an important role in classification. In this study, we are focusing on nine types of erythrocytes (teardrop, target, elliptocyte, hypochromic, microcytic, normocytic, burr and keratocyte) shapes that are commonly seen in iron deficiency anaemia (IDA) and thalassemia patient (Fig 1).

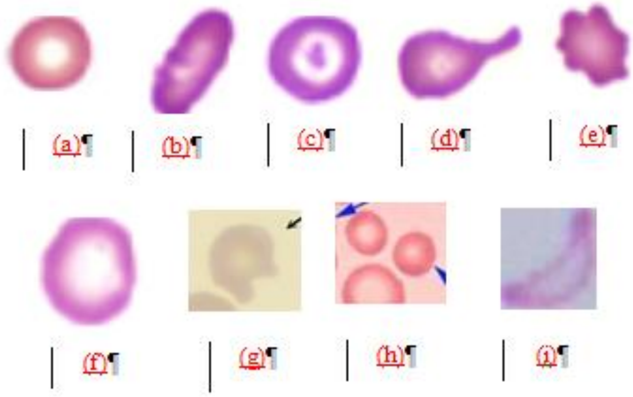


Fig. 1. Erythrocyte Cell Seen in IDA and Thalassemia Patient (a) Normal (b) Elliptocyte (c) Target (d) Teardrop (e) Burr (f) Hypocromic (g) Keratocyte (h) Microcyte and (i) Sickle.

A research reported [4] an approach to classify between malaria, thalassemia and normal patient using backpropagation neural network, while [5] proposed 271 features of color, texture and geometrical with PCA adopted. Author in [6] and [7] presented a rule-based technique for four classes of erythrocyte such as elliptocyte, macrocyte, microcytic, sphrerocyte and thalassemia beta major. Other author proposed cancer cell classification using geometric mean transform and dissimilarity metrics [8]. While some others, proposed shape geometric features into 8 features [9], 4 features [6] and 6 features [7] including of area, perimeter, diameter, shape geometric feature, area proportion, deviation, central pallor, form factor, shape area factor and diameter area factor. Meanwhile, other author proposed moment invariant [10] and geometric features in their classification [11], [12]. Recently, [13] proposed a new shape descriptors using generalized support functions to describe convex figures. In conclusion, recent cell detection and recognition research are still focusing on morphological features to boost up the performance of overall digital pathology image analysis.

Our objective is to classify nine different categories of abnormal and normal erythrocyte in IDA and Thalassemia using a set of 24 and 31 features (Hue's moment, Zernike moment, Fourier descriptor and geometrical features). The gold standard analysis of the images is performed by two experts. Then, evaluation and statistical analysis are conducted for both set of features and compared. The rest of this paper is arranged as follows: the proposed morphological feature analysis in IDA and Thalassemia blood smear images in Section II, results and discussion in Section III, and the conclusion of this work in Section IV.

II. Material and Methods

A. Data Collection and Image Acquisitions

In this study, we have considered about seven healthy, nine Iron Deficiency Anaemia and seven Thalassemia blood smear images. These patients' peripheral blood smear were captured under light microscope from Pathology Department, Hospital Canselori Tuanku Muhriz (HCTM), Cheras, Malaysia. The process for displaying RBC image involved digitization of image from optical image with 40 times (40X) objective which equals to approximately 400 magnifications. Image acquisition has been guided by Hematologist (Fig. 2) in HCTM. From this study, the effective resolution and pixel size were 150dpi and 1920 X 1440 respectively. These images were captured using Olympus Digital Camera model DP22.

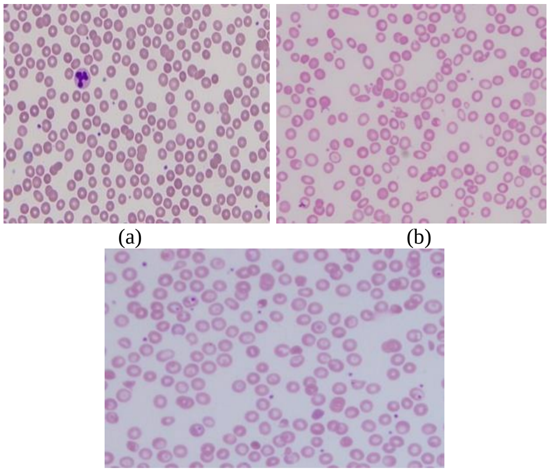


Fig. 2. Image Acquisition of (a) Normal (b) IDA (c) Thalassemia Source Real Blood Smear Dataset from Hospital Canselori Tuanku Muhriz (HCTM).

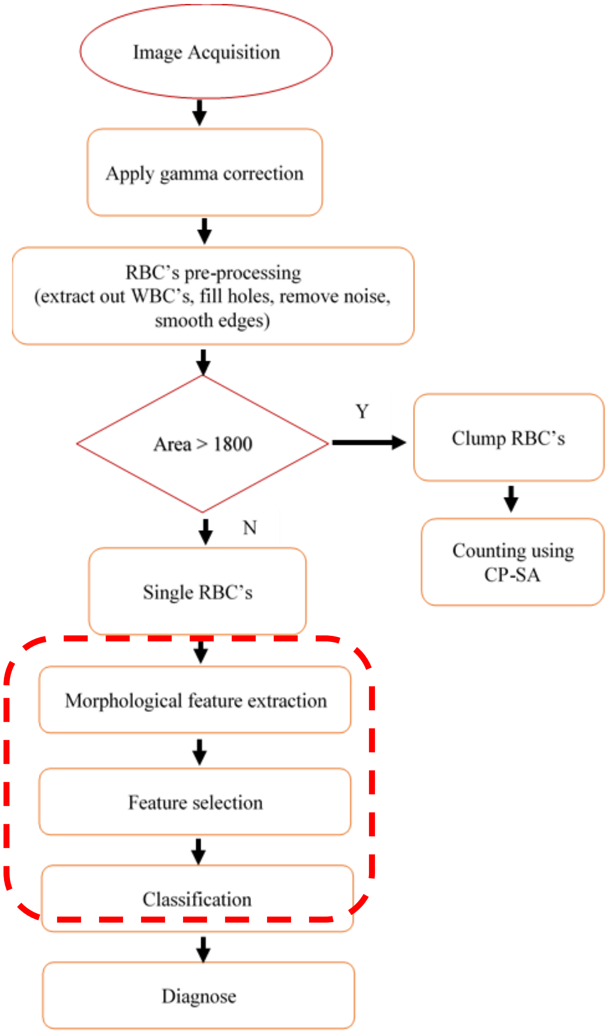


Fig. 3. General Methodology for the Proposed Method.

B. Erythrocyte Segmentation

Based on Fig. 3, we apply gamma correction to get sharpen edges. Followed by a few pre-processing step for erythrocyte segmentation. Hence, erythrocyte segmentation is vital for distinguishing abnormality present in IDA and Thalassemia disease. We have used *connected component labelling* method to segment erythrocyte from peripheral blood smear images. Later, each segmented erythrocyte was separated to clump and single cell based on their area. Here, clump cells will be counted by CP-SA method. CP-SA used skeleton algorithm in getting the backbone of each clump RBC, and analysed it with the concavity point pixel to detect and count the RBC. While single cell will be input to next level.

C. Morphological Feature Extraction

In IDA and Thalassemia condition, erythrocyte morphology varies from the normal cell. Thus, mathematical model was chosen to characterizing abnormal RBC's. Here, two separated experiment was conducted. First experiment consists of 24 morphological features in Table I, while another experiment includes 31 morphological features as shown in Table II.

TABLE I. 24 EXTRACTED FEATURES FOR IDA AND THALASSEMIA CATEGORIZATION

NO. OF FEATURES	FEATURES
7	HUE'S MOMENT
1	ZERNIKE MOMENT
4	FOURIER DESCRIPTOR
12	PERIMETER, AREA EQUIVALENT DIAMETER(AED), ECCENTRICITY, COMPACTNESS, MAJOR AXIS, MINOR AXIS, SOLIDITY, PERIMETER EQUIVALENT DIAMETER (PED), AREA, ROUNDNESS, CONVEX AREA, CONCAVITY

TABLE II. 31 EXTRACTED FEATURES FOR IDA AND THALASSEMIA CATEGORIZATION

NO. OF FEATURES	FEATURES
7	HUE'S MOMENT
1	ZERNIKE MOMENT
4	FOURIER DESCRIPTOR
19	PERIMETER, AREA EQUIVALENT DIAMETER(AED), ECCENTRICITY, COMPACTNESS, MAJOR AXIS, MINOR AXIS, SOLIDITY, PERIMETER EQUIVALENT DIAMETER (PED), AREA, ROUNDNESS, CONVEX AREA, CONCAVITY, FILL AREA, CENTRAL NODE, CENTRAL PALLOR, AREA PROPOTION (AP), AP EQUIVALENT CIRCLE (APEC), AREA MINOR , RADIAN

a) Hue's moments

This moment was introduced by Hu (1962). It manipulates the moment invariant within translation, scaling and rotational. This method calculate central moments, and the next seven invariant moments were created based on it [12].

b) Zernike moments

Zernike moments are based on a unit circle of a complete orthogonal set of complex polynomial. It is better than other methods in terms of information redundancy, reconstruction capability and noise resilience.

In this method, region of interest is plotted using polar coordinate in a unit circle. The origin of circle is also the center of the region of interest. Here, ϴ is the angle of polar coordinate and *r* is the radius of polar coordinate. The plotting to the polar coordinate is:

$$x = r\cos\emptyset \; and \; y = r\sin\emptyset \qquad (1)$$

where

$$r = \sqrt{x^2+y^2} \text{ and } \emptyset = tan^{-1}\left(\frac{y}{x}\right)$$

Next, Zernike moments A(x,y), was calculated by normalized the region of interest using polar coordinates. Equation as following is considered for achieving both properties.

$$A(x,y) = f\left(\frac{x}{a}+\bar{x}, \frac{y}{a}+\bar{y}\right) \qquad (2)$$

Where

$$a = \sqrt{\frac{\beta}{m_{00}}},\; \bar{x} = \frac{m_{10}}{m_{00}} \; and \; \bar{y} = \frac{m_{01}}{m_{00}}$$

Note that, $m_{00,}m_{10,}m_{01}$ are taken from Hue's moments. In binary image, $m_{00} = \beta$.

c) Fourier descriptor

A region of interest image consists of M-point pixel boundary [14]. The boundary points coordinate is $(x_0, y_0), (x_1, y_1), \ldots\ldots.(x_{k-1}, y_{k-1})$. These coordinates can be stated as x(k) = x_k and y(k) = y_k.. It can be defined in complex number as Eq. (3).

$$s\,(k) = x_k + jy_k \qquad (3)$$

where $k = 0, 1, 2,3 \ldots M-1$. In complex number plane, real axis is stated as x-axis and imaginary axis as y-axis

The Fourier descriptor of $s(k)$ is defined in Eq. (4) as below,

$$a(u) = \frac{1}{M}\sum_{k=0}^{M-1} s\,(k)e^{-j2\pi uk/M} \qquad (4)$$

For $k = 0, 1, 2,3 \ldots M-1$.

Fourier descriptor of the boundary is defined as *a(u).* The four features of fd1, fd2, fp1 and fp2 were calculated based on this descriptor [15].

d) Other morphological features

Other morphological features consist of eccentricity, roundness, area, fill area, area proportion have also been considered in [16], [17]. Part of them are detailed as following

Fill Area: Specifies the number of pixels in the region, with all holes filled in.

Centre Node (CN): Specifies value of center point in binary (1 or 0).

Central Pallor (CP)

$$CP = Fill\ Area - Area \qquad (5)$$

Area Proportion (AP)

$$AP = \frac{CP}{Fill\ Area} \qquad (6)$$

(i) Area Proportion Equivalent Circle (APEC):

$$APEC = \frac{\text{Fill Area}}{\text{EC}} \qquad (7)$$

where, Equivalent Circle, $EC = \pi r^2$

(i) Area Minor (AM)

$$AM = \pi\left(\frac{MinorAxisLength}{2}\right)^2 \qquad (8)$$

where, Minor Axis Length = length of the minor axis of the ellipse.

D. Selection of Significant Feature for IDA and Thalassemia

The feature selection is playing an important role in machine learning process. It has impact to increase the potentiality of each classifier. Some are suitable to differentiate textural and morphological features for describing normal and abnormal erythrocytes. Though, for predicting this situation, a set of features needs to be identified, in which is given the highest prediction accuracy. Plus, reduced the computational complexity and non-selected features. Here, statistical process were employ to the extracted features using one-way ANOVA (analysis of variance) for pinpointing a set of significant features [18]. Next, the information gain value is computed for each significant features using Weka version 3.7.5 and ranked based on it [19]. One way ANOVA calculated and compared the means of two or more instance using F-test statistic [18] as denoted below

$$F = \frac{between\ class\ variance}{within\ class\ variance} > 0 \tag{9}$$

The lower F value indicates lower discrimination potentiality and bigger F value indicates the better discrimination potentiality. Table III displays the analysis of one-way ANOVA from the extracted morphological erythrocyte images for both 24 and 31 features. Each feature was modeled for nine types of erythrocyte in IDA and Thalassemia representation. From this table, we can observe that all 24 and 31 features are statistically significant.

TABLE III. ONE WAY ANOVA ANALYSIS OF IDA AND THALASSEMIA FEATURES

FEATURES	P VALUE	FEATURES	P VALUE
ROUNDNESS	< 0.0001	MAJOR	< 0.0001
APEC	< 0.0001	CP	< 0.0001
ECCENTRICITY	< 0.0001	HUE5	< 0.0001
MINOR	< 0.0001	HUE4	< 0.0001
AREA MINOR	< 0.0001	AP	< 0.0001
COMPACTNESS	< 0.0001	AREA	< 0.0001
SOLIDITY	< 0.0001	CONVEXAREA	< 0.0001
PED	< 0.0001	HUE3	< 0.0001
CN	< 0.0001	HUE1	< 0.0001
RADIAN	< 0.0001	ZERNIKE	< 0.0001
AED	< 0.0001	FD1	< 0.0001
HUE2	< 0.0001	FP2	< 0.0001
HUE6	< 0.0001	FP1	< 0.0001
FILL AREA	< 0.0001	FD2	< 0.0001
HUE7	< 0.0001	PERIMETER	< 0.0001
CONCAVITY	< 0.0001		

a) Features ranking

Here, process for ranking the significant features based on their information gain was done. This study compared the ranked features into two conditions: (i) for 24 features (Table IV) and (ii) for 31 features (Table V). These ranked features were allocated into five subsets (24 features) and seven subsets (31 features) based on their ranks.

b) Information gain ranking

Information gain is designed via entropy theory of Shannon's [19]. The top-ranked features have the highest value of information gain. All significant features were set from maximum to minimum value of their information gain.

TABLE IV. FEATURE RANKING BASED ON INFORMATION GAIN FOR 24 FEATURES

RANK	INFO GAIN VALUE	RANK FEATURES	RANK	INFO GAIN VALUE	RANK FEATURES
1	0.959	COMPACTNESS	13	0.42	HUE7
2	0.959	SOLIDITY	14	0.395	ZERNIKE
3	0.882	ECCENTRICITY	15	0.389	CONVEXAREA
4	0.875	ROUNDNESS	16	0.38	HUE5
5	0.823	CONCAVITY	17	0.35	HUE4
6	0.8	MINOR	18	0.347	HUE3
7	0.55	MAJOR	19	0.303	FD2
8	0.511	HUE2	20	0.268	FP1
9	0.511	PED	21	0.266	FP2
10	0.511	AED	22	0.252	AREA
11	0.479	HUE1	23	0.247	FD1
12	0.426	HUE6	24	0.236	PERIMETER

TABLE V. FEATURE RANKING BASED ON INFORMATION GAIN FOR 31 FEATURES

RANK	INFO GAIN VALUE	RANK FEATURES	RANK	INFO GAIN VALUE	RANK FEATURES
1	0.959	SOLID	17	0.446	FILLAREA
2	0.959	COMPACTNESS	18	0.426	HUE6
3	0.889	APEC	19	0.42	HUE7
4	0.882	ECCENTRICITY	20	0.395	ZERNIKE
5	0.875	ROUNDNESS	21	0.389	CONVEXAREA
6	0.823	CONCAVITY	22	0.38	HUE5
7	0.8	MINOR	23	0.35	HUE4
8	0.753	AREA MINOR	24	0.348	CN
9	0.647	CP	25	0.347	HUE3
10	0.55	MAJOR	26	0.303	FD2
11	0.55	RADIAN	27	0.268	FP1
12	0.526	AP	28	0.266	FP2
13	0.511	HUE2	29	0.252	AREA
14	0.511	AED	30	0.247	FD1
15	0.511	PED	31	0.236	PERIMETER
16	0.479	HUE1			

E. IDA and Thalassemia Cell Identification

Our main objective was to classify and compared nine different categories of abnormal and normal erythrocyte based on two conditions (24 and 31 features). Abnormal erythrocyte characterization in IDA and Thalassemia patients such as teardrop, target, elliptocyte, acanthocyte, hypochromic, echinocyte, microcytic, sickle cell and normal erythrocyte were identified. Here, we setup five classification approach namely radial basis function network, logistic regression, Naïve Bayes classifier, classification, regression trees and multilayer perceptron by using Weka version 3.7.5.

a) Logistic regression (LR)

Logistic regression is a methodology under supervised classification which determines the membership of each dataset. As our experiment involved nine classes of erythrocyte, multiclass logistic regression model with one against all algorithm was used [20].

b) Radial basis function network (RBF)

The radial basis function network is a sort of feed forward network [21]. It contents of input layer, output layer and single hidden layer. Let assume that x = $[x_i: i = 1,2,3, \ldots., d]$ *and d* dimensional input feature space, while target output, $t = [t_z = 1, \ldots., k]$. The output of the RBF is defined by Eq (6) as below,

$$y_k(x) = \sum_{j=1}^{m} w_{kj} \emptyset_j(x) + w_{k0} \quad (6)$$

Here, w_{k0}= the basis input, w_{kj}= weights of hidden node.

The basis function, $\emptyset_j(x)$ is explained in Eq. (7) as the following:

$$\emptyset_j(x) = exp\left(-\frac{\|x-\mu_j\|^2}{2\sigma^2{}_j}\right) \quad (7)$$

where μ_j is the centre of the radial basis function, x is input vector, σ_j is the width of $\emptyset_j$. In order to determine the basis function parameters, K-means clustering algorithm is applied. Let say, σ_j =0.1.

c) Multilayer perceptron (MLP)

This methods is one type of back propagation neural network algorithm with multiple layers [21]. This algorithm will update the weights of each input and output layer in order to minimize output error and give better accuracy. Eq. (8) is applied to calculate the training error,

$$J(w) = \frac{1}{2} \sum_{k=1}^{c} (t_k - z_k)^2 \quad (8)$$

where the network output z_k and target output is t_k. The number of node is depicted by C. Here, we have nine output nodes with single hidden layer.

d) Naïve bayes classifier (NB)

This classifier is rooted from Bayesian theoretic method for feature set independent assumption. This method will guess possession of specific feature set via posterior probability. Here, posterior probability is estimated and forecast for each class label based on maximum value [21].

e) Classification and regression tree (CART)

CART is commonly used in machine learning with different classification problem. This algorithm construct classification and regression tree via top down recursive and divide-and-conquer approach. The CART comprises of root node, internal node and leaf node. The bottom node is leaf node and top most node is root node. Class label is allocated by leaf node. Here, the splitting criterion for data partition is done by Gini index measure [19].

III. Result and Discussion

Based on both Table VI (31 features) and Table VII (24 features), it is perceived that sometimes, whole features might not contribute the highest prediction performance. Hence, we decided to diversify them into subsets via their information gain value. Each subset consists of incremental of five features separately that sum up to 7 subsets for 31 features and 5 subsets for 24 features. Then, all subset will go through the five selected classifiers. This research includes 725 abnormal and 99 normal erythrocytes.

As tabulated in Table VI and Table VII, it shows variation performance of different subsets of morphological features ranked using five different classifiers. As in Table VI, the Accuracy value varies from 68.2 to 78.3 for Naïve Bayes classifiers, 73.0 to 80.5 for RBF, 75.8 to 83.3 for MLP, 75.6 to 83.5 for LR and 71.2 to 78.5 for CART respectively. While in Table VII, this Accuracy value is much lower within 46.9 to 58.6 (Naïve Bayes), 50.4 to 63.7 (RBF), 54.3 to 68.0 (MLP), 54.6 to 72.1 (LR) and 52.6 to 63.7 (CART). It is also noticeable that top fifteen features (from 31 features) with (Sensitivity:83.5%, PPV:83.3% and Accuracy:83.5%) via logistic regression give better performance compared to all features (from 24 features) (Sensitivity:72.2%, PPV:72.2% and Accuracy:72.1%) using the same classifier.

We also observed a fluctuation trend in these tables with different subset with different classifier. This can be concluded that each of feature have their unique value that led to better or worst result. Lesser feature is worth to get better performance and accuracy, in which approximately subset of ten, fifteen and twenty are worth than 31 features ranked. However, more features need to get more accurate result based on 24 features ranked. Hence, it is notable that each feature will get more precise and optimized result with the introduction of information gain ranking.

TABLE VI. Performance Statistics of 31 Features for Erythrocyte Characterization (A) Sensitivity (B) PPV and (C) Accuracy

(A)

Set of ranked features	**Sensitivity**				
	NB	**RBF**	**MLP**	**LR**	**CART**
FIVE	68.2	73.0	75.9	75.5	71.2
TEN	71.8	76.5	75.7	81.3	73.2
FIFTEEN	75.9	75.3	79.9	83.5	75.1
TWENTY	78.1	80.5	77.5	82.3	77.7
TWENTY FIVE	78.3	80.3	81.1	80.3	78.5
THIRTY	77.1	76.5	83.3	78.3	78.3
THIRTY ONE	77.1	76.3	80.3	81.5	78.3

(B)

Set of ranked features	PPV				
	NB	RBF	MLP	LR	CART
FIVE	69.3	71.8	75.2	75.3	70.1
TEN	72.9	76.5	74.5	81.0	73.1
FIFTEEN	76.4	75.1	79.6	83.3	74.9
TWENTY	78.0	80.6	77.1	82.1	77.5
TWENTY FIVE	78.1	80.4	80.8	80.1	78.5
THIRTY	77.1	76.2	83.2	78.6	78.2
THIRTY ONE	77.2	76.2	80.0	81.5	78.2

(C)

Set of ranked features	Accuracy				
	NB	RBF	MLP	LR	CART
FIVE	68.2	73.0	76.0	75.6	71.2
TEN	71.8	76.5	75.8	81.3	73.2
FIFTEEN	76.0	75.4	79.9	83.5	75.2
TWENTY	78.1	80.5	77.5	82.3	77.7
TWENTY FIVE	78.3	80.3	81.1	80.3	78.5
THIRTY	77.1	76.5	83.3	78.7	78.3
THIRTY ONE	77.1	76.3	80.3	81.5	78.3

TABLE VII. PERFORMANCE STATISTICS OF 24 FEATURES FOR ERYTHROCYTE CHARACTERIZATION (A) SENSITIVITY (B) PPV AND (C) ACCURACY

(A)

Set of ranked features	Sensitivity				
	NB	RBF	MLP	LR	CART
FIVE	46.9	50.4	54.3	54.6	52.6
TEN	53.9	56.9	59.8	62.3	56.5
FIFTEEN	57.6	63.7	63.1	67.2	62.5
TWENTY	58.6	63.1	65.7	69.3	62.1
TWENTY FOUR	57.2	61.4	68.0	72.2	63.7

(B)

Set of ranked features	PPV				
	NB	RBF	MLP	LR	CART
FIVE	44.0	49.5	53.3	52.6	50.2
TEN	52.5	56.7	58.6	61.7	54.9
FIFTEEN	56.2	63.6	62.7	67.2	62.1
TWENTY	57.6	63.3	65.6	69.2	61.5
TWENTY FOUR	56.4	60.9	67.9	72.2	63.3

(C)

Set of ranked features	Accuracy				
	NB	RBF	MLP	LR	CART
FIVE	46.9	50.4	54.3	54.6	52.6
TEN	53.9	56.9	59.8	62.3	56.5
FIFTEEN	57.6	63.7	63.1	67.2	62.5
TWENTY	58.6	63.1	65.7	69.3	62.1
TWENTY FOUR	57.2	61.36	68.0	72.1	63.7

IV. CONCLUSION

As a conclusion, we have characterized and classified nine types of erythrocyte for identification of IDA and thalassemia condition. A set of morphological features were extracted from segmented blood smear images and ranked via information gain value into two conditions: (i) 24 features and (ii) 31 features. The experimental amongst the five classifiers has shown that logistic regression had given the optimal value of accuracy, sensitivity and positive predictive value (83.5%, 83.5% and 83.3%) respectively. Thus, we proposed these geometrical features to characterize each of the shape.

It also detected that a fewer feature in subset (for 31 feature) provides better performance in most classifier. This result proven that each of features has different significant value depend on the segmented image and it is important to ranked this feature correctly. This research could be extended by using image dataset from other blood based disease for future work.

ACKNOWLEDGMENT

This work obtained ethical approval KM1.5.3.5/244/FRGS/1//2014/ICT07/UKM/02/5/Dr.SitiNorul HudaSheikhAbdullah starting from 19 October 2014.

REFERENCES

[1] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Zaharatul Azma, and K. Omar, “Automatic Detection and Quantification of WBCs and RBCs Using Iterative Structured Circle Detection Algorithm.,” Comput. Math. Methods Med., vol. 2014, p. 979302, 2014.

[2] D. Albashish, A. Sahran, S., M. A., Alweshah, and A. A., “A hierarchical classifier for multiclass prostate histopathology image gleason grading.,” Pertanika J. Sci. Technol., vol. 2, no. 2, pp. 323–346, 2018.

[3] A. Rz et al., “A Comparative Study of Red Blood Cell Parameters of Alpha and Beta Thalassaemia Patients Diagnosed in a University Hospital in,” vol. 3, no. 1, pp. 23–27, 2018.

[4] Y. Hirimutugoda and G. Wijayarathna, “Image Analysis System for Detection of Red Cell Disorders Using Artificial Neural Networks,” Sri Lanka J. Bio-Medical Informatics, vol. 1, no. 1, pp. 35–42, 2010.

[5] J. A. Alkrimi, L. E. George, A. Suliman, A. R. Ahmad, and K. Al-jashamy, “Isolation and Classification of Red Blood Cells in Anemic Microscopic Images,” Int. J. Medical, Heal. Pharm. Biomed. Eng., vol. 8, no. 10, pp. 686–689, 2014.

[6] S. Chandrasiri and P. Samarasinghe, “Morphology Based Automatic Disease Analysis through Evaluation of Red Blood Cells,” 2014 5th Int. Conf. Intell. Syst. Model. Simul., pp. 318–323, 2014.

[7] E. Suryani and K. N. Wahyudiani, “Berdasarkan Morfologi Sel Darah Merah,” Fakt. thalasemia pada anak, vol. 2, no. 1, pp. 15–28, 2015.

[8] S. M. M. Kahaki, M. J. Nordin, W. Ismail, S. J. Zahra, and R. Hassan, “Blood cancer cell classification based on geometric mean transform and dissimilarity metrics,” Pertanika J. Sci. Technol., vol. 25, no. S6, pp. 223–234, 2017.

[9] V. Acharya and P. Kumar, “Identification and red blood cell classification using computer aided system to diagnose blood disorders,” 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017–Janua, pp. 2098–2104, 2017.

[10] S. J. Zahra, R. Sulaiman, and A. S. Prabuwono, “SCIENCE & TECHNOLOGY Invariant Feature Descriptor based on Harmonic Image Transform for Plant Leaf Retrieval,” vol. 25, pp. 107–114, 2017.

[11] N. F. Hassan and A. I. Hussein, “Efficient method to Recognition of Anemia Images based on Moment Invariants and Decision tree classifier روص ضرم رقف مدلا د لع ساسا موزعلا ةتذ باڈ لا ف ذصملاو رجش ةرق لا را ةق يرطةءوف ك زي يم تل,” vol. 57, no. 3, pp. 2360–2370, 2016.

[12] D. Das, M. Ghosh, C. Chakraborty, M. Pal, and A. K. Maity, “Invariant moment based feature analysis for abnormal erythrocyte recognition,” Int. Conf. Syst. Med. Biol. ICSMB 2010 - Proc., no. December, pp. 242–247, 2010.

[13] X. Gual-Arnau, S. Herold-García, and A. Simó, “Erythrocyte shape classification using integral-geometry-based methods,” Med. Biol. Eng. Comput., vol. 53, no. 7, pp. 623–633, 2015.

[14] R. C. Gonzales and R. E. Woods, Digital Image Processing, 2nd editio. Prentice Hall, 2002.

[15] M. M. R. Krishnan et al., “Automated characterization of sub-epithelial connective tissue cells of normal oral mucosa: Bayesian approach,” TechSym 2010 - Proc. 2010 IEEE Students’ Technol. Symp., no. April, pp. 44–48, 2010.

[16] M. Gonzalez-Hidalgo, F. a Guerrero-Pena, S. Herold-Garcia, A. Jaume-I-Capo, and P. D. Marrero-Fernandez, “Red Blood Cell Cluster Separation from Digital Images for use in Sickle Cell Disease.,” IEEE J. Biomed. Heal. informatics, vol. 2194, no. c, pp. 1–11, 2014.

[17] H. Berge, D. Taylor, S. Krishnan, and T. S. Douglas, “Improved Red Blood Cell Counting in Thin Blood Smears,” in 978-1-4244-4128-0/11/$25.00 © IEEE, 2011, pp. 204–207.

[18] F. Statistics, A. M. Goon, and M. K. Gupta, “BOOK-REVIEWS,” vol. 2, pp. 0–1.

[19] M. a Karaolis, J. a Moutiris, D. Hadjipanayi, and C. S. Pattichis, “Assessment of the risk factors of coronary heart events based on data mining with decision trees.,” IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 3, pp. 559–566, 2010.

[20] D. K. Das, C. Chakraborty, B. Mitra, A. K. Maiti, and A. K. Ray, “Quantitative microscopy approach for shape-based erythrocytes characterization in anaemia,” J. Microsc., vol. 249, no. 2, pp. 136–149, 2013.

[21] R. .Duda, Pattern Classification, 2nd edition. wiley, 2001.

Enhanced Analytical Hierarchy Process for U-Learning with Near Field Communication (NFC) Technology

Huzaifa Marina Osman[1], Manmeet Mahinderjit Singh[2],
Azizul Rahman Mohd Shariff[4], Anizah Abu Bakar[5]
School of Computer Sciences, University of Science
Malaysia, Penang,
Malaysia

Manuel Serafin Plasencia[3]
Universidad Nacional Experimental Politechnica
Bolivar,
Venezuela

***Abstract*—Integration of current Virtual Learning Environment (VLE) system with the Near Field Communication (NFC) technology provides Ubiquitous Learning Environment (ULE) in education. The utilization of NFC technology in U-Learning concept will help to improve accessibility and encourage collaborative learning methods in the education sector. In this paper, we conduct a study to investigate eleven (11) adoptions factors of U-Learning with NFC and ranking them using Analytical Hierarchy Process (AHP) a multicriteria decision-making (MCDM) approach. We also utilized Technology Acceptance Model (TAM), Technology Readiness (TR), and combination of TAM and TR (TRAM) as theoretical framework. We have identified TRAM as the best tool based on literature review and utilized the theory to propose an NFC-Enabled Ubiquitous Technology model. The model was utilized to design a questionnaire for survey about user acceptance. Results from the online survey were analyzed using AHP in an absolute measurement approach method. Results from AHP show that optimism is the most influencing factor in adoption of U-Learning using NFC technology followed by innovativeness and accessibility. Finally, this paper contributes in designing an NFC research model.**

Keywords*—Ubiquitous learning (U-Learning); virtual learning; multi criteria decision making (MCDM); Analytical Hierarchy Process (AHP)

I. Introduction

The emergence of wireless and sensor-based technology such as Near Field Communication (NFC) offers great potential to be utilized in education system. The NFC technology provides advantages in the context of teaching and learning it allows a better condition in the data exchange process [1].

Integration of current Virtual Learning Environment (VLE) system with the NFC technology provides Ubiquitous Learning Environment (ULE) in education. Both NFC technology and U-Learning are ubiquitous and pervasive concept, which is interconnected with each other. The utilization of NFC technology in U-Learning concept will help to improve accessibility and encourage collaborative learning methods in the education sector [2]. Besides, this utilization will give valuable benefits for education because it could offer an active learning and enhance interaction between students and teachers.

The need of innovation and new approaches are essential in education in order to increase quality of education [3]. There are challenges in current teaching and learning technologies such as: accessibility issues in current VLE system, low student motivation, as well as lack of ICT equipment with current VLE approach.

We proposed three research questions in order to achieve the objectives of the study: i) what is user's acceptance about U-Learning using NFC technology in education?; ii) What are the factors that most impacts the adoption of U-Learning using NFC technology?; iii) What are the applications of U-Learning using NFC technology for education? In consequence three objectives have been identified: i) to study the user acceptance towards the usage of Ubiquitous Learning with NFC technology in education; ii) to rank adoption factors of U-Learning using NFC technology in education using AHP methods in MCDM and iii) to propose a framework for U-Learning using NFC technology applications in education. The outline of the paper is as the following. Section II demonstrates the state of art. Section III and IV present the research methodology and proposed work respectively. Section V and VI focus on the discussions, findings and conclusion.

II. Literature Review

A. Factors in Influencing usage of U-Learning

Six characteristics of U-Learning influences its usage for education: mobility of learning environment, urgency of learning requirements, situating of instructional task, interaction in the learning process, initiative in obtaining knowledge, and combination of instructional content [4]. In contrast, three characteristics of U-Learning that influenced its use was identified: accessibility, immediacy and permanency [5]). However, [6] have combined these characteristics with interactivity and situating of instructional tasks that was proposed by [4] as U-Learning characteristics.

The factors that lead users to use U-Learning are whose provides users with the characteristics of context-awareness, seamless services as well as adaptive services [7]. It is a concept of pervasive and omnipresent whereby it allows content or information to be accessed in the right context [7]. It is clearly demonstrated that the characteristic of U-Learning which is 'anywhere' and 'any time' influenced others to use U-Learning. However, [8] indicate as characteristics of U-Learning: accessibility, immediacy, interactivity, context-awareness and permanency.

Another factor is the integration of wireless technology with web technology and the use of Internet sources that enable users to obtain information according their free time and from where [9], [10]. Previous studies showed that characteristic of mobile device [11] and ease of use are also factors that influenced utilization and usage of U-Learning. [4] and [5] also proposed immediacy, interactivity, accessibility, permanency and situating instructional activities as major characteristics of U-Learning. Accordingly, in this research five characteristics of U-Learning as influencing factors of U-Learning usage was proposed based on the similarities in literature findings.

B. Multi Criteria Decision Making (MCDM)

MCDM refers to a process of decision making with the presence of multiple criteria, which are usually conflicting [12]. Analytic Hierarchy Process (AHP) proposed by [13], [14], [15], is a MCDM effective tool. AHP breaks down complicated MCDM problem by means of a hierarchy to elicit pairwise comparisons judgments that facilitates the evaluation [16], [17]. The judgments are arranged in a pairwise comparison matrix and priorities are derived from its main eigenvector enabling to compare and rank the alternatives. A consistency index is estimated to check transitive consistency of the derived priorities. This method is simple, systematic, dependable, and user friendly and several suitable software options are available to decision makers [18]. According to [19] there are four main steps in AHP process: define problem; formulate hierarchy structure consisting of a goal, criteria and alternatives; elicit pairwise comparisons; and utilize priorities based on comparisons.

C. Middleware for Internet of Things (IoT)

Middleware that need to be considered in IoT with utilization NFC technology is Open NFC because it is appropriate and compatible with any other NFC compliant. Any NFC applications developed using Open NFC middleware is applicable and appropriate to be utilized by any smartphone brands [20]. Open NFC is used as middleware to connect the Network Layer and Application Layer in the architecture of IoT. The research presented Open NFC architecture to propose a framework for NFC application.

Mobile RFID-Enabled Android (MORENA) is the model of the Google Android NFC API. It is the most progressive NFC API for smart devices available in the market. MORENA is modeled to grant the developer to execute applications for NFC-enabled devices without dealing with hardware details [21]. MORENA minimize the complexness in designing an application for RFID-enabled Android devices. In contrast to utilize the Android NFC API in which provides better experience for users, ambient-oriented programming is considered as an alternative for distributed computing, which allows MORENA to distribute spaces [21]. The abstraction provides MORENA with ability to create applications for RFID-enabled Android in five various ways: asynchronous interaction, tracking via connectivity, allows support of beam, First-Class References for IoT objects, and unpaired in time.

D. User Acceptance Models

The user acceptance models discussed in this research are Technology Acceptance Model (TAM) and Technology Readiness (TR). TAM is a model that defined the use of technology and technology acceptance [22]. TAM emerged from the Theory of Reasoned Action [23] with the purpose to provide a better explanation about computer acceptance and the user's attitude about computing technologies [24]. TAM was introduced by [24]. It has two key attributes which are Perceived Ease of Use (PEOU) and Perceived Usefulness (PU). TAM is a theoretical model that helps in explaining and predicting user's behavior towards IT [25]. Davis insisted that PU can best described as key variables that convinced people to utilize a particular system or application which can enhance their work proficiency [24], [26]. TAM has analytically proved in distinctive research studies in diverse contexts [27].

TR model is developed by [28] to describe the user's tendency to accept and use recent technologies to complete their task. TR construct can be described as states of mind as it is closely related with an overall variables and obstacles which resulted in how people's tendency in facilitating new technology [28], and includes four aspects: inconvenient, innovative, insecurity, and optimism. From TR, the Technology Readiness Index (TRI) was developed as an appropriate tool to identify early technology users. TRI is a condition of mind refined from understanding of mental enablers as well as inhibitors of user's willingness regarding current technologies [29].

III. Research Methodology

This research was accomplished by phases. Phase 1: designing a taxonomy of five categories (NFC versus RFID, NFC operation modes, ISO standards, NFC applications, and user acceptance model) for reviewing NFC technology and user acceptance model literature. A website to give basic information about the NFC technology and its implementation in U-Learning was developed. A video presentation about U-Learning with NFC technology was produced and uploaded to the website. A questionnaire was designed and uploaded on a free online survey platform. The questionnaire was prepared based on TR model, TAM model, and some Technology Characteristics, which emerges as TRAM model, the approach currently being implemented [30]. The TRAM model is utilized to propose a model for NFC-enabled Ubiquitous Technology characteristics. Components of NFC-enabled Ubiquitous Technology suitable and appropriate with characteristics of U-Learning using NFC technology were identified by means of a literature review.

Based on the analysis of adopting factors of sensor-based technology such as NFC technology, utilized in current educational technology such as U-Learning, it is essential for individuals to trust in U-Learning using NFC technology will improve teaching and learning in term of quality and efficiency, which is considered as PU. Besides, it allows users to complete their task in right time, which refer to PEOU.

TR model is also appropriate because it is capable to evaluate current technology such as NFC technology because of it constructs appropriateness (innovativeness, optimism, discomfort, and insecurity). Components of NFC-Enabled Ubiquitous Technology Characteristics were adopted because its potential effect on PU. The integration of TAM, TR and proposed NFC-Enabled Ubiquitous Technology Characteristics is considered as a TRAM model, and it is adopted in this research.

Twenty-two (22) questions were laid in two sections survey questionnaire: Demographic Section (8 questions) and Adoption Factors of U-Learning using NFC Technology Section (14 questions associated to any of three constructs taken from the TAM model). Table I display all the determinants used in designing the survey questions. All the questions were designed as multiple choice with a Likert-type scale. The scores ranged from: 1 – Strongly Disagree to 5 – Strongly Agree. The survey questionnaire was designed using KwikSurveys (an online free platform). The online survey was conducted within four weeks' time span and its weblink was distributed via e-mail, word-of-mouth and WhatsApp

TABLE I. DETERMINANTS IN SURVEY QUESTIONNAIRE

Determinant	Item number
Age	1
Gender	2
Level of Education	3
Status	4
Having smartphone	5
Knowledge about sensor-based technology	6-8
Innovativeness	9
Optimism	10
Discomfort	11
Insecurity	12
Responsiveness	13
Smartness	14
Permanency	15
Accessibility	16
Immediacy	17
Interactivity	18
Context-awareness	19
Ease of Use	20
Usefulness	21
Intention to Adopt	22

In this research the NFC technology and U-Learning website was developed using a free platform. The intention of the NFC website has provided the fundamental information about the NFC technology such as operation modes of NFC, differentiation between NFC and RFID and basic information about U-Learning. A video presentation about U-Learning with NFC technology was also included on the website. The website also contains link to the survey questionnaire providing direct access to the respondents once they accessed the website.

The video was created using free video application software. The video focused on issues regarding teaching and learning, the comparison of teaching and learning for teachers and students with and without technology utilization, and how the NFC technology brings great potential, which can overcome existing issues in teaching and learning. The video highlighted the effectiveness of U-Learning using NFC technology.

A pretesting session of the survey was executed. The purpose of pretesting questionnaire was to identify either respondents really comprehend with the concept, terminology and words used in the survey. In the pretest session a brief explanation of the research model was given at the beginning. The outcomes enabled the research team to amend the questionnaire and website by means of analyzing the pretesting session participant's feedback. Based on the participant's feedback new technical terminology, clearer statement of the complicated questions, and simplified instructions was provided.

Target populations in this research are teachers and students in education institutions. The non-random sample was defined by means of respondent self-selection during the time frame. The final sample size was of 125 respondents. The data collected was analyzed using SPSS based on the selected research model and hypothesis.

In phase 2, the data collection from the online survey was processed to become the input to MCDM using AHP in a very new approach. The outcome was an importance ranking of the adoption factors identified in the literature and included in the model towards the idea of U-Learning using NFC technology. This ranking is derived from the synthetized weights obtained from the survey respondents by means of an absolute measurement approach in AHP. In phase 3, a simple framework for U-Learning using NFC technology application was designed using the research generated knowledge.

A. *Sample Description*

The observation on the two determinants in demographic section indicates that respondents are mostly females and are having smartphone as listed in Table II.

Results from determinant of respondents' point of view about sensor-based technology from survey questionnaire indicate that most respondents have some knowledge about sensor-based technology as listed in Table III. The percentage of respondents aware about sensor-based technology is 50.4% and respondents that only heard about sensor-based technology is merely 3.2%.

TABLE II. DEMOGRAPHIC PROFILE OF RESPONDENT

Variables	Categories	Frequencies	Percentage (%)
Gender	Male	58	46.4
	Female	67	53.6
	Total	**125**	**100**
Having smartphone	Yes	125	100
	No	0	0
	Total	**125**	**100**

TABLE III. QUESTION ON RESPONDENTS' POINT OF VIEW ABOUT SENSOR-BASED TECHNOLOGY

Categories	Frequencies	Percentage (%)
I did not know about it at all.	1	0.8
I have only heard about it.	4	3.2
I am aware with the sensor-based technology	63	50.4
I have some knowledge of what it is.	55	44.0
I know all about sensor-based technology.	2	1.6

Table IV indicates how long respondents know about sensor-based technology. The highest percentage of how long respondents know about sensor-based technology is within 6 months to 1 year with 44%, followed by 27.2% of respondents know for 1 to 2 years, 16.8% know for 2 to 3 years and 9.6 % of respondents know for less than 6 months. Only 2.4% of respondents know about sensor-based technology for more than 3 years.

TABLE IV. QUESTION ON HOW LONG RESPONDENTS KNOW ABOUT SENSOR-BASED TECHNOLOGY

Categories	Frequencies	Percentage (%)
Less than 6 months	12	9.6
6 months to 1 year	55	44.0
1 year to 2 years	34	27.2
2 years to 3 years	21	16.8
Over 3 years	3	2.4

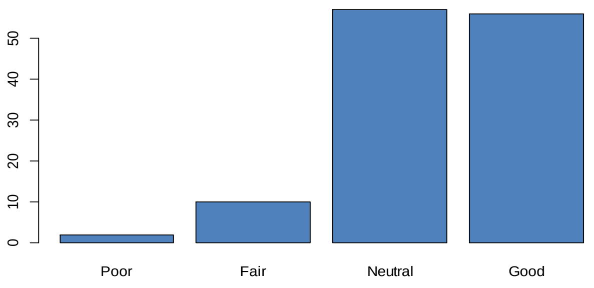


Fig. 1. Survey Respondents' Knowledge about Sensor-based Technology.

The following question examined how respondents rate their knowledge about sensor-based technology in a qualitative scale consisted of: (1) Poor, (2) Fair, (3) Neutral, (4) Good and (5) Excellent. None respondent judges their knowledge as excellent and 45.6% of respondents answered Neutral. Fig.1 shows the responses spread.

IV. RESULTS

All data gathered from survey linked with a research model for Adoption Factors of U-Learning using NFC Technology showed in Fig.2. There are fourteen (14) determinants in the model of Fig.2. For each one of them three (3) items were presented to the respondents in the survey (adding up to 42 items). The respondents could choose one of five categories of agreement (Strongly Disagree, Disagree, Neutral, Agree, or Strongly Agree). Then, each question is a bipolar attribute to express a belief including neutral or zero degree of belief [31]. The answer choices were coded to suitable represent the expressed belief and for further processing as shows in Table V.

TABLE V. CODING VALUE TO INTERPRET ANSWER CHOICE

Answer Choice	Value
Strongly Disagree	-3
Disagree	-1
Neutral	0
Agree	1
Strongly Agree	3

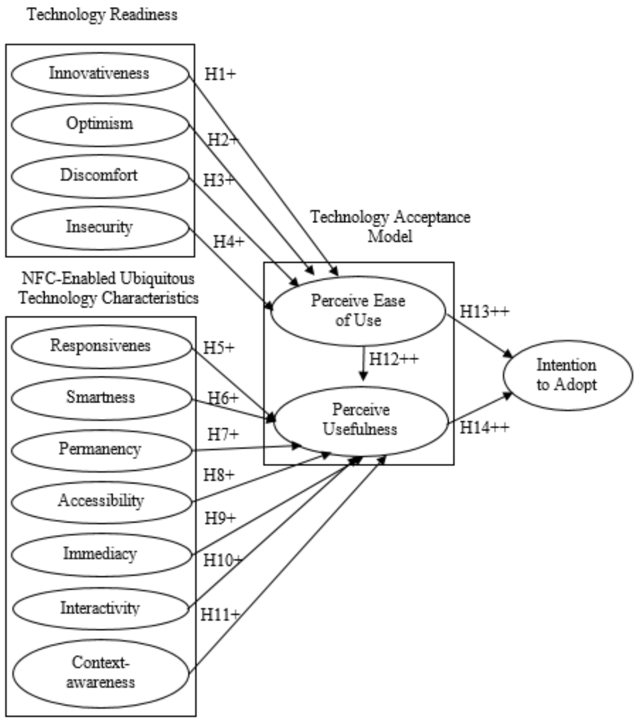


Fig. 2. Model of Research Study.

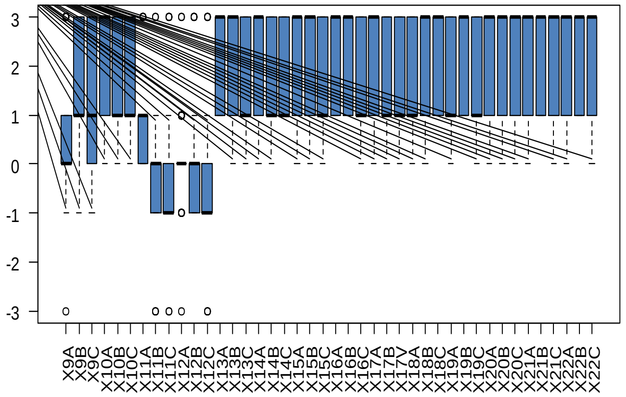


Fig. 3. Boxplot of Answers to Research Model Items.

Fig.3 summarizes the answers of the 42 items showing a high concentration of agreement. Exceptions are showed in items 11 and 12, both in parts B and C. The circles in fig.3 depict outliers in items 9 part A, 11 parts A, B and C, and 12 parts A, B and C. The most intriguing item is 12 part A where the respondent judged risk regarding use of wireless devices in contrast with wired ones, the modal answer for this item was neutral (89 of 125 – 71.4%) than any other answer is atypical.

A. *Multi Criteria Decision Making*

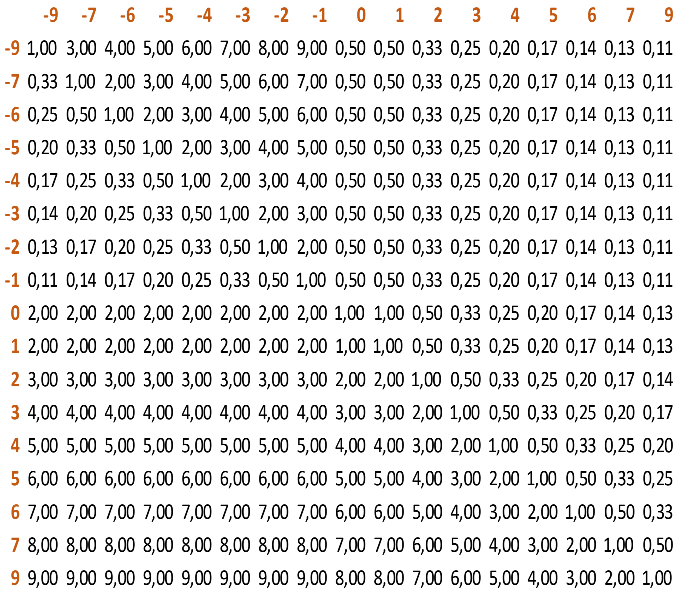
Results from the online survey were utilized in the MCDM process with AHP approach. AHP is a theory of relative measurement able to deal with intangible criteria [32]. The ultimate scope of the AHP is that of using pairwise comparisons between alternatives as inputs, to produce a rating of alternatives, compatibly with the theory of relative measurement [33]. When using relative measurement, the interest is on the proportions between some quantities instead of the exact measurement of them, but from the very beginning [34] attend to absolute and relative measurement as well. The absolute measurement in AHP is called the rating mode [19].

AHP is a method for building an evaluation model with the following main characteristics: (1) evaluation model is structured in a hierarchical way; (2) same assessment technique is used at each node of the hierarchy; (3) assessment of the sub nodes of a common node is based on pairwise comparisons. It works with a minimum of three levels, the top of the hierarchy represented by the goal, in this case it is Intention to Adopt. It is the parent node of the criteria level that presents in this research two elements: PEOU and PU. Finally, the alternative level disclosed four sub nodes under PEOU (Innovativeness, Optimism, Discomfort, and Insecurity) and other seven children nodes to PU (Responsiveness, Smartness, Permanency, Accessibility, Immediacy, Interactivity, and Context-awareness). Each node implies a decision matrix of order nxn where n is the number of sub nodes. There are three matrices one for the goal node (2x2 dimension) and one for each node at the criteria level, at the PEOU node is an order 4 matrix (4x4 dimension), and an order 7 matrix (7x7 dimension) linked to PU node.

TABLE VI. SCALE VALUES FOR EXPERT JUDGMENT ON PROPOSED MODEL ELEMENTS

Combinations of Item's Answers	Value
3 Strongly Disagree	-9
2 Strongly Disagree & 1 Disagree	-7
2 Strongly Disagree & 1 Neutral	-6
2 Disagree & 1 Strongly Disagree	-5
2 Strongly Disagree & 1 Agree	-5
1 Strongly Disagree & 1 Disagree & 1 Neutral	-4
1 Strongly Disagree & 1 Disagree & 1 Agree	-3
2 Neutral & 1 Strongly Disagree	-3
2 Strongly Disagree & 1 Strongly Agree	-3
3 Disagree	-3
1 Strongly Disagree & 1 Neutral & 1 Agree	-2
2 Disagree & 1 Neutral	-2
1 Strongly Disagree & 1 Disagree & 1 Strongly Agree	-1
2 Agree & 1 Strongly Disagree	-1
2 Disagree & 1 Agree	-1
2 Neutral & 1 Disagree	-1
1 Neutral & 1 Disagree & 1 Agree	0
1 Neutral & 1 Strongly Disagree & 1 Strongly Agree	0
3 Neutral	0
1 Strongly Disagree & 1 Agree & 1 Strongly Agree	1
2 Agree & 1 Disagree	1
2 Disagree & 1 Strongly Agree	1
2 Neutral & 1 Agree	1
1 Disagree & 1 Neutral & 1 Strongly Agree	2
2 Agree & 1 Neutral	2
1 Disagree & 1 Agree & 1 Strongly Agree	3
2 Neutral & 1 Strongly Agree	3
2 Strongly Agree & 1 Strongly Disagree	3
3 Agree	3
1 Neutral & 1 Agree & 1 Strongly Agree	4
2 Agree & 1 Strongly Agree	5
2 Strongly Agree & 1 Disagree	5
2 Strongly Agree & 1 Neutral	6
2 Strongly Agree & 1 Agree	7
3 Strongly Agree	9

The numerical coding of the answers of the three items used to survey each of the fourteen determinants was added transforming each factor in a scale of three items. Then, the extreme value in both side of the answers' choices can be associated with a plus or minus sign pointing to the agree side or the disagree side, respectively. Table VI shows the scale score.



	-9	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	9
-9	1,00	3,00	4,00	5,00	6,00	7,00	8,00	9,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-7	0,33	1,00	2,00	3,00	4,00	5,00	6,00	7,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-6	0,25	0,50	1,00	2,00	3,00	4,00	5,00	6,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-5	0,20	0,33	0,50	1,00	2,00	3,00	4,00	5,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-4	0,17	0,25	0,33	0,50	1,00	2,00	3,00	4,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-3	0,14	0,20	0,25	0,33	0,50	1,00	2,00	3,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-2	0,13	0,17	0,20	0,25	0,33	0,50	1,00	2,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
-1	0,11	0,14	0,17	0,20	0,25	0,33	0,50	1,00	0,50	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11
0	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	1,00	1,00	0,50	0,33	0,25	0,20	0,17	0,14	0,13
1	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	1,00	1,00	0,50	0,33	0,25	0,20	0,17	0,14	0,13
2	3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00	2,00	2,00	1,00	0,50	0,33	0,25	0,20	0,17	0,14
3	4,00	4,00	4,00	4,00	4,00	4,00	4,00	4,00	3,00	3,00	2,00	1,00	0,50	0,33	0,25	0,20	0,17
4	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	4,00	4,00	3,00	2,00	1,00	0,50	0,33	0,25	0,20
5	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	5,00	5,00	4,00	3,00	2,00	1,00	0,50	0,33	0,25
6	7,00	7,00	7,00	7,00	7,00	7,00	7,00	7,00	6,00	6,00	5,00	4,00	3,00	2,00	1,00	0,50	0,33
7	8,00	8,00	8,00	8,00	8,00	8,00	8,00	8,00	7,00	7,00	6,00	5,00	4,00	3,00	2,00	1,00	0,50
9	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	8,00	8,00	7,00	6,00	5,00	4,00	3,00	2,00	1,00

Fig. 4. Comparison Matrix for the New Absolute Scale.

TABLE VII. IDEALIZED PRIORITIES FROM SCALE SCORES

Scale Score	Idealized Priority
-9	0,1920
-7	0,1300
-6	0,1047
-5	0,0847
-4	0,0694
-3	0,0580
-2	0,0501
-1	0,0450
0	0,1238
1	0,1238
2	0,1867
3	0,2624
4	0,3551
5	0,4697
6	0,6118
7	0,7873
9	1,0000

In order to get the priorities from the scale a decision matrix is built by means of researcher judgment about the relative classical AHP reciprocal measurement between the different scale scores as shows in Fig.4. The weights extracted from the matrix (Fig.4) as its principal right eigenvector using R software to calculate it. Then, idealized priorities were calculated by means of dividing each weight value by the biggest one. Table VII shows the idealized priorities obtained.

From the respondent ratings from each item, the scale values for each factor and each individual was determined by adding them, then change the scale value by the correspondent idealized priority using Table VII.

TABLE VIII. CONTRIBUTING WEIGHTS FOR ADOPTION FACTORS DERIVED FROM ALL INDIVIDUALS RESPONSES

Criteria	Criteria Weight	Factors	Weights Regarding Criteria	Final Factor Weight
Perceived Ease of Use	0,487	Innovativeness	0,289	0,1406
		Optimism	0,535	0,2603
		Discomfort	0,096	0,0468
		Insecurity	0,080	0,0389
Perceived Usefulness	0,513	Responsiveness	0,141	0,0725
		Smartness	0,144	0,0737
		Permanency	0,144	0,0737
		Accessibility	0,145	0,0746
		Immediacy	0,142	0,0729
		Interactivity	0,145	0,0744
		Context-awareness	0,140	0,0717

In order to aggregate the 125 opinions about the eleven factors we use the geometric mean following [35] whose suggest the geometric mean as the unique suitable procedure for aggregate values expressed in ratio scales. Finally, the weights are the result of normalize those averages in any set of each criteria PEOU and PU respectively. This procedure was done again to get the relative weights of each criteria. The final weights are then calculated by multiplying weights of the factor by weight of the criteria associated with them. Final and intermediate weights are showed in Table VIII.

The main assumption using AHP is that each value judgment comes from an expert about the issue of concern. The sample that answered this research survey was not homogeneus on their expertise and knowledge, then it is reasonable to suppose that not all of them could provide equal responses in terms of informative value. All the respondents were consulted with three items adressing Intention to Adopt directly then it is possible to derive statistical measures about the predictive power of the model about intention to adopt using the final factor's weights obtained by using AHP and the survey answers to the intention to adopt. Both are very strength associated as its correlation coefficient shows (r=0.875).

It is also possible to determine the residuals from the predictions for each individual's Intention to Adopt in relation to their actual scale score for the same variable resulting in a maximum relative residual of 170,8% with a mean of 21,2%. More than 80% of residuals are in the fewest residual classes. Applying the Pareto principle [36] is possible to make a second iteration in the AHP procedure excluding all individual's responses, which predicted intention to adopt relative residual is bigger than 36,0%. 106 individuals remain in the reduced sample of more accurate individuals and their relative residual mean drops to 12,2%. The resulting adjusted weights are presented in Table IX. The new correlation coefficient increases to r = 0,982 both measures validate predictive power of the adjusted weights of factors to evaluate the intention to adopt.

Based on results from AHP method, Table X shows the ranking of all adoption factors for U-Learning using NFC Technology.

TABLE IX. ADJUSTED WEIGHTS FOR ADOPTION FACTORS DERIVED FROM THE REDUCED SAMPLE OF MORE PRECISE INDIVIDUAL'S RESPONSES

Criteria	Criteria Weight	Factors	Weights Regarding Criteria	Final Factors Weight
Perceived Ease of Use	0,503	Innovativeness	0,307	0,1547
		Optimism	0,524	0,2637
		Discomfort	0,094	0,0473
		Insecurity	0,074	0,0374
Perceived Usefulness	0,497	Responsiveness	0,142	0,0705
		Smartness	0,145	0,0721
		Permanency	0,142	0,0708
		Accessibility	0,145	0,0721
		Immediacy	0,144	0,0714
		Interactivity	0,143	0,0709
		Context-awareness	0,139	0,0691

TABLE X. RANKING OF ADOPTION FACTORS IN EDUCATION

Ranking	Factors	Final Factors Weight
1	Optimism	0,2637
2	Innovativeness	0,1547
3½	Accessibility	0,0721
3½	Smartness	0,0721
5	Immediacy	0,0714
6	Interactivity	0,0709
7	Permanency	0,0708
8	Responsiveness	0,0705
9	Context-awareness	0,0691
10	Discomfort	0,0473
11	Insecurity	0,0374

Results from AHP approach show that Optimism is the most important factor in adoption of U-Learning using NFC technology followed by Innovativeness. Factors of Accessibility and Smartness have the same final factors weight and the ranking shows the tie by half numbers, followed by Immediacy, Interactivity and Permanency, Responsiveness, Context-awareness, Discomfort, and Insecurity accordingly. Based on the findings, factors of TR have high impact in adopting sensor-based technology in education. It shows that users are optimistic and innovate with NFC technology utilization in education. Factor of Accessibility and Smartness also affected adoption of U-Learning using NFC technology. Results from AHP approach illustrated that the characteristics of NFC-Enabled Ubiquitous Technology are important in the adoption of U-Learning using NFC technology. The results support that Accessibility as one of the factors which influence the adoption of U-Learning using NFC technology, which this research refer as an unsolved issue with current VLE.

V. Proposed Implementation

This paper proposed an NFC framework integrated with the conventional repository system such as Frog VLE based on literature review and key stakeholder's opinion analysis. The users are system administrators, teachers and students. The system administrator will be using the web application, whereas teachers and students will be using both web application and NFC application. There are three modules proposed which are administrator, teacher, and student module.

A. Administrator Module

System administrator of current Frog VLE enrolls students according to class application. Teachers will be assigned with particular subjects. Students Registration System administrator will register students' information into the system database. Teachers Registration-System administrator will register lecturers' information into the system database.

B. Teacher Module

1) Teaching material management: Teachers can upload the learning materials such as notes and exercises for the specific subjects into the system database. Teachers also enable to edit the content of learning materials. Access control is applied in this module to protect learning material from being edited by students.

2) Interactive learning management: Teachers can store learning materials from system database such as notes into NFC tags or NFC stickers and can attach the URL of the web application for specific subjects for additional information in NFC tag or sticker.

C. Student Module

1) Learning material management: Students can download the learning materials such as notes and exercises for specific subjects from the system database.

2) Interactive learning management: Students can access learning materials from NFC tag or NFC sticker and access the URL of the web application for specific subjects for additional information provided by teachers.

Fig.5 illustrates the use case diagram for NFC application development. The information from use case is useful for development phase of the System Development Life Cycle (SDLC).

Fig.6 illustrates the proposed NFC application framework which consists of System, Applications, Middleware, NFC Controller and NFC Features. The System layer consists of

VLE system, while Applications consist of modules from use case diagram. The middleware layer consists of two choices such as Open NFC and MORENA. The NFC Controller layer consists of NFC modes of operation and NFC Features layer consist of NFC devices.

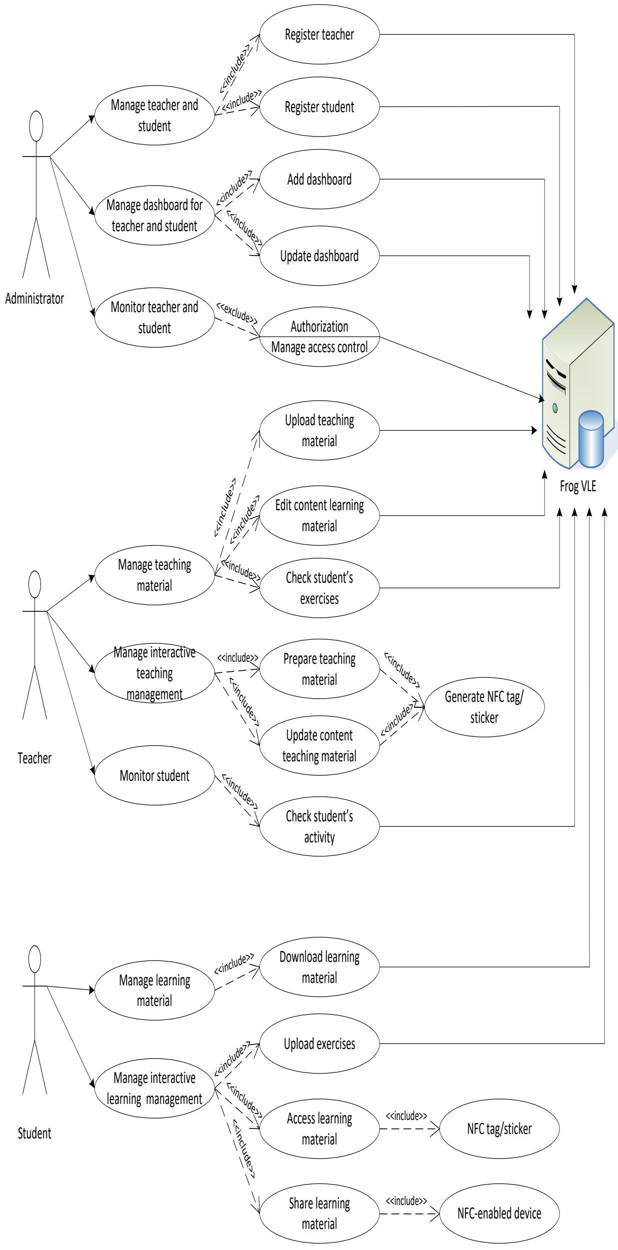


Fig. 5. Use Case Diagram for NFC System.

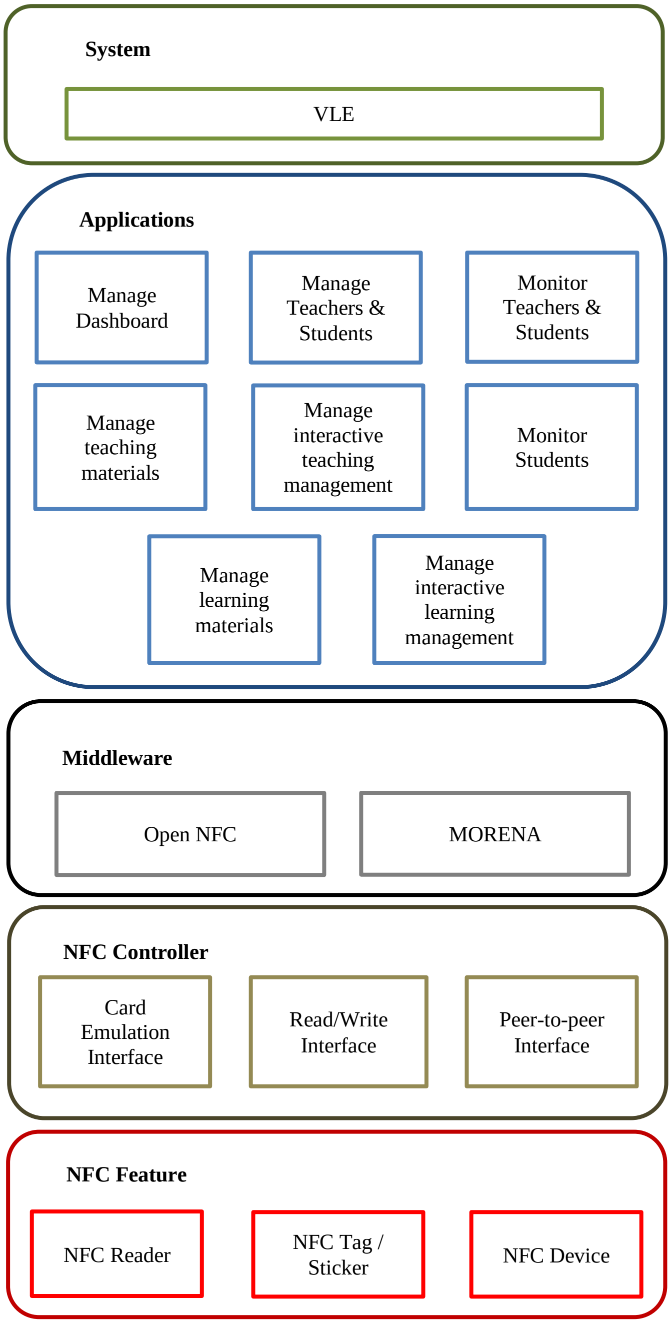


Fig. 6. Proposed NFC Application Framework.

Fig.7 illustrates the proposed NFC application framework for the NFC system using NFC application. It consists of various functions such as accessing, sharing, viewing and downloading learning materials. The web application is used for configuration of user's information, uploading and editing content of learning materials.

The factor of accessibility is an element in the research problem which motivated this research study. The idea of integrating current VLE system with NFC technology is suitable and appropriate to provide U-Learning concept for education sector. Therefore, the researcher has proposed VLE as centralized database and NFC application as framework through the NFC system integration framework.

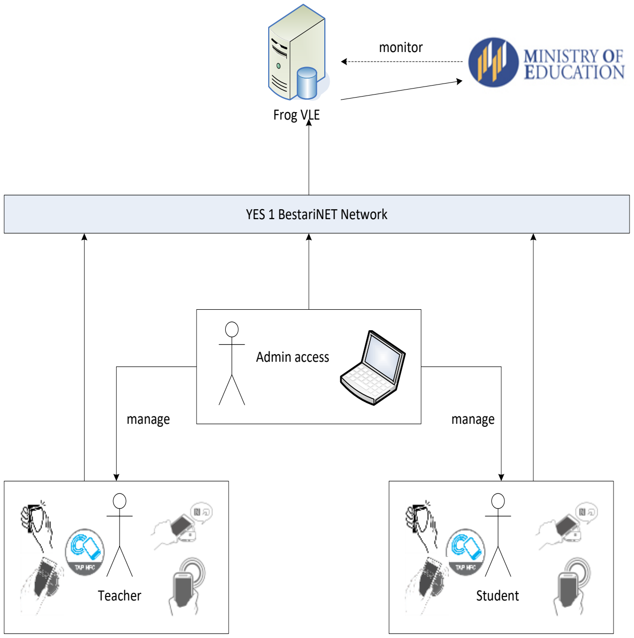


Fig. 7. Proposed NFC Application Architecture.

Based on findings from online survey, it can be claimed that accessibility of using NFC technology for U-Learning has positive impact on user's PU. The technology utilization will provide better solution in accessing current VLE and provide U-Learning environment simultaneously. The proposed NFC application framework will enhance the concept of U-Learning and provide solution for current VLE systems.

VI. Contribution and Discussion

The research contributions are:

1) Design an NFC research model: The theoretical study regarding the U-Learning and NFC technology, which enables the researcher to propose NFC-Enabled Ubiquitous Technology Characteristics research model based on TRAM model. The proposed research model was used to analyze the adoption factors of U-Learning using NFC technology.

2) Ranking of adoption factors: The theoretical study about MCDM helps the researcher to rank adoption factors of U-Learning using NFC technology with AHP approach.

The factors of Optimism and Innovativeness have significant impact for U-Learning using NFC technology. Optimism and Innovativeness are constructs under TR. According to Parasuraman [28] Optimism is important when introducing new technology such as NFC technology. In this research, Optimism has the most impact factor by mean the respondents have positive view about using NFC technology in education. Innovativeness shows that respondents know that this factor allow them to shift from conventional to an advanced technique of teaching and learning. Innovativeness allows users to become pioneer with utilization of NFC technology in education [28].

Accessibility also provides significant impact on adoption of U-Learning using NFC based on survey and AHP results. This factor is one issue in current VLE implementation, whereby it only allows users to access VLE using specific devices such as laptop and tablet. Still, it needs users to key-in to access VLE in which becomes a constraint for users with disabilities such as visually impaired. According to [8] accessibility is one of U-Learning characteristics, which allow users more flexible access to information. By using NFC technology, it allows normal users and users with disabilities to access current VLE using NFC-enabled device with simple touch paradigm.

VII. Conclusion and Future Work

The evolution in education has transformed these processes from conventional technique to an advanced technique. It has been proven by the emergence of E-Learning, M-Learning and U-Learning concepts in education. Advancement in computing technology has brought potential to utilize computing technology within teaching and learning. As a result, the NFC technology could provide benefits for both teachers and students.

This research results support the suggestion of implementation of NFC application for U-Learning in education based on the proposed framework in the future. The implementation will provide benefits for education enhancing the process of teaching and learning and achieving the target to provide education sector with 21st century learning environment. This research shows some evidence about the implementation of the proposed framework could overcome constraints and challenges in current VLE.

The modular architecture in the proposed framework could be enhanced and further extended in the next phase of the implementation by including a module for people in administration, management of education institution and module to record attendance for students and academic staffs in future works. The research results advance the case for U-Learning implementation in real educational system using NFC technology.

References

[1] Ebner, M., & Maierhuber, M. (2013). Near Field Communication-Which Potentials Does NFC Bring for Teaching and Learning Materials?. International Journal of Interactive Mobile Technologies (iJIM), 7(4), 9-14.

[2] Zahrani, M. S. (2010). The benefits and potential of innovative ubiquitous learning environments to enhance higher education infrastructure and student experiences in Saudi Arabia. Journal of Applied Sciences(Faisalabad), 10(20), 2358-2368.

[3] Hoic-Bozic, N., Mornar, V., & Boticki, I. (2009). A blended learning approach to course design and implementation. IEEE transactions on education, 52(1), 19-30.

[4] Chen, Y. S., Kao, T. C., Sheu, J. P., & Chiang, C. Y. (2002). A mobile scaffolding-aid-based bird-watching learning system. In Wireless and Mobile Technologies in Education, 2002. Proceedings. IEEE International Workshop on (pp. 15-22). IEEE.

[5] Curtis, M., Luchini, K., Bobrowsky, W., Quintana, C., & Soloway, E. (2002). Handheld use in K-12: A descriptive account. In Wireless and Mobile Technologies in Education, 2002. Proceedings. IEEE International Workshop on (pp. 23-30). IEEE.

[6] Ogata, H., & Yano, Y. (2004). Context-aware support for computer-supported ubiquitous learning. In Wireless and Mobile Technologies in Education, 2004. Proceedings. The 2nd IEEE International Workshop on (pp. 27-34). IEEE.

[7] Hwang, G. J., Chin-Chung, T., & Yang, S. J. (2008). Criteria, strategies and research issues of context-aware ubiquitous learning. Journal of

Educational Technology & Society, 11(2).

[8] Yahya, S., Ahmad, E. A., & Jalil, K. A. (2010). The definition and characteristics of ubiquitous learning: A discussion. International Journal of Education and Development using Information and Communication Technology, 6(1), 1.

[9] Tatar, D., Roschelle, J., Vahey, P., & Penuel, W. R. (2003). Handhelds go to school: Lessons learned. Computer, (9), 30-37.

[10] Churchill, D., & Churchill, N. (2008). Educational affordances of PDAs: A study of a teacher’s exploration of this technology. Computers & Education, 50(4), 1439-1450.

[11] Tsai, P. S., Tsai, C. C., & Hwang, G. H. (2010). Elementary school students' attitudes and self-efficacy of using PDAs in a ubiquitous learning context. Australasian Journal of Educational Technology, 26(3).

[12] Xu, L., & Yang, J. B. (2001). Introduction to multi-criteria decision making and the evidential reasoning approach (pp. 1-21). Manchester: Manchester School of Management.

[13] Saaty, T. L. (1977). A scaling method for priorities in hierarchival structures. Journal of Mathematical Psychology, 15(3), 234-281.

[14] Saaty, T. L. (1980). The Analytic Hierarchic Process. Nueva York: McGraw-Hill.Saaty, T. L. (1980). The analytic hierarchy process: planning, priority setting, resources allocation. New York: McGraw, 281.

[15] Saaty, T. L. (2006). Fundamentals of Decision Making and Priority Theory with The Analytic Hierarchy Process (2 ed.). Pittsburgh, PA, USA: RWS Publications.

[16] Aruldoss, M., Lakshmi, T. M., & Venkatesan, V. P. (2013). A survey on multi criteria decision making methods and its applications. American Journal of Information Systems, 1(1), 31-43.

[17] Alonso, J. A., & Lamata, M. T. (2006). Consistency in the analytic hierarchy process: a new approach. International journal of uncertainty, fuzziness and knowledge-based systems, 14(04), 445-459.

[18] Kumar, S., Luthra, S., Haleem, A., Mangla, S. K., & Garg, D. (2015). Identification and evaluation of critical factors to technology transfer using AHP approach. International Strategic Management Review, 3(1-2), 24-42.

[19] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International journal of services sciences, 1(1), 83-98

[20] Vazquez-Briseno, M., Hirata, F. I., Sanchez-Lopez, J. D. D., Jimenez-Garcia, E., Navarro-Cota, C., & Nieto-Hipolito, J. I. (2012). Using RFID/NFC and QR-code in mobile phones to link the physical and the digital world. Interactive Multimedia. Dr. Ioannis Deliyannis (Ed.) InTech, 219-242.

[21] Carreton, A. L., Pinte, K., & De Meuter, W. (2012). MORENA: a middleware for programming NFC-enabled Android applications as distributed object-oriented programs. In Proceedings of the 13th International Middleware Conference (pp. 61-80). Springer-Verlag New York, Inc.

[22] Park, S. Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. Journal of Educational Technology & Society, 12(3), 150.

[23] Bradley, J. (2009). The technology acceptance model and other user acceptance theories. In Handbook of research on contemporary theoretical models in information systems (pp. 277-294). IGI Global.

[24] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

[25] Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. Information & management, 40(3), 191-204.

[26] Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. Information systems research, 6(2), 144-176.

[27] Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies?. Decision sciences, 30(2), 361-391.

[28] Parasuraman, A. (2000). Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. Journal of service research, 2(4), 307-320.

[29] Matthing, J., Kristensson, P., Gustafsson, A., & Parasuraman, A. (2006). Developing successful technology-based services: the issue of identifying and involving innovative users. Journal of Services Marketing, 20(5), 288-297.

[30] Lin, C. H., Shih, H. Y., & Sher, P. J. (2007). Integrating technology readiness into technology acceptance: The TRAM model. Psychology & Marketing, 24(7), 641-657.

[31] Rossiter, J. R. (2010). Measurement for the social sciences: The C-OAR-SE method and why it must replace psychometrics. Springer Science & Business Media

[32] Saaty, T. L. (2016). The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In Multiple Criteria Decision Analysis(pp. 363-419). Springer, New York, NY.

[33] Brunelli, M. (2015). Introduction and Fundamentals. In Introduction to the Analytic Hierarchy Process (pp. 1-15). Springer, Cham.

[34] Saaty, T. L. (1986). Absolute and relative measurement with the AHP. The most livable cities in the United States. Socio-Economic Planning Sciences, 20(6), 327-331.

[35] Aczél, J., & Saaty, T. L. (1983). Procedures for synthesizing ratio judgments. Journal of Mathematical Psychology, 27(1), 93-102. doi:10.1016/0022-2496(83)90028-7.

[36] Koch, R. (1998). The 80/20 Principle: The Secret of Achieving More with Less, Bantam Doubleday.

Automation Lecture Scheduling Information Services through the Email Auto-Reply Application

Syahrul Mauluddin[1]
Departement of Informatic Management
Universitas Komputer Indonesia
Bandung, Indonesia

Leonardi Paris Hasugian[2], Andri Sahata Sitanggang[3]
Departement of Information System
Universitas Komputer Indonesia
Bandung, Indonesia

***Abstract*—The study program of information systems is one of the largest studies programs at Indonesian Computer University (UNIKOM). In the process of scheduling lectures in the study program of information systems, it has already information systems of used desktop-based lecture scheduling. Lecture schedules that have been created are then informed through various media such as trust online, social media, email and bulletin board. With so many media which are used in the delivery of lecturers schedule it is expected that lecturers, students and laboratory staff can obtain schedule information properly. However, this also frequently causes problems in learning activities like misplaced of room, time, class, and so on. This usually occurs because the schedule in one of the communication media about lecture schedule is not updated when there is a change of schedule, so there are differences in the schedule information among lectures, students and laboratory staff. To overcome these problems, it needs a service center of lecture schedules information to facilitate lecturers, students and laboratory staff in obtaining the latest lecture schedules information. Related to this, in this study we propose a design of email auto-reply application that will be the service center of lecturer schedules information. In this study, the research method is the method of object-oriented approach and the method of prototype system development. In building email auto-reply application, we are using the Java programming language with MySQL database. With the applications, it is expected that lecturers, students and laboratory staff can obtain the latest lecture schedule easily and from the same source, so different lecture schedules among lectures, students and laboratory staff do not happen again.**

Keywords—Email; auto-reply; service center; lecture schedule; java programming

I. Introduction

The lectures scheduling process is one activity that is very important in the academic activities of a college. A good academic activity should be supported by a good lecture scheduling system starting from the process of scheduling until the deployment of schedule information that has been made so that academic activities can be run according to the academic calendar which has been specified.

The study program of information system is one of the largest studies programs at Indonesian Computer University. In the process of lecture scheduling in the study program of information systems, it has been already using the desktop-based information system of lecture schedule that can help to overcome the problems of conflicting schedules. The lecture schedules that have been created are then informed through various media such as trust online, social media, email and bulletin board. With so many media which is used in the delivery of schedule information, it is expected that all lecturers and students can get timetable well. However, this is also frequent to cause any problems in the lecturing process.

In the lecture scheduling, there are still common problems, such as the wrong room, wrong time, wrong class, and so on. The problems usually occur because of the schedule information in one of the communication media that the schedule is not updated when there is a change of schedule, resulting in differences in the schedule information between lecturers and students. If there is a scheduling problem usually students meet the secretariat of the study program to ask for the phone number of lecturers and also ask for the latest schedule. To overcome these problems, it needs service center of lecture schedule information to facilitate lecturers, students and laboratory staff in obtaining the latest lecture schedule information.

Based on the evaluation result of the methods of lectures schedule information dissemination that are running and the evaluation of information system of lecture scheduling application, then to meet these needs, it is proposed to create an application that will be the information service center of lecture schedule in the form of email auto-reply applications that will send lecture schedule information automatically according to the request which is sent via email.

With the above application, it is expected to lecturers, students and laboratory staff can obtain lecture schedule information easily without having to visit the secretariat and obtain lecture schedule information from the same source so that it does not happen anymore the schedule difference between lecturers, students and laboratory staff.

II. Literature Review

A. Application

The application is “software that is created to work on solving specific problems” [1].

B. Email

“Electronic mail (Email) is a letter which is delivered through an electronic device that is called a computer. To send and to access email on the mail server is required protocol “[2].

C. Simple Mail Transport Protocol (SMTP)

Simple Mail Transport Protocol (SMTP) is “used to communicate and send emails to server. This protocol function is only focused as the email sender to server. It does not function as the email recipient of the server. So, SMTP could not find any message in the mailbox or create an email in the email application directory. So, this protocol is known as Mail Transfer Agent (MTA). This protocol runs on port 25 by its default”[2].

D. Post Office Protocol Version 3 (POP3)

“Version 3 of the Post Office Protocol (POP3) is comparatively simple, and only allows the user to download emails from the server to the client. The user can log in to an account, view the contents of the mailbox, transfer and delete emails, and log out, all via server port 110. This requires few resources, and there is little to configure, which means few sources of error. Emails are stored locally on the user’s PC, which saves precious storage space on the server and reduces backup times. The user usually has to download all emails before deciding which ones are worth reading, based on the subject and/or the sender, although by now most clients support filters for screening incoming mail messages”[3].

E. Internet Message Access Protocol (IMAP)

“It is the development of POP3 protocol which is the system works is more complex. This protocol works when the mail client user connects to the mail server and successfully verify based on the user and password. Then users read the email that come that is when IMAP protocol works, this protocol that visualizes to users the email which were read by the users of the email client”[2].

F. Email Gateway

“Email Gateway is an application-based functionality that allows someone to start a script by sending an email message to the address that we want to go. This application can automatically process email messages from customers and provide feedback process by sending an email notification. Email Gateway acts as an SMTP email recipient. However, in addition to storing or forwarding received email messages, email gateway can specify the script that should be used to process email messages. The script is used to receive and process the content of email messages or to respond to email messages, logs, and to create automatic error handling procedures, which can be changed in accordance with the user wishes. By using an email gateway, transaction reports can be sent automatically to the email administrator” [4].

G. Unified Modelling Language

“Unified Modeling Language (UML) is a family of graphical notation that is supported by single meta-model, which enables the description and design of software systems, especially systems which are built using object-oriented programming (OO). UML is a relatively open standard that is controlled by Object Management Company (OMG), an open consortium that consists of many companies”[5].

H. Use Case Diagram

“Use Case is a technique to record the functional requirements of a system. Use Case describes a typical interaction among the system users with the system itself, by giving a narrative of how the system is used. Use Case diagram shows which the actors are using use case that includes another use case and the relation between actor and use case”[5].

I. Activity Diagram

“Activity diagram is a technique to describe procedural logic, business processes, and work flow. In some cases, activity diagram plays a role which is similar to flowcharts, but the principle difference is between flow diagram notations that activity diagram supports behavior parallel. Node on an activity diagram is called the action, so the diagram displays an activity that is composed of action”[5].

J. Sequence Diagram

“A sequence diagram is a dynamic diagram that shows what happens during the time. In this diagram, all the details of the operations are specified, and the messages that each object involved in the operation sends to the others is detailed together with the time instant at which it happens. A sequence diagram is a time-based representation on messages in the system because it shows time line of events that happen in the system. Sequence diagram shows what happens when a particular flow through a use case or activity diagram is executed. It also shows set of collaborating objects, in addition to showing the messages that pass between them”[6].

K. Lecture Schedule

The definition of lecture schedule refers to the definition of learning activities scheduling according to Tomhart. “Scheduling of teaching and learning activities is arrangements of teaching and learning planning which include subjects, lecturers, time and place at university. In general schedule of teaching and learning activities are presented in the days of the week table which consists of the time slot. It is comprised of subjects, days, hours, and teachers in accordance with the teaching subjects”[7].

III. Research Method

A. Data Collecting Method

The data collected in this study came from two sources as follows:

1) Primary data sources: The data is derived from primary data source which is obtained by using two ways:

a) Observation

Observation is a data collection technique through direct observation of symptoms or events that occur on the research object. In this case writer made some observations to observe the physical state, the location or place of study, namely Studies Program of Information Systems at Indonesian Computer University.

b) Interview

Interview is data collection technique through face to face and direct discussion between data collector (writer) with parties related to the research object. In this case interviews are conducted by secretary and secretariat of information systems study program.

2) Secondary data sources: The data are derived from secondary data sources that are obtained by documentation technique. Documentation is data collection techniques by collecting documents which are related to the research object. In this study, the necessary documents are Lecture Schedule, Minutes of Class, Lecturers' and Students' Phonebook as well as Class Use Table.

3) Research methods: In this study the research method which is used is the method of object-oriented approach. "Object-oriented analysis and design can offer an approach that facilitates logical, rapid, and thorough methods for creating new systems responsive to a changing business landscape" [8]. Method of system development, it is a prototype model. The steps of prototype models can be seen in Fig. 1.

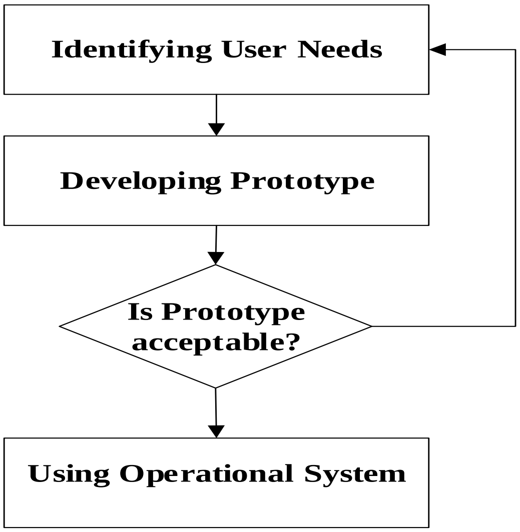


Fig 1. Prototype Model [9].

Here is the explanation of each step of the prototype models:

1) Identifying the user needs. At this step, we interviewed users to get an idea of what it is wanted by the user to the system/application.

2) Developing Prototype. At this step, we do the designing of email auto-reply application with conducting the applications functionality designing, database designing, interface application designing, format designing of email delivery.

3) Determining whether the prototype is acceptable. Users provide input to the analyst whether the prototype is already as the needs or not. If it is not then going back to an earlier step.

IV. Result

This section will explain the steps which are undertaken in designing email auto-reply applications and explain the achievements of each steps. The following steps refer to the prototype model system development methods.

A. Identifying User Needs

The end result of this first step is known that study program of information system requires the service center of lecture schedule information to facilitate all parties who are interested in the lectures schedule.

Based on the evaluation results to the methods of information dissemination on lectures' schedules that are running and the evaluation of lectures scheduling information system application, so to meet the needs of information study program it is proposed to create an application that will be the central of information service to lectures schedule in the form of email auto-reply applications which will send lecture schedule information automatically according to the request which is sent via email. Several reasons make email as a request media and information delivery of lectures schedule as follows: 1) All lecturers, students and laboratory staff have email as it is facilitated by university. 2). The use of email is easier with the email's application on smartphones, tablets and ipad. 3). Email is capable of sending large file size for sending the schedule file in the form of .pdf and .xls files.

B. Developing Prototype

At the step of developing this prototype it will explain the results of the applications functionality design, database design and email delivery format design.

1) Draft of application functionality: At the design step, it will be explained the draft of application functionality is through use case diagram and it will be explained the overview of user activity to joint application on system reaction through activity diagram.

a) Use Case Diagram of Email Auto-reply Application

The functionality of email auto-reply application consists of two that can be seen in Fig. 2.

b) Activity Diagram of Email Auto-reply Applications

The flow picture of process that occurs in each use case can be seen in Fig. 3 and 4.

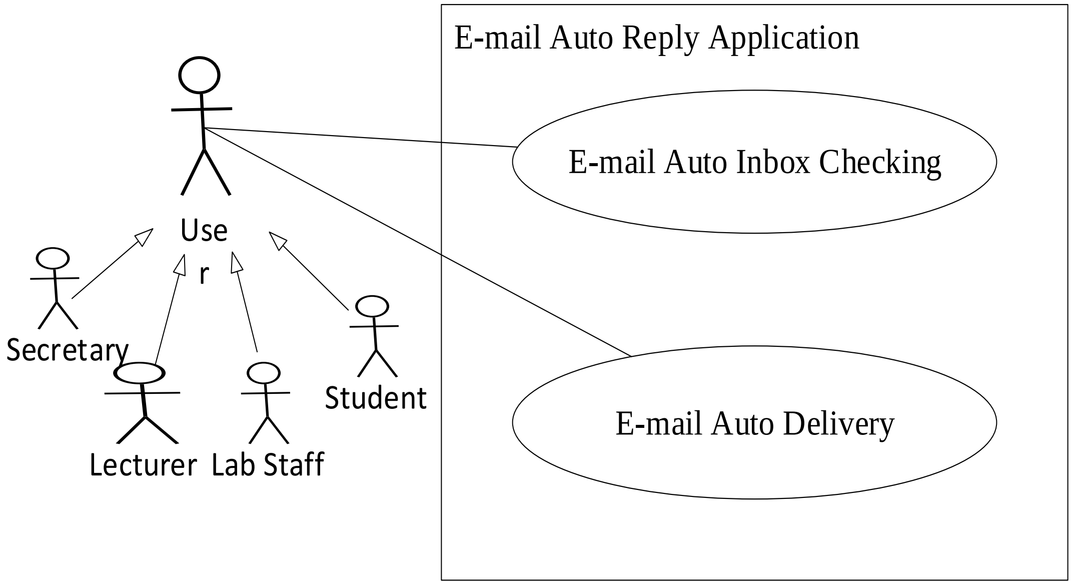


Fig 2. Use Case Diagram of Email Auto-Reply Application.

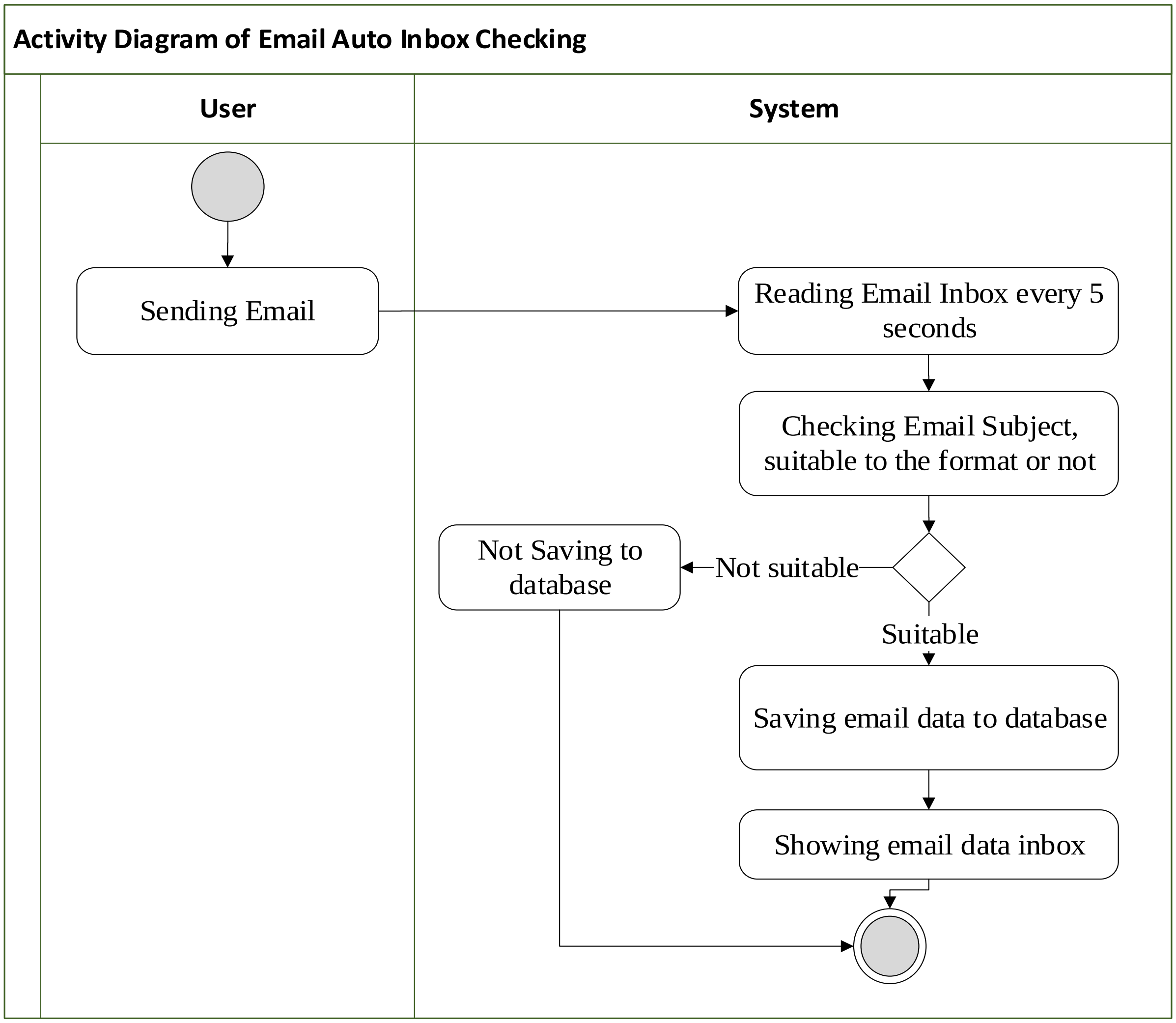

Activity Diagram of Email Auto Inbox Checking
User
System
Sending Email
Reading Email Inbox every 5 seconds
Checking Email Subject, suitable to the format or not
Not Saving to database
Not suitable
Suitable
Saving email data to database
Showing email data inbox

Fig 3. Activity Diagram of Email Auto Inbox Checking.

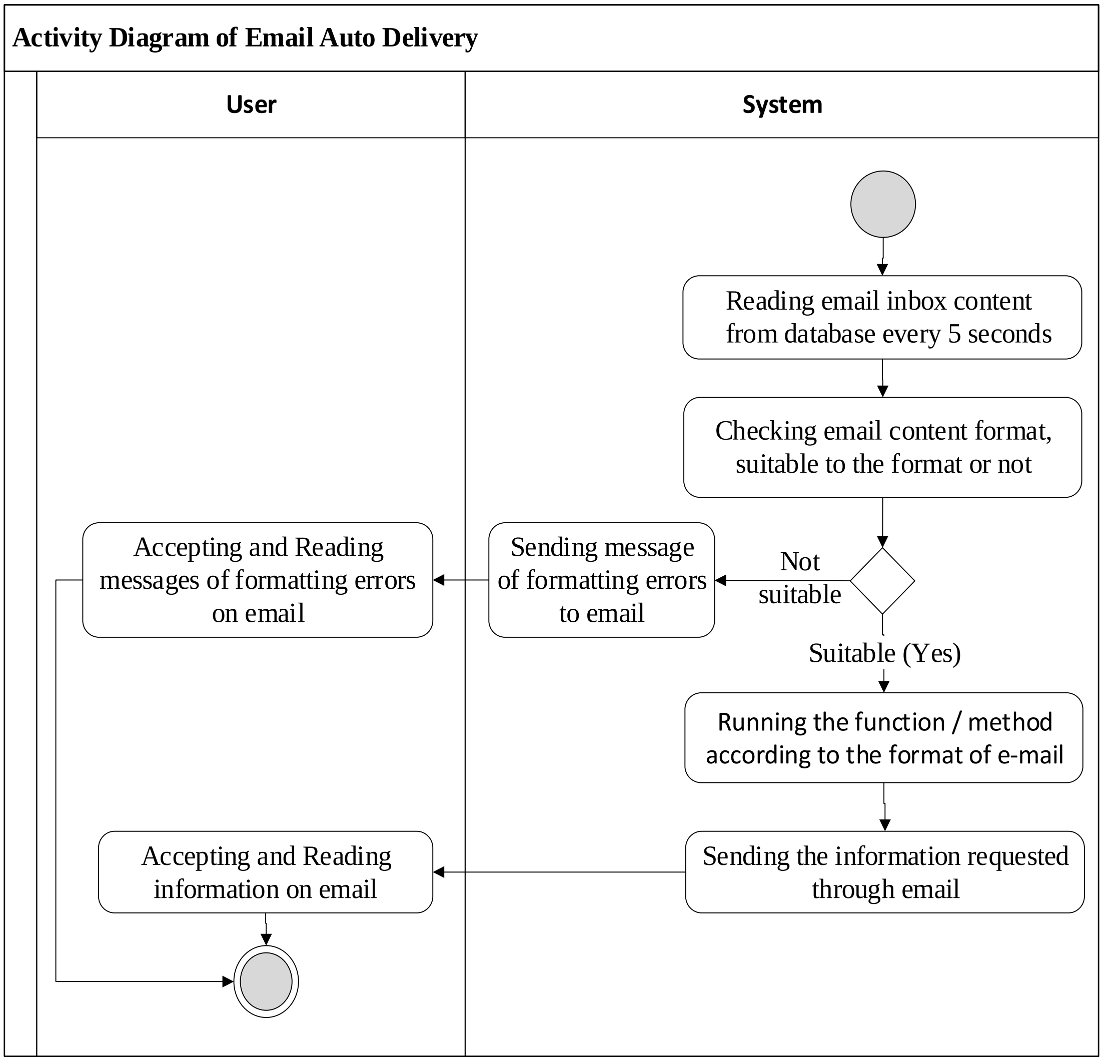

Activity Diagram of Email Auto Delivery
User
System
Reading email inbox content from database every 5 seconds
Checking email content format, suitable to the format or not
Not suitable
Sending message of formatting errors to email
Accepting and Reading messages of formatting errors on email
Suitable (Yes)
Running the function / method according to the format of e-mail
Sending the information requested through email
Accepting and Reading information on email

Fig 4. Activity Diagram of Email Auto Delivery.

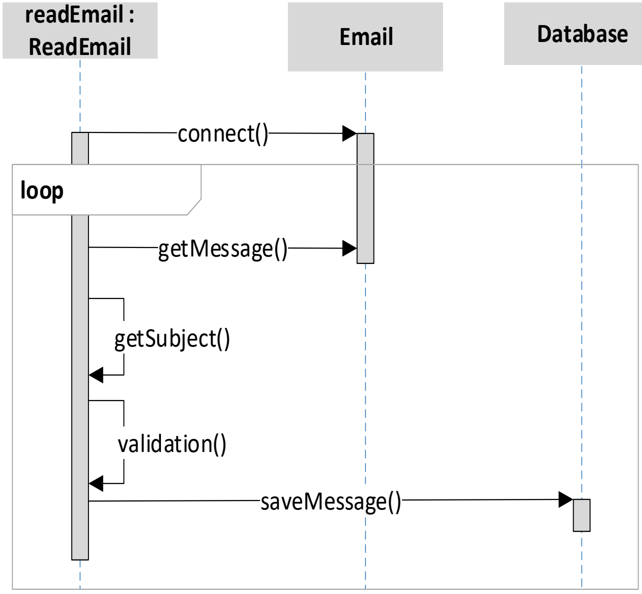


Fig 5. Sequence Diagram of Email Auto Inbox Checking.

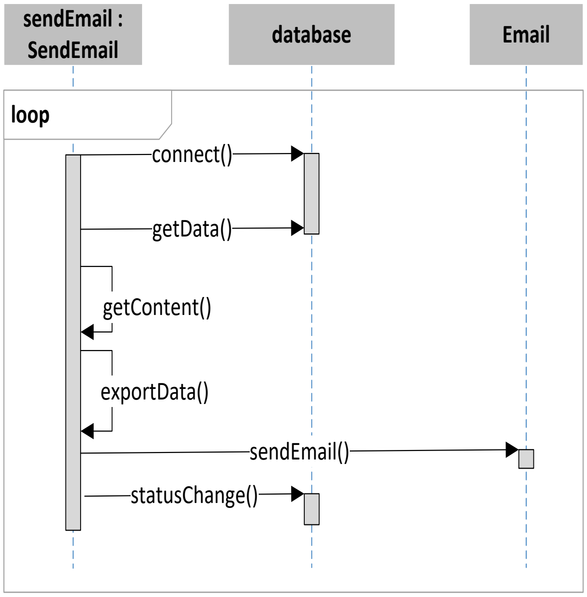


Fig 6. Sequence Diagram of Email Auto Delivery.

c) Sequence Diagram of Email Auto-reply Application

The next step is to create a sequence of diagram to illustrate the logical flow of the developed application. Sequence diagram of email auto-reply application can be seen in Fig. 5 and 6.

d) Format Email Designing to Information Request of Lecture Schedule

To get the schedule of lectures, lecturers, students and laboratory staff should send an email to the specified format. Here is the design format of the email to request lecture schedule information:

1) Email Subject must be filled with the word JADWAL

2) Format of Email Format Guidelines Request
PANDUAN

3) Format of Lecture Schedule Request
JADWAL#SEMESTER#LEVEL
Example: JADWAL#3#S1

4) Format of Schedule for Middle Exams (UTS) Request
UTS#SEMESTER#LEVEL,
Example: UTS#5#S1

5) Format of Schedule for Final Exams (UAS) Request
UAS#SEMESTER#LEVEL
Example: UAS#7#S1

6) Format of Minutes of Class Request (BAP)
BAP#LECTURERCODE
Example: BAP#SYAHRUL

7) Format of Class Use Schedule Request
RUANG # CLASSNAME
Example: RUANG#5204

8) Format of Laboratory Schedule Request
LAB # LABNAME
Example: LAB#LAB4

9) Format OF Lecturer Phone Number Request
TELEPON#LECTURERCODE
Example: TELEPON#SYAHRUL

e) Email Format Design of Reports Inquiry on Email Entry and Email Delivery Data.

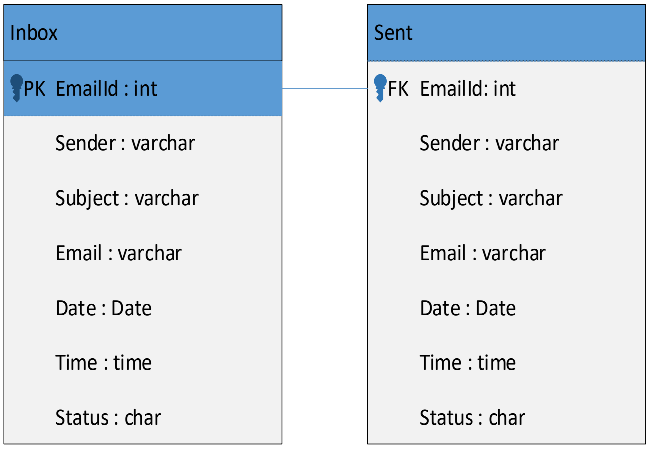
To get email reports of email entry data and email delivery, secretary should send email to the specified format. Here is the design of the email format for information requests on the report data:

1) Subject of the email must be filled with the word JADWAL

2) The format of report requests
LAP# START DATE-END DATE
Example: LAP#30/12/2018-31/12/2018

C. Database Design

Email auto-reply application requires a database to store email data in accordance with a format which is determined. As for the number of tables which are needed for this application is two tables as follows in Fig. 7.



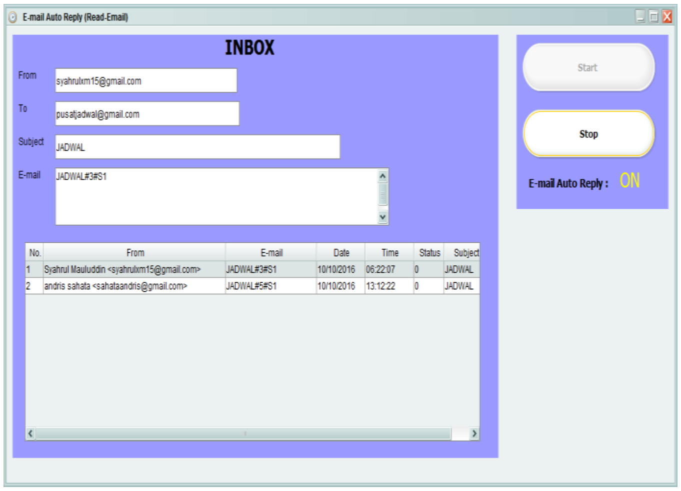
Inbox
PK EmailId : int
Sender : varchar
Subject : varchar
Email : varchar
Date : Date
Time : time
Status : char

Sent
FK EmailId: int
Sender : varchar
Subject : varchar
Email : varchar
Date : Date
Time : time
Status : char

Fig 7. Database Design.

f) Applications Designing of Email Auto-reply

Email auto-reply application consists of two forms: one form to read incoming emails and then save the email data to the database and another for reading email format and then sends the schedule information in accordance with the request format. Both of the forms can be seen in Fig. 8 and 9.



E-mail Auto Reply (Read-Email)

INBOX

From: syahrulxm15@gmail.com
To: pusatjadwal@gmail.com
Subject: JADWAL
E-mail: JADWAL#3#S1

No.	From	E-mail	Date	Time	Status	Subject
1	Syahrul Mauluddin <syahrulxm15@gmail.com>	JADWAL#3#S1	10/10/2016	06:22:07	0	JADWAL
2	andris sahata <sahataandris@gmail.com>	JADWAL#5#S1	10/10/2016	13:12:22	0	JADWAL

E-mail Auto Reply : ON

Fig 8. Form of Email Auto Inbox Checking.

As for the algorithm design of the email auto inbox checking, it can be explained descriptively as follows:

Step 1: Click the start button to activate the timer

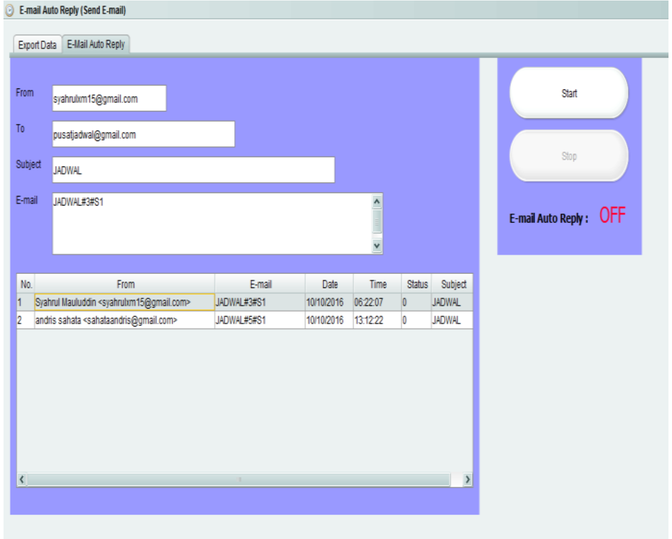
Step 2: Connection to email and inbox folder

Step 3: Every five seconds, take and check the number of emails in inbox

Step 4: If the number of emails is now larger than the number of previous emails then read the last email

Step 5: Check the last email subject if the format is appropriate then save it to database

Step 6: Do repetition of step 3 to 5



From: syahrulxm15@gmail.com

To: pusatjadwal@gmail.com

Subject: JADWAL

E-mail: JADWAL#3#S1

E-mail Auto Reply : OFF

No.	From	E-mail	Date	Time	Status	Subject
1	Syahrul Mauluddin <syahrulxm15@gmail.com>	JADWAL#3#S1	10/10/2016	06:22:07	0	JADWAL
2	andris sahata <sahataandris@gmail.com>	JADWAL#5#S1	10/10/2016	13:12:22	0	JADWAL

Fig 9. Form of Email Auto Delivery.

As for the algorithms design of email auto delivery, it can be explained descriptively as follows:

Step 1: Click the start button to activate the timer

Step 2: Every five seconds do connection to database, show new email data from database to table component

Step 3: Take email data on the first line, take emails contents on email column

Step 4: Split or extract Email format

Step 5: Run method of schedule data export in accordance with email format

Step 6: Send email with attachments

Step 7: Change email data status in database into already sent.

Step 8: Do repetitions of step 2 to 7

V. Conclusion

With the email auto-reply application, it is expected to lecturers, students and laboratory staff can easily obtain lecture schedule information. Then they can get the latest lecture schedule information from the same source so that there is no difference in class schedules between them.

Acknowledgment

This research project is supported by Directorate of Research and Community Service, Ministry of Research, Technology and Higher Education of Indonesia in 2018 Fiscal Year.

References

[1] Daryanto, Keterampilan Dasar Pengoprasian Komputer. (Basic Skills of Computer Operating). Bandung: Yrama Widya, 2004.

[2] A. S. Hidayat, "Perancangan Mail Server Intranet Berbasis Web Base Dengan Optimalisasi Operasi Sistem Client. (Web-Base Based Intranet Mail Server Designing with System Client Operating Optimization)," J. Tek. Komput., vol. 1, no. 1, hal. 1–10, 2015.

[3] P. Heinlein dan P. Hartleben, The Book of IMAP: Building a Mail server with Courier and Cyrus. Munich: Press GmbH, 2008.

[4] T. M. Zakaria dan O. Wongso, "Studi dan Implementasi Teknologi Flashdisk dan Email Gateway dalam Penyewaan Alat pada Perusahaan X (The Study and Implementation on Technology of Flashdisk and Email Gateway in X Equipment Rental Company)," J. Inform., vol. 6, no. 2, hal. 161–174, 2012.

[5] M. Fowler, UML Distilled 3rd Edition. Yogyakarta: Andi, 2005.

[6] M. Grgec dan R. Mužar, "Role of UML Sequence Diagram Constructs in Object Lifecycle Concept," J. Inf. Organ. Sci., vol. 31, no. 1, hal. 63–74, 2007.

[7] F. Tomhart, Afriyudi, dan M. Bakti, "Optimasi Penjadwalan Perkuliahan Di Universitas Tridinanti Palembang". Jurnal Ilmiah Teknik Informatika Ilmu Komputer (Lectures Scheduling Optimization at the University of Tridinanti Palembang," J. Ilm. Tek. Inform. Ilmu Komput., vol. 11, no. 2, hal. 1–11, 2013.

[8] E. K. Kendall dan E. J. Kendall, Systems Analysis and Design, 7th ed. New Jersey: Pearson Prentice Hall, 2008.

[9] I. Ikbal dan S. Mauluddin, "Classroom Booking Information System Integrated with Course Scheduling Information System," in IOP Conference Series: Materials Science and Engineering, 2018, vol. 407, no. 1.

RSSI and Public Key Infrastructure based Secure Communication in Autonomous Vehicular Networks

K. Balan[1], A. S. Khan[3], A. A. Julaihi[4], S. Tarmizi[5], K. S. Pillay[6]
Faculty of Computer Science and Information Technology
University Malaysia Sarawak
Kota Samarahan, Sarawak, Malaysia

L. F. Abdulrazak[2]
Research Management Center
Computer Science Department,
Cihan University Sulaimani, Iraq

H. Sallehudin[7]
Faculty of Information Science and Technology
University Kebangsaan Malaysia
UKM Bangi, Selangor, Malaysia

***Abstract*—Autonomous Vehicular Ad hoc Networks (A-VANET) is also known as intelligent transportation systems. A-VANET ensures timely and accurate communications between vehicle to vehicle and Vehicle to Roadside Unit (RSU) to improve road safety and enhance the efficiency of traffic flow. Due to open wireless boundary and high mobility, A-VANET is vulnerable to several security threats especially impersonation, denial of service, pollution attacks. This paper presents a novel Received Signal Strength Indicator (RSSI) based public key infrastructure (PKI) to address the above-mentioned attacks. Each incoming signal will be authenticated based on RSSI value and digital signal (obtained using PKI) is utilized for cryptography and communication within the insecure channel. The proposed solution is verified with and without the presence of attacker by evaluating the packet delivery ratio and packet overhead.**

Keywords*—Autonomous; vehicular ad hoc networks; public key infrastructure; received signal strength indicator

I. Introduction

Vehicular Ad Hoc Networks is also known as intelligent transportation systems. The aim is to provide inter-vehicle communication and roadside to vehicle communication to in-creasing road safety, improving local traffic flow and the efficiency of road traffic by providing accurate and timely information to road users [1], [2]. In VANET, vehicles are used as network nodes as seen in Figure 1. Security necessities are imperative to provide safe communication in VANET. Due to high mobility, security is more challenging because nodes constantly change network topology. Due to its open-access nature, additionally, VANET is powerless against pollution, Denial of Service (DoS), impersonation, and message fabrication attacks. Thought process of the attackers is to alter the message content, to occupy from different malicious attack, to get the information required, send false message and make network resources become unavailable to others. The various security attacks adopted by an attacker such as pollution attack, impersonation attack, DoS attack and fabrication attack do not only invade driver's confidentiality but also pose risks to the driver, which can cause serious harms/injuries or, worse, loss of lives.

Autonomous vehicles are a type of self-driving car in this current technology of world [3]. The number of autonomous vehicles being used on the road is increasing day by day. Self-driving car, driver-less car or robotic cars are capable to perform an action and navigate without human input or driver responsibility. The autonomous car will have its own GPS (Global Positioning System) as function to locate the user destinations [4]. Alongside other technology that has emerged during this time, the idea of autonomous car can be categorized as an excellent idea, however there are still draw back happened. As the autonomous car is wireless, it depends on a system for exchanging data or information between other vehicles in order to avoid collision on the road [5], [6]. The use of wireless sensor network (WSN) system creates an opportunity to the attacker such as hacker to attack the car system and function of the autonomous car.

An attacker could launch pollution attack by sending malicious or useless data to the target vehicles in order to reduce the vehicles performance. The attacker also created a great number of fake messages to interrupt the vehicle and make it malfunction. In addition, some attackers will distract those good vehicles from malicious attacks in order for other attackers to attack successfully.

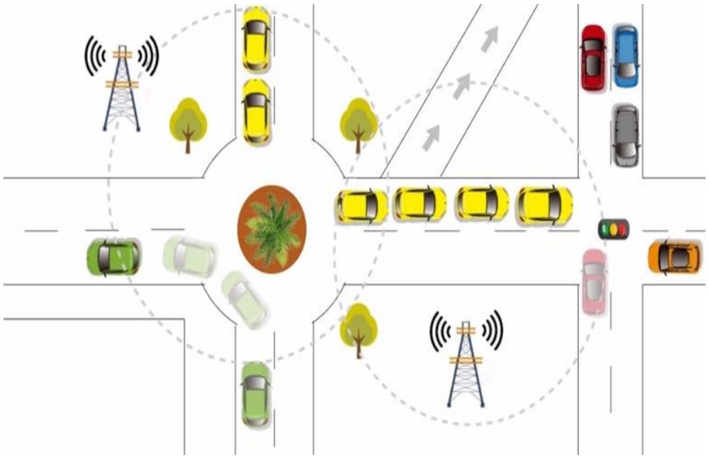


Fig. 1. Basic VANET Structure.

The suggested solution for controlling the vulnerability of VANET system is road side unit that depends on public key infrastructure. Encrypted message can only be decrypted using the owner's private key. Meantime, the proposed solution will request the vehicles to register and road side unit will validate their personal information before accessing to the communication network. The technique used to evaluate the proposed solutions is the measurement of Packet Delivery Ratio and Packet Overhead. It clearly shows the measurement of it based on the proposed solutions with and without attacks. It shows the difference when there is and there is none of the attackers that attacks and disturb the communication network. Firstly, the message is encrypted by public key and then it will decrypt using the owner's private key to read the information or data in it. The technique has enabled the message transfer between two parties in a secure manner, indirectly mitigate the various security attacks based on public key infrastructure. With the aid of road side unit and public key infrastructure, the defense and security of the communication network will increase significantly by allowing the vehicles to register themselves as a valid vehicle in the network. Moreover, the VANET system will become robust and no longer vulnerable to various malicious attacks.

II. Related Works

Timed Efficient Asymmetric Cryptography (TEAC) uses asymmetric cryptography to protect them from attacks [7]. Asymmetric basically uses public and private key in VANET system to validate any user before granting them access to it. Timed Loss Tolerant Authentication (TESLA) is utilized as an authentication technique for broadcast network communication. Although Public Key Infrastructure (PKI) is one of the ways to authenticate different user, but in this case, TESLA can be used to authenticate user instead of using PKI. TESLA is not only systematic for signatures but also more efficient and secure. TESLA also discards used messages if the verification is the out of date.

To enhance In-Vehicle Network Security [8], authentication and encryption is one of the most effective way to provide secureness to a controller are network (CAN) frame. Before that, there is four important issues that needs to be considered. First, key management is essential to provide a good authentication and encryption as well as key update. Secondly, the more the message authentication code (MAC), the more it corrupts the transmission effectiveness. Next, by giving authentication and encryption for CAN frames, it will give the frames a serious delay that will affect the effectiveness the vehicles flexibility. Lastly, the compatibility between the CAN bus that has been improved by authentication and encryption and existing devices must be well label.

Resilient Control Strategy [9] has been presented to refine and enhance the vehicles execution. The strategy that used to control the resilient is a totally new version of Cooperative Adaptive Cruise Control (CACC) where extra estimation algorithm is added into the CACC. This estimation algorithm has three main important components and that is a model-based witness, a slower estimator for the vulnerable attack situation and a Luenberger observer. This algorithm has good functions in it as it can detect the DoS attack as soon as it disrupts the communication network. Thus, it can estimate the number of vehicles that are proceeding in the communication network.

For the transmission of messages [10] in VANET to be transmitted safely, anonymous message authentication has been used in this case. To provide efficiency in authentication, cooperative authentication protocols will be used here. Although there are still several reports of cooperative authentication protocols that fail, success report will still be chosen based on the method of cooperative authentication protocols. Moreover, there is no matching problem here when it comes to cooperative and non-cooperative modes. By using a simulation and analyze it, this protocol does not need mode synchronization while there is no message losses as well even if the density is assigned with 200/km2. Lastly, a binary tree can always decrease the transmission size when the messages are updated for a brand-new group key.

Distributed Aggregate Privacy-Preserving Authentication (DAPPA) [11] manage to obtain an improved version of privacy, good and quick message processing and key guarantee freeness without using a perfect tamper-proof devices (TPD). To solve the message broadcast problem in serious cases, a good solution is that to utilize flexible beacon frequencies for message transmission within a communication network.

An Integrated Circuit Metrics (ICMetric) based on Micro Electro-Mechanical System (MEMS) gyroscope has been proposed [12]. The ICMetrics techniques will generate the symmetric key for data communication and attack detection based on gyroscope sensor device reading. The detection system named as ICMetricIDS is a novel intelligent intrusion detection that is based on ICMetrics in VANETs which able to secure external communication system of the self-driving and able to identify the existing and previously unseen attack for instance fabrication, modification and interception. Advantageously, the experiment demonstrates the proposed security system such as Feed-Forward Neural Network (FFNS)-IDS and k-Nearest Neighbors (k-NN)-IDS will able to identify and block the malicious vehicles in VANETs of self-driving and semi-self- driving vehicles.

INTERLOC has been proposed to demonstrate the detect Sybil attack that gives false information about the vehicle location [13]. The INTERLOC will estimate the area of vehicle that has no error and required an observer to process all the data it received. The vehicles must send the exact location and estimating distance to the observer and the observer will calculate the polygon of intersection point of the vehicles. The result of polygon may be varying due to environment. In preventing any error or false information, the estimated distance between the vehicle and observer will always be updated. Advantageously, the experiment demonstrates that INTERLOC performs better in localization and accurately detect Sybil attack. The high accuracy can improve the traffic safety significantly and making INTERLOC a reliable alter- native to GPS.

An Efficient Anonymous Authentication with Conditional Privacy Preserving scheme (EAAP) has been proposed to prevent harmful vehicles or RSU to enter the VANET as well as securing the vehicular communication [14]. The purpose of this scheme is to identify and track any vehicles and RSU which had assault the VANET. Besides, EAAP will provide anonymous authentication via five parts which are registration and key generation, Anonymous Certificate Generation, Signature generation, message verification, and conditional tracking. Based on the proposed scheme, they had proved that it give an authentication with low signature verification cost and certificate. Moreover, it able to provide an efficient conditional privacy tracking system to determine the original identity of the harmful vehicles and provide rapid verification for certificates and signature compared to the other previous reported scheme.

An anonymous and lightweight authentication based on Smart Card protocol (ASC) to enhance the performance of authentication in VANET has been proposed [15]. To secure the VANET, ASC has implemented a low-cost cryptographic for validation of messages data and verification of the real identity of vehicles. Besides, they also had proved that ASC can reduce the cost of computational and communication by 50% compared to the other technique. In this scheme, dynamic change of login identity and password change phase without rely on TA (trusted authority) has been provided by ASC to avoid harmful attack while formal security model is used to prove ASC is secure from computational Diffie-Hellman problem.

III. Security Challenges

In any autonomous VANET, the motive of the attackers is to send fake information to other vehicles and get useful information or data for personal benefit. They can flood the memory of the vehicle full of useless data and also send invalid information to other vehicles in the communication network. To confuse the vehicles on the road, the attacker may also distract the vehicles to allow other malicious attack to attack the vehicles in the communication network. The attacker convinces the neighboring vehicles that there is considerable congestion ahead, then enforced them clear path for the attacker. In addition, the attacker impersonates any priority vehicle (ambulance, public servant protocol etc.) so that they can move more affluent and faster in the right side road.

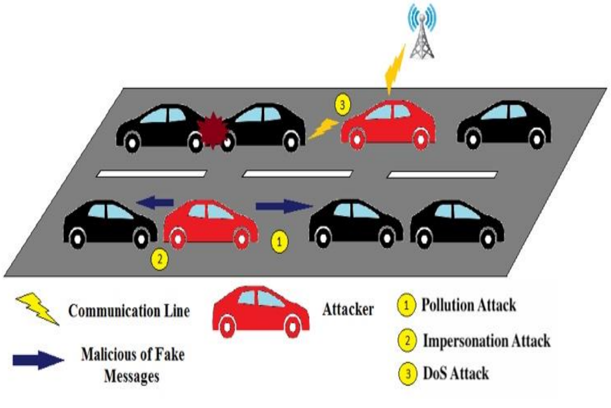


Fig. 2. Scenario on Types of Attack in VANET.

Thus, the attackers may generate Pollution attack, where the attack will only happen if the attacker sends a malicious or useless data to the vehicles until it floods the vehicles memory with bad packet that reduces the vehicles performance (Figure 2). Whereas impersonation attack happens when the attacker steals the identity of the any high priority vehicles on road and try to get benefit of the stolen identity [16]. Next, DoS Attack is an attack that allows the attacker to create large number of fake messages or identities to disturb the data transfer happening between two vehicles that causes jamming and flooding in that particular area. The detailed illustrations of the attacks can be seen in figure 2.

IV. Security Requirements

Based on the above security challenges, it is mandatory to ensure that the communication links are secure enough to perform cryptanalysis. The key three requirements are the confidentiality, integrity and availability to counter the pollution attack, impersonation attack and the denial of service attacks [17].

V. Proposed Solution

This paper proposed detailed architecture for secure communication amongst autonomous vehicles, specifically between the vehicles and vehicles and vehicles to road side units. The idea is to utilize the lightweight and distributed concepts of RSSI localization technique. In this signal strength-based position verification technique, autonomous vehicles will verify the location of other participating vehicles and verifies that at particular time, each physical location is occupied by only one identity. Second, public key infrastructure is utilized during registration process of autonomous vehicles with road side unit. To ensure the freshness of the messages, timestamps are utilized which is first generated by the RSU, and later utilized by all other participating autonomous vehicles lies under the coverage of that particular road side units. The inclusion of timestamp is to ensure that the message is fresh, which consequently minimize the changes of replay or denial of service attacks.

For the proposed solution, road side unit (RSU) will be used to contribute in providing the information of exact location within the database of the system such as bus station, ATM, train station and more. Moreover, vehicle that is in the communication network can communicate with the RSU through vehicle-to-Road Side Unit (V2R) communication. So, in order to discover RSU nearby, vehicle will first send out solicitation packet for discovery. If RSU is successfully discovered, then the RSU will reply by sending a replay in the form of advertisement packet format. Vehicles in the communication network would need to validate their own personal access instead of depending on other communication system. This means that if there are users who want access to the communication network, they would need a proper authorization with an identification for their own following permission. By this, a secure and safe access process will have a positive result in it and this allows the system to get the information or data of the vehicles that enters and exits in the communication network. If there is any abnormality in the information or data of the vehicles that is trying to enter and exit the communication network, this means that there is a fraud within the vehicles or users. The information of data can also be used to characterize and enhance the access for the vehicles that needs to communicate using the communication network.

There are a few phases and operations that the vehicles need to follow in order to prevent the attacker from altering the message or send fake messages to other vehicles. The first phase is the registration phase [18] where the vehicle must complete the registration by providing a complete an accurate personal data and checked by RSU. The vehicle has to submit important vehicular details such as the car registered number plate, model, etc. and the system would need to verify the information given by the vehicle to see if the vehicle is a genuine node. Once the vehicle registration is complete, the system will send a certificate signed by the certificate authority for communication purpose [19]. This will enable the vehicle to have access to the VANET communication network. During the registration phase for the vehicle to enter the communication network, RSU will check if any of the vehicle in the network is blacklisted. In addition, if there is any unregistered vehicle in the network, that particular vehicles information will be pass to the police to track unregistered vehicles. After that, RSU will be ready to serve any vehicles in the network with all the information RSU had after the registration phase.

In order to overcome the attacks before it takes over the system, RSSI (Received Signal Strength Indication) based localization will be implemented into the system. RSSI is a mechanism to secure the communication channels of vehicles by detecting the attack and provide the location to vehicle when the GPS of the vehicles is not working properly. When a vehicle sends signal to another vehicle for location request, RSSI will authenticate by checking each signal in order to detect any sign of attacks before proceeding by enabling sharing locations. Advantageously, the use of RSSI can help in detecting the attacks and remove it from the system in the earlier stage. The RSSI is based on the localization algorithm using the ratio to overcome the attack [20]. Based on theorem 5 in [20], at least 4 sensors that will monitor the radio signals can prevent user from hiding their location. Suppose that node *i* had obtained a radio signal from node 0, the RSSI will be defined as in equation 1.

$$R_i = \frac{P_0 . K}{d_i^{\alpha}} \tag{1}$$

where P_0 is the transmitter power, R_i is RSSI, K is a constant, d_i is the Euclidean distance and □□is the distance-power gradient.

The proposed algorithm is possible to be used to detect attacks such as impersonation. During the receiving of the message, four receiving nodes calculate the physical location of the sender with the sender's ID using equation 2. At any time, if any message from the same physical location with different ID is received, which means that an impersonation attack on VANET. Unfortunately, it is troublesome to calculate the location using equation 2 for each of the participating vehicles to detect the attacks. Considering the fact that all *x*, *y* and x_i, y_i locations are the exact same point, attack can be detected by only through continuous recording of the signal strength in any table and later analyzing the ratio of RSSI by comparison to receive the message.

Figure 3 shows the topology for RSSI in an autonomous vehicular communication scenario. Therefore, we let each of the 4 monitoring nodes having an ID as B1, B2, B3, B4 and the impersonation attack node as S1 and S2 in time.

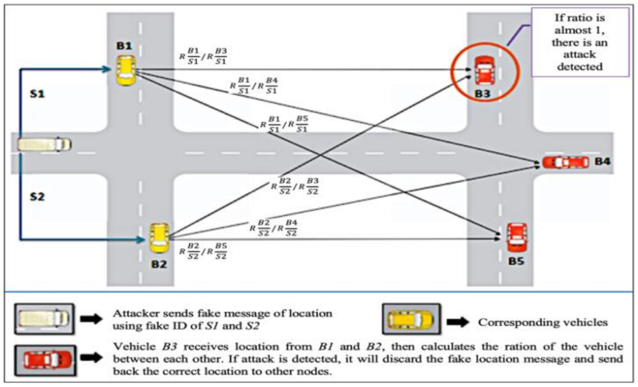

If ratio is almost 1, there is an attack detected
B1
$R\frac{B1}{S1}/R\frac{B3}{S1}$
$R\frac{B1}{S1}/R\frac{B4}{S1}$
$R\frac{B1}{S1}/R\frac{B5}{S1}$
S1
B3
B4
S2
$R\frac{B2}{S2}/R\frac{B3}{S2}$
$R\frac{B2}{S2}/R\frac{B4}{S2}$
$R\frac{B2}{S2}/R\frac{B5}{S2}$
B2
B5
Attacker sends fake message of location using fake ID of *S1* and *S2*
Corresponding vehicles
Vehicle *B3* receives location from *B1* and *B2*, then calculates the ration of the vehicle between each other. If attack is detected, it will discard the fake location message and send back the correct location to other nodes.

Fig. 3. VANET Scenario with RSSI Localization Implementation.

Let's suppose, there exist an impersonation attacker at any time t_1, it will broadcast a malicious message with its impersonated ID as S1. The monitoring node will record the signal strength value and the impersonated ID. For each of the monitoring node, it will send a message to B1 that consisting the received signal strength from S1. Then, we allow R_i^k to represent the value of signal strength when at *i*, it received the message from sender *k*. After that, the messages are collected from the monitors where B1 computes every ratio as

$$\frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_2}}, \frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_3}} and \frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_4}} \tag{2}$$

and locally stores the values. While at time t_2, the malicious node once again broadcasts the message using an alternate ID which is S2. The nodes that monitor the network will then record the signal strength value from S2 and send it to B1. Then, B1 calculates the equivalent ratio (3) which is

$$\frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_2}}, \frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_3}} and \frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_4}} \tag{3}$$

Therefore, B1 now can identify the attack by analyzing the ratio by comparison at time t_1 and t_2. When the value of the difference in the ratio of the two is close to zero, B1 then determines that an impersonation attack has happened in the area. If the ratios of the signal strength are the same, it indicates that the point of attack to fake the multiple IDs is also the in the same place. Due to that, B1 concludes that there is no attack node. If

$$\frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_2}} = \frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_2}}, \frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_3}} = \frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_3}}, \frac{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_1}_{\mathrm{B}_4}} = \frac{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_1}}{\mathrm{R}^{\mathrm{s}_2}_{\mathrm{B}_4}} \tag{4}$$

is true, as in equation 4, B1 had detects an impersonation attack. The most important is when the attacker tries to attack the GPS, which is the critical function for the autonomous car, RSSI can detect it and cut the attacks and send back the location to the vehicle so that the vehicle can arrive at the destination safely.

Secondly, PKI with public key technology is utilized to securely transmit the message from vehicle to vehicle or vehicle to road side unit [21]. The information or data of the vehicle is encrypted in a secure form so that nobody can decrypt the data or the information of the vehicle without knowing their own private key. This helps to improves vehicles satisfaction by enabling communications from all around the network. PKI basically generate a digital certificate. By signing the certificate means that the certificate authority (CA) has verified the private key that corresponds to the public key in the certificate and this will be in the hand of the subject that is named in the certificate.

CA is usually a third-party company that provides a completely trusted vehicle a certificate that contains a public key that used to encrypt the data. The certificate can be distributed freely to anyone and it can only be decrypt using the correlated vehicles private key [22]. The private key must be kept securely to avoid anyone else having access to it. For instance, if vehicle A in the communication network wants to send message that contain some important information or data to vehicle B, then vehicle A can use the public key from a trusted association to encrypt the message. The encrypted message can only be decrypted using the vehicles B private key.

To be precise, PKI manage information and data in a way that it will identifies and authenticate the vehicle simultaneously to prevent attackers from hacking into the communication network because it makes it difficult for the attackers to intercept identities. In this proposed scheme, the value of signal strength encrypted with digital certificate of that autonomous vehicle is transmitted through insecure to ensure that the message came from the legitimate sender as shown below.

$$\text{Authentication Message} = [\text{RSSI}_{\text{value}}]_{\text{digital signature}} \qquad (5)$$

Once the authentication message is digitally signed by the sender digital certificate, which means that the message belongs to the legitimate sender, thus, chances of impersonate is negligible. In conclusion, with the proposed solution suggested above, it will be able to solve all the three attacks stated above and the vehicles that are communicating between each other in the network would be more secure.

VI. Results and Discussion

This study proposed the RSSI localization algorithm to identify the attack autonomous vehicular networks and provide location to the vehicle when the GPS is not working properly. By using the ratio of RSSI localization from multiples receivers, it is possible to overcome the attack from the attacker by detecting the attack through recording and comparing the ratio of RSSI to receive the messages even though RSSI is time-varying, unreliable in general and non-isotopic radio transmission. The RSSI which is based on the localization algorithm which using the ratio to overcome the attack in this project is presented.

A. Evaluation Parameters

In this paper, Packet Delivery Ratio (PDR) and Packet Overhead (PO) are used as the observed parameters for the proposed solution. Both of these parameters are obtained against the different number of nodes in the network.

PDR is the amount of ratio in packets that are successfully transmitted or send to a goal target that sent by the sender [23]. Moreover, packet delivery ratio is also important to identify problems that might cause a poor throughput. PDR will be used to measure the performance of the vehicles using the proposed solution with attacks and using the proposed solution without attacks in the communication network.

PO is the additional cost incurred by the network due to maintaining latest information in the data packet. While additional information may affect the throughput of the network, in reality, the overhead offers actual information of the network availability.

B. Conceptual Results and Analysis

When there is no attack during the exchange of packet between vehicles, the receiver will receive the same amount of packet send by the sender. As shown in Figure 4, Car A successfully delivered the 10 packets to Car B without any loss of packets.

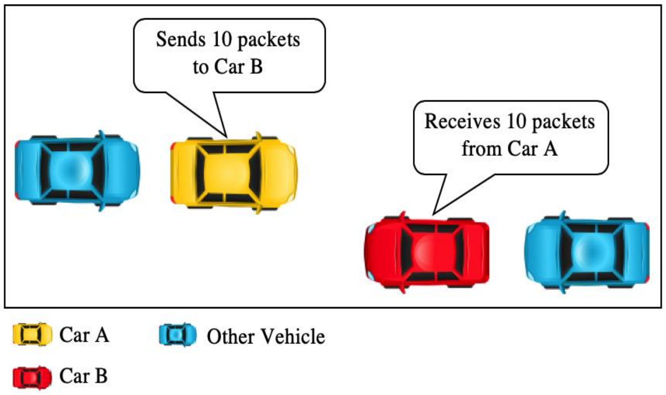


Fig. 4. Packet Transmission without an Attack.

When there is an attack during transmission of the packet between vehicles, the receiving node will receive either a reduced amount or more packet than actual packet transmitted by the sender node. As shown in Figure 5, Car A send the 10 packets to Car B, but when there is an attack, Car B only received 5 packets from Car A.

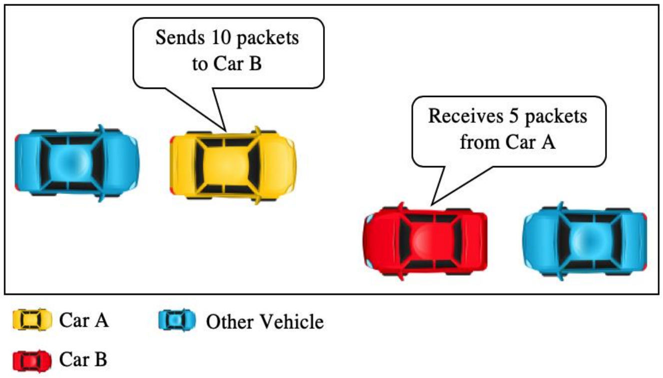


Fig. 5. Car a Sends 10 Packets to Car B but Car B Only Receives 5 Packets.

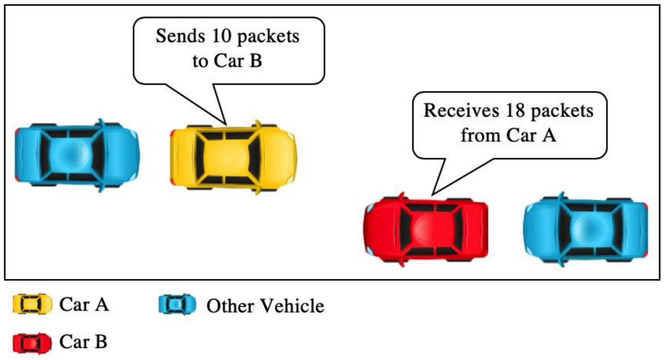


Fig. 6. Car a Sends 10 Packets to Car B but Car B Receives Additional Packets.

Figure 6 depicts another scenario when there is additional packet transmitted during the exchange of information between vehicles the delivery ratio seem to be more than one, clearly indicating that there is a malicious packet inserted by an attacker and the network is compromised.

The graph in Figure 7 indicates where Car A, C, D, F and H sends the specific number of packets to B, D, E, G and I respectively, the receiver will receive the same amount of packet delivered by the sender.

In Figure 7, the graph shows the percentage of packet delivery ratio against the number of nodes during no attack. The proposed solution used RSSI localization where it compares the ratio of RSSI during time, t_1 and t_2 to receive the packet. If

$$Ri - Rj = 0 \quad (5)$$

then there is an attack at the area and the location of the fake attack is at the same place. When the proposed solution is used, it will compare the ratio of RSSI at time, t_1 and t_2 to receive the packet. If

$$Ri - Rj > 0 \quad (6)$$

Then there is no attack and the network is secure for communication with other nodes.

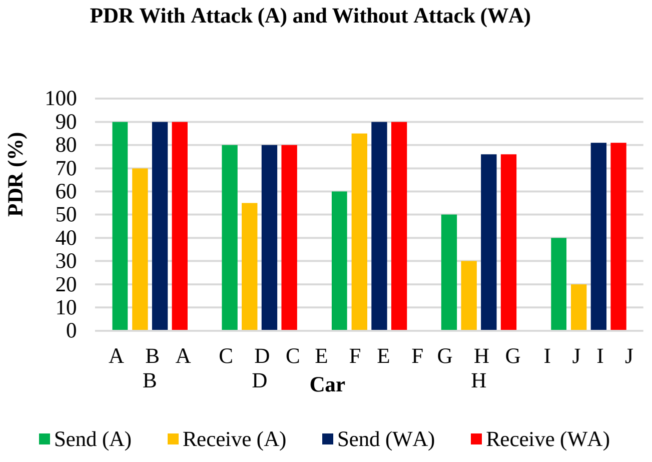


Fig. 7. PDR of Proposed Solution with RSSI Localization During Attack.

Based on Figure 8, the graph shows the percentage of packet overheads against the number of nodes with and without RSSI localization during and attack. It is essential to understand how the increasing number of nodes and speed can affect the packet delivery ratio. The conceptual result that is expected to have a slight difference in the packet overhead in an unsecured VANET network is compared to the one with the implementation of the RSSI. As the number of nodes increase, the gap between the secured and unsecured network increases and we would be able to see a much significant difference in the packet overhead. This increase in the packet overhead could possibly due to the understanding that the attack has widespread among the nodes in a much-congested environment, causing the packet overhead to increase. However, with the implementation of the RSSI localization in the network, the packet overhead seems to be much steadier and kept at a manageable level even at greater traffic congestion. Therefore, we can conclude that there is a significant difference of packet overhead with and without the implementation of RSSI Localization during an attack in the VANET network.

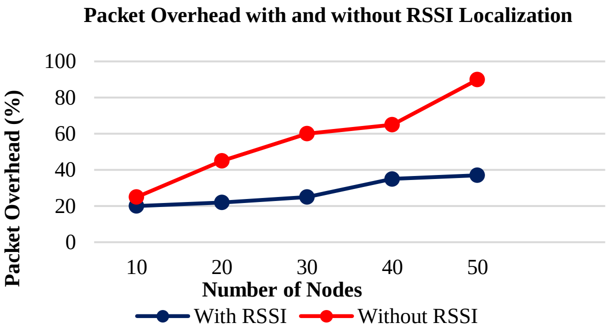


Fig. 8. Packet Overhead with and without RSSI Localization.

VII. Conclusion

In conclusion, a conceptual framework is proposed for autonomous VANET system to overcome denial of service (DoS), pollution and impersonation using signal strength and digital certificate. When vehicle sending the signal to other vehicle for the location request, signal strength will authenticate by checking each signal in order to detect any sign of attacks before proceeding by enabling sharing locations. With the road side unit, user's information security can be enhanced. The efficiency of the proposed solution is conceptually measured using Packet Delivery Ratio and Packet Overhead. With these conceptual measurements, the security of the vehicle and user information is expected to be more secure. The proposed solution is hoped to encounter the various attacks presented above effectively. Lastly, road side unit provides more secure and reliable communication from one vehicle to another vehicle. With road side unit and public key infrastructure, security of the message will increase significantly. For future work, this proposed solution will be implemented and simulated for its performance in terms of packet delivery ratio and packet overhead. In addition to that, the performance of the proposed solution will also be evaluated against other current existing solutions to determine its efficiency.

ACKNOWLEDGMENT

This research is the outcome of collaborative research between University Malaysia Sarawak (UNIMAS) and Cihan University, Sulaimaniya Campus and is fully funded by Research and Innovation Management Centre (RIMC-UNIMAS) and Research Management Center, Cihan University, Sulaimaniya Campus.

REFERENCES

[1] A. Abbasi and A. S. Khan, "A Review of Vehicle To Vehicle Communication Protocols for VANETs In The Urban Environment", Future Internet, 10(2), art. no. 14.

[2] A. Mehmood, A. Khanan, A. H. H. M. Mohamed, S. Mahfooz, H. Song, and S. Abdullah, "ANTSC: An Intelligent Naïve Bayesian Probabilistic Estimation Practice for Traffic Flow to Form Stable Clustering in VANET," IEEE Access, vol.6, pp. 4452-4461, 2017.

[3] Y. R. B. Al-Mayouf, M. Ismail, N. F. Abdullah, S. M. Al-Qaraawi, and O. A. Mahdi, "Survey on VANET technologies and simulation models," ARPN Journal of Engineering and Applied Sciences, vol. 11, no. 15, pp. 9414-9427, 2016.

[4] A. Idris, M. Ismail, M.H Hamdan, M. R. Baharon, N. L. Abdullah, M. H. A. Hamid, A. S. H. Basari, "A Comparative Study between Bluetooth and GPS Tracking System," Journal of Advanced Research in Dynamical & Control Systems, vol. 10, no. 02-Special Issue, pp. 1327-1335, 2018.

[5] Y. R. B. Al-Mayouf, N. F. Abdullah, O. A. Mahdi, S. Khan, M. Ismail, M. Guizani, S. H. Ahmed, "Real-Time Intersection-Based Segment Aware Routing Algorithm for Urban Vehicular Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 7, pp. 2125–2141, 2018.

[6] Y. R. B. Al-Mayouf, O. A. Mahdi, N. A. Taha, N. F. Abdullah, S. Khan, and M. Alam, "Accident Management System Based on Vehicular Network for an Intelligent Transportation System in Urban Environments," Journal of Advanced Transportation, vol. 2018, Article ID 6168981, 11 pages, 2018.

[7] S. V. Mahagaonkar and N. Dongre, "TEAC: Timed efficient asymmetric cryptography for enhancing security in VANET," in 2017 International Conference on Nascent Technologies in Engineering (ICNTE), New Mumbai, India, 2017.

[8] J. Liu, S. Zhang, W. Sun, and Y. Shi, "In-vehicle network attacks and countermeasures: Challenges and future directions," IEEE Network, vol. 31, no. 5, pp. 50-58, 2017.

[9] Z. A. Biron, S. Dey and P. Pisu, "Resilient Control Strategy under Denial of Service in Connected Vehicles," American Control Conference (ACC), 2017.

[10] H.J. Jo, I.S.Kim, and D.H. Lee, "ReliableCooperativeAuthentication for Vehicular Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 4, pp. 1065–1079, 2017.

[11] L. Zhang, Q. Wu, J. Domingo-Ferrer, B. Qin and C. Hu, "Distributed Aggregate Privacy-Preserving Authentication in VANETs," IEEE Trans- actions on Intelligent Transportation Systems, vol. 18, no. 3, pp. 516– 526, 2017.

[12] K. M. A. Alheeti, R. Al-Zaidi, J. Woods and K. McDonald-Maier, " An intrusion detection scheme for driverless vehicles based gyroscope sensor profiling," in 2017 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2017.

[13] M. T. Garip, P. H. Kim, P. Reiher and M. Gerla, "INTERLOC: An interference-aware RSSI-based localization and sybil attack detection mechanism for vehicular ad hoc networks," in 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2017.

[14] M. Azees, P. Vijayakumar and L. J. Deboarh, " EAAP: Efficient anonymous authentication with conditional privacy-preserving scheme for vehicular ad hoc networks," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 9, pp. 2467–2476, 2017.

[15] B. Ying and A. Nayak, "Anonymous and lightweight authentication for secure vehicular networks," IEEE Transactions on Vehicular Technology, vol. 66, no. 12, pp. 10626–10636, 2017.

[16] Irshad Abbasi, A. S. Khan, Shahzad Ali, "A Reliable Path Selection Packet Forwarding Routing Protocol for Vehicular Ad hoc Networks", Eurasip Journal On Wireless Communication and Networking, 2018(1), 236.

[17] A. S. Khan, H. Halikul, M.N.Jambli, R. Thangaveloo, Mitigation of Non-Transparent Rouge Relay Stations in Mobile Multi hop Relay Networks, Advanced Science Letters, 2017 , 23 (6), pp. 5246-5250,

[18] Y. Park, C. Sur, S. W. Noh, and K. H. Rhee, "Secure vehicle location-sharing for trajectory-based message delivery on VANETs," in 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, UK, 2017.

[19] A. H. A. Halim, M. Elshaikh, M. N. M. Warip, and R. B. Ahmad, "Validation of performance analysis for optimized vehicular ad hoc network using Taguchi method," Jurnal Teknologi, vol. 77, no. 32, pp. 133–140, 2015.

[20] S. Zhong, L. Li, Y. G. Liu, and Y. R. Yang, "Privacy-Preserving Location Based Services For Mobile Users In Wireless Networks", Technical Report YALEU/DCS/TR-1297, Yale Computer Science, July 2004.

[21] A. H. Salem, A. Abdel-Hamid, and M. A. El-Nasr, "The case for dynamic key distribution for PKI-based VANETs," International Journal of Computer Networks & Communications (IJCNC), vol. 6, no. 1, pp. 61–78, 2014.

[22] Irshad Abbasi, A. S. Khan, Shahzad Ali, "A Reliable Path Selection Packet Forwarding Routing Protocol for Vehicular Ad hoc Networks", Eurasip Journal On Wireless Communication and Networking, 2018(1), 236.

[23] Irshad Abbasi, A. S. Khan, Shahzad Ali, "Dynamic Multiple Junction Selection based Routing Protocol for VANETs in City Environment", Applied Sciences, 8(5), 687.

Weighted Minkowski Similarity Method with CBR for Diagnosing Cardiovascular Disease

Edi Faizal[1]
Department of Informatics Management
STMIK AKAKOM Yogyakarta
Yogyakarta, Indonesia

Hamdani Hamdani[2]
Department of Computer Science
Universitas Mulawarman
Samarinda, Indonesia

***Abstract*—This study implements Case-Based Reasoning (CBR) to make the early diagnosis of cardiovascular disease based on the calculation of the feature similarity of old cases. The features used to match old cases with new ones were age, gender, risk factors and symptoms. The diagnostic process was carried out by entering the case feature into the system, and then the system searched cases having similar features with the new case (retrieve). The level of similarity of each similar case was calculated using weighted Minkowski method. Cases with the highest level of similarity would be adopted as new case solutions. If the value of similarity was <0,8, the revision would be conducted by an expert. The tests result conducted by the expert showed that the system was able to perform the diagnosis correctly. The test results are performed on the sensitivity of 100% and specificity of 83,33%. Meanwhile, the accuracy of 95,83% and the error rate of 4,17% is so that this research is relevant enough to be implemented in the medical area.**

***Keywords*—CBR; cardiovascular; similarity; weighted minkowski**

I. Introduction

Weighted Minkowski similarity method Cardiovascular Disease (CVD) is the term for a series of heart and blood vessel disorders. World Health Organization data of 2012 shows that CVD is the number one cause of death in the world. In 2008 there were 17,3 million people died from CVD, these numbers represent 30% of the cause of death in the world. There were 7,3 million people died because of coronary heart disease and 6,2 million because of stroke [1].

On this issue, it is necessary for Diagnosing Cardiovascular Disease using Case-based Reasoning (CBR) approach with Weighted Minkowski Similarity. Many of the early systems attempted to apply pure Rule-Based Reasoning (RBR) as reasoning by logic in the expert system [2]. However, for broad and complex domains where knowledge cannot be represented by rules (i.e. IF-THEN), this pure rule-based system encounters several problems [3]. Due to the difficulty of the knowledge acquisition process, computer experts have tried to learn other problem solving methods known as CBR using lambda value analysis on the weighted Minkowski distance model [2].

The knowledge representation of CBR is a case base occurred previously. CBR uses a solution from an earlier case similar to the current case to solve the problem. The method that can be used to calculate similarity is weighted Minkowski [4]. If a new case has a resemblance to the old one, CBR will reuse the old case solution as a recommendation for the new case solution. But if there is none match, CBR will do adaptation by retaining the new case into the case database, so the CBR knowledge will increase [2]. The more cases stored in the case base, the smarter the CBR system will be.

Based on the above facts, it is necessary to establish a system capable of diagnosing cardiovascular disease. The built system is an implementation of CBR in which the problems in new cases are solved by adapting solutions from old cases that have occurred and CBR is an important technique in artificial intelligence, which has been applied to various kinds of problems in a wide range of domains [3]. We use this weighted Minkowski similarity method because it is very good in our case for completing the diagnosis of CVD.

II. Related Work

Several studies in the domain of cardiovascular have been conducted by [5] used a structured poly tree concept and directed acyclic graphical model (DAG) to predict all cases that could cause coronary heart disease. Tests showed that the applied concept was more accurate and efficient in predicting heart disease before the physical examination. In the same years [6] proposed a new algorithm to predict heart disease using CBR techniques. Meanwhile, most of the algorithms are based on Binary data only. The system was implemented using Java and had successfully predicted different levels of risk of heart attack effectively [6].

The application of CBR in the field of cardiovascular disease has been done by [7] by building case-based expert system prototype of heart disease diagnosis. While [8] used CBR in building a multimedia decision support system (MM-DSS) of heart disease diagnosis. [6] used 110 cases for 4 types of heart disease. The two retrieval methods were used namely induction and nearest-neighbour. It showed that the accuracy of using the nearest neighbour method is better than that of the induction method, i.e. 100% and 53%. Meanwhile, [8] medical multimedia based clinical decision support system for operational chronic lung diseases diagnosis and training with 97,36% Sensitivity, 97,77% Specificity, 96,85% positive predictive value (PPV) and 93,90% negative predictive value (NPV).

CBR for diagnosing heart failure in children done by [9]. The research conducted by [10] was for face recognition using 3 (three) different local features namely Manhattan distance, weighted angle distance and Minkowski distance. The results showed that Minkowski distance provided better results in terms of time to recognize faces that were 0.46, 0,45 and 0,43 seconds for Minkowski, Weighted angle and Manhattan respectfully. [11] has developed a mobile cancer management system (MCSM) prototype to diagnose cancer patients. The system developed was a combination of CBR and CBIR with similarity measure using weighted Minkowski method. Based on 600 images of breast cancer radiology tested, it resulted in 90% accuracy.

Based on the explanation above, a great number of researches to diagnose cardiovascular disease have been conducted. In fact, the process of cardiovascular disease diagnosis needs to involve some risk factors, gender and age of patients to improve the accuracy of the diagnosis. Specifically, there has been no research conducted to diagnose the type of Acute Myocardial Infarction (I21) disease. Meanwhile, [12] conducted research on Individual risk prediction model for incident cardiovascular disease using the Bayesian approach.

Research in the application of CBR to make a diagnosis has also been conducted with various degrees of accuracy, while the application of Minkowski method has been performed for certain purposes with a fairly good level of accuracy. This research conducted in this paper applied CBR to diagnose type I21 cardiovascular disease. The diagnostic process involved symptoms, risk factors, age and gender of the patient. The calculation of similarity used weighted Minkowski method. The research was expected to generate a system capable of diagnosing cardiovascular disease, especially type I21 with a good level of accuracy.

Meanwhile, Minkowski central partition model by [13] for the pointer to a suitable distance exponent and consensus partitioning using developed clustering algorithm capable of computing feature weights. In [14], Minkowski metric for feature weighting and anomalous cluster is initializing in K-Means Clustering.

III. Proposed Method

The system input was in the form of risk factors and symptoms data of the patient's disease, and then the data were made into a case. There were two types of cases namely target case and source case. The source case was the case data entered into the system that served as knowledge for the system, while the target case was a new case data of which the solution to be sought.

The diagnostic process began with inserting patients' data, risk factors and symptoms experienced by the patient and then the similarities to the case stored were counted. Each feature had a certain value of weight obtained from the experts. The similarity between features was calculated using local similarity formula, and then calculated as a whole using global equality formula.

The calculation resulted in each case was sorted from the highest value to the lowest value. The highest value was the case most similar to the new case. The value of similarity ranged from 0 to 1 (in the percentage from 0% to 100%). If the value was smaller than the threshold value that was ≥ 0.8, the solution of the case must be re-shared by the expert. The system output was the name of the disease most similar to the new case.

A. Knowledge Acquisition

Case base would be formed from a collection of medical record data of cardiovascular disease inpatient of Dr. Sardjito public Hospital, Yogyakarta. The next stage was to make knowledge acquisition process to collect knowledge data from the knowledge source. The source of knowledge was obtained from an expert (cardiovascular disease specialist / SpJP). In addition to the expert, knowledge material was also derived from the literature related to the problem, such as books, journals, articles, etc.

B. Case Representation

The representation is intended to capture the essential properties problems and make that information accessible to the problem-solving procedure [3]. Case data obtained from medical records were stored in a case base. The collected cases were represented in the form of a frame. The frame contained the relation among the patient data, the illness, the risk factors and the symptoms of the case. Levels of confidence/trust were given on the relationships of these data so that the case for the CMB system could be made based on the representation in which problem space was the risk factor and the symptoms of the disease and solution space where the name of a disease.

Every risk factor and symptom has a weight that indicates the level of importance of the disease. The weight value ranges from 1 to 10 and the greater the value of the weight, the more important the risk factor or symptom determine the patient's disease. The level of confidence showing the sureness of the diagnosis of the expert is based on the risk factors and symptoms experienced by patients.

C. Indexing

The index on a record consists of two parts, search-key (value) and pointer. The search key is the value of a record while a pointer is index position of the search key. Case data searching in the retrieving process requires one or more search keys. In the development of CBR for Cardiovascular disease, two search keys have been developed namely risk code and symptom code.

D. Retrieval and Similarity

Retrieval is the core of the CBR – the process found in the case-base, the cases closest to the current case. The most commonly investigated retrieval techniques so far are the k-nearest neighbour, decision tree and its derivative. This technique uses a similarity metric to determine the size of similarity between cases [1]. In this study, the similarity method used referred to Equation (1) [10].

$$\mathrm{d}(\mathrm{C_i},\mathrm{C_j}) = \left(\frac{\sum_{k=1}^{n} w_k^r * |d(C_{ik},C_{jk})|^r}{\sum_{k=1}^{n} w_k^r}\right)^{\frac{1}{r}} \qquad (1)$$

With $\mathrm{d}(\mathrm{C_i},\mathrm{C_j})$ is similarity value between case C_i (new case) and case C_j (old case), n is number of attributes in each case, k is individual attributes, ranging from 1 to n, w is weight given 1 to *k* attribute and *r* is Minkowski factor (positive integer).

The *r* value was the positive number ≥ 1, (from 1 to infinite). The research presented in this paper used r=3. The previous research conducted by [11] showed that with the use of *r*=3 resulted in maximum accuracy.

The weighted of features in diagnosing cardiovascular disease was necessary because of the difference between particular features. Weight value was obtained from experts/cardiovascular disease specialists.

Due to the weight difference given to features for each case and the handling of new symptoms that may arise in the new case, the equation (1) introduced by [10] needs to be modified. Modifications to deal with similar problems have been carried out by [3] by adding the value of trust and handling of new symptoms as shown in equation (2).

$$T_k(T_i, S_i) = \mathrm{Sim}(S, T) * T_k(S_i) * \frac{n(S_i, T_i)}{n(T_i)} \qquad (2)$$

With $T_k(T_i,\ S_i)$ is similarity normalization with the level of trust, $T_k(S_i)$ is expert confidence level on a case in source case,$n(S_i, T_i)$ is number of symptoms of target case appear in source case and $n(T_i)$ is the number of symptoms in the target case [14].

The modification was made in equation (1) with reference to equation (2) so that the research conducted used equation (3).

$$E(C_i, C_j) = \left[\frac{\sum_{k=1}^{n} \omega_k{}^r * |d_k(C_{ik}, C_{jk})|^r}{\sum_{k=1}^{n} \omega_k{}^r}\right]^{1/r} * T(C_j) * \frac{n(C_i, C_j)}{n(C_i)} \qquad (3)$$

Where,

$E(C_i, C_j)$	:	The global similarity value between case of C_i (target case) and C_j (source case)
ω_k	:	Weight value given to-k attribute
$d_k(C_{ik}, C_{jk})$	:	Local similarity value between attribute C_i to-k and attribute C_j to-k
r	:	Minkowski factor (positive integer)
$T(C_j)$	:	Percentage of case trust level in case base
$n(C_i, C_j)$	:	Number of new case attributes (C_i) and appear in old case (C_j)
$n(C_i)$	:	Number of attributes in new case (C_i)

The similarity of each aspect in two cases is computed by a particular local similarity function. The local similarity values are aggregated by means of a sum of weighted aspects [15]. Local similarities are divided into two types namely symbolic and numeric. Features involved in symbolic is symptoms while in numeric are age, gender, smoking habit, body weight and so forth. The symbolic feature was calculated by using an equation (4) [12].

$$d_k(C_{ik}, C_{jk}) = \begin{cases} 1, \text{if } C_{ik} = C_{jk} \\ 0, \text{if } C_{ik} \neq C_{jk} \end{cases} \qquad (4)$$

The numeric feature was calculated by using an equation (5) [13].

$$d_k(C_{ik}, C_{jk}) = \begin{cases} 1\,, & \text{if } C_{ik=}C_{jk} \\ 0\,, & \text{if } C_{ik} \vee\ C_{jk} = 0\ \vee\ C_{ik}\ \vee\ C_{jk} = \perp \\ \frac{\min(C_{ik}, C_{jk})}{\max(C_{ik}, C_{jk})}, & \text{else} \end{cases} \qquad (5)$$

If the similarity level is high, the case will be reused, in which the old case solution will be reused as a new case solution. If there is no case that meets the threshold value, the expert needs to give a conclusion to the new case.

E. Case Revision

The case revision is the part of system adaptation performed by an expert. The expert would revise the name of the disease and the level of confidence of the disease as the result of the diagnosis which has similarities lower than 0,8. After being revised, the case becomes a new case base.

F. System Implementation

The system is divided into 2 categories based on user types namely expert and paramedic. Each category of the user has access to a system with different facilities. Expert administration has access to add new users to the system, enter knowledge data, enter and revise cases as the result of a diagnosis and to diagnose them. Users with paramedic type have access to input patients' data, diagnose new cases and store new cases.

G. System Assessment

System assessment is carried out by performing diagnostic tests to measure the system's ability to detect disease or exclude a person without the disease. In [2] explained that sensitivity and specificity are used to determine the accuracy of diagnostic tests. Predictive values can be used to estimate disease probabilities, but positive predictive values and negative predictive values vary according to the prevalence of disease.

The analysis was conducted by using 4 parameters namely TP, FP, TN and FN and then they subsequently were used in calculating sensitivity, specificity, PPV and NPV. Calculation of the values using equations (6), (7), (8), (9) [16].

$$\text{sentitivity} = \left[\frac{TP}{TP+FN}\right] * 100\% \qquad (6)$$

$$\text{specificity} = \left[\frac{TN}{TN+FP}\right] * 100\% \qquad (7)$$

$$\text{PPV} = \left[\frac{TP}{TP+FP}\right] * 100\% \qquad (8)$$

$$\text{NPV} = \left[\frac{TN}{TN+FN}\right] * 100\% \qquad (9)$$

$$\text{Accuracy} = \text{sentitivity}\frac{\text{P}}{(\text{P}+\text{N})} + \text{specificity}\frac{\text{N}}{(\text{P}+\text{N})} \quad (10)$$

$$\text{Error Rate} = \frac{\text{FP}+\text{FN}}{(\text{P}+\text{N})}\,\text{x}\;100\% \quad (11)$$

With TP (True Positive) is Positive diagnosis results for positive data samples, TN (True Negative) is Negative diagnosis Results for negative data samples, FP (False Positive) is Positive diagnosis results for negative data samples, FN (False Negative) is Negative diagnosis results for positive data samples, P is Total of positive diagnosis results and T is the total of negative diagnosis results. These values will appear in the Confusion matrix.

According to [16], the Confusion matrix is a useful way to analyze how well the system recognizes the tuples of different classes. TP and TN provide information when the system is correct, while FP and FN notify when the system is incorrect. Sensitivity and specificity can be used for the classification of accuracy. Sensitivity can be designated as true positives (recognition) rate (the proportion of correctly identified positive tuples). While specificity is true negatives rate (the proportion of the correctly identified negative tuples). The function of sensitivity and specificity can be used to show the accuracy level by equation (10) and the level of the system error rate can also be calculated by equation (11).

IV. Result and Discussion

A. *Case Base Filling Process*

The initial stage of the use of the CBR system was the preparation and filling of the base case. The case data inputted in the case base were medical records of inpatients obtained from the medical records installation of Dr. Sardjito Public Hospital, Yogyakarta. There were 126 cases with 74 symptoms and 9 types of risk factors that of class I21 disease (Acute Myocardial Infarction).

Symptoms and risk factors have a weight that indicates the level of importance of the symptoms or the risk factors. The weight of symptoms or risk factors were obtained based on expert data ranging from 1 to 10. Before filling the case base, users must first input patients' data, disease data, Symptom data and risk factor data into the system.

B. *Diagnostic Process*

Generally, the process of diagnosis of cardiovascular disease can be performed by doctors in several ways. The first way is to consider the risk factors and symptoms (felt by the patient). It is known as anamnesis. Another way is to perform laboratory tests to ensure the diagnosis. In the CBR system, the system performs the diagnostic process by anamnesis.

The diagnostic process began with selecting patient data to be diagnosed then entering symptoms data and risk factors felt by a patient into the system by utilizing the input facilities provided. Having all data entered, the system would perform retrieve process and calculate the similarity level between the new case and the case in the case base using weighted Minkowski.

Each case was calculated based on 4 components namely symptoms, risk factors, gender and age. Symptoms were assessed on the basis of the appearance of a symptom in a case. If a certain symptom appeared then it was valued 1 and 0 if otherwise. Gender was categorized into male and female, while patient age was grouped in 6 age ranges. Several risk factors were assessed by their appearance in a case (such as symptom assessment), but there were two risk factors calculated based on a certain range namely disease history and smoking.

For example, there was a case in the case base as shown in Table 1. The user diagnosed a new patient with data entered into the system as shown in Figure 1. Based on the case example, the system performed the process of calculating both cases after the user had clicked the Diagnose Result button.

The process of calculating the local similarity in the case was divided into 4 (four) sections namely of age, gender, risk factors and symptoms. The calculation of age and risk factors used equation (5), while the proximity of sex and symptoms used equation (4), and the global similarities were calculated by using equation (3).

TABLE I. Case base sample

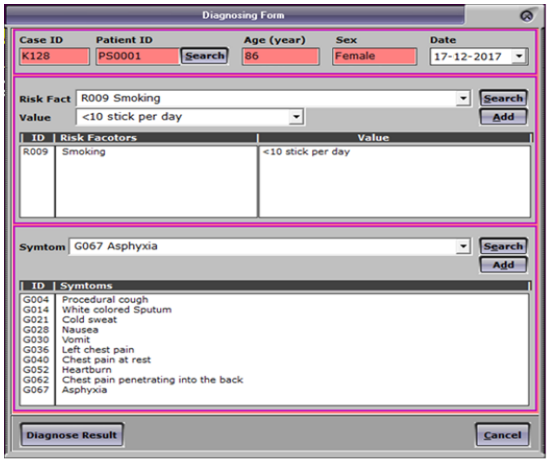
New Case	Description	Value	Weight
Age	86	6	3
Sex	Female	0	2
Risk Fact:			
R009	Smoking < 10 stick per day	1	4
Symtomps:			
G004	Procedural cough	1	3
G014	White colored Sputum	1	2
G021	Cold sweat	1	5
G028	Nausea	1	2
G030	Vomit	1	4
G036	Left chest pain	1	7
G040	Chest pain at rest	1	6
G052	Heartburn	1	7
G062	Chest pain penetrating into the back	1	8
G067	Asphyxia	1	6
Disease I21.1, with 100% confidence level			

a) Local similarities

Age proximity: (Min(5,6))/(Max(5,6)) = 5/6 = 0,83.

Risk Factor proximity: R004 =(Min(0,1))/(Max(0,1)) = 0/1 = 0 and R009 =(Min(1,2))/(Max(1,2)) = 1/2 = 0,5.

Gender proximity = 0, due to the difference of old case and a new case. Symptoms proximity: Symptoms G004, G014, G021, G028, G030, G036, G040, G052, G062 are valued 1 because both cases have the symptoms. Symptom G062 is valued 0 because only old case (case base) has the symptom.



Case ID	Patient ID	Age (year)	Sex	Date
K128	PS0001	86	Female	17-12-2017

Risk Fact: R009 Smoking
Value: <10 stick per day

ID	Risk Facotors	Value
R009	Smoking	<10 stick per day

Symtom: G067 Asphyxia

ID	Symtoms
G004	Procedural cough
G014	White colored Sputum
G021	Cold sweat
G028	Nausea
G030	Vomit
G036	Left chest pain
G040	Chest pain at rest
G052	Heartburn
G062	Chest pain penetrating into the back
G067	Asphyxia

Fig 1. Diagnosis Process into the System.

1) Global Similarities: Based on the calculation of local similarities explained above, global similarities were calculated by using equation (3).

$$\mathrm{E}(\mathrm{C_i}, \mathrm{C_j}) = \left[\frac{\begin{array}{c}(3*0{,}83)^3 + 0 + (4*0{,}5)^3 + \\ (3*1)^3 + \\ (2*1)^3 + (5*1)^3 + \\ (2*1)^3 + (4*1)^3 + \\ (7*1)^3 + (6*1)^3 + (7*1)^3 + \\ (8*1)^3 + 0\end{array}}{\begin{array}{c}(3)^3 + (2)^3 + (4)^3 + (3)^3 + \\ (2)^3 + (5)^3 + (2)^3 + (4)^3 + (7)^3 + \\ (6)^3 + (7)^3 + (8)^3 + (6)^3\end{array}}\right]^{1/3} * 100 * \frac{11}{13}$$

The similarity value of $\mathrm{E}(\mathrm{C_i}, \mathrm{C_j})$ was 0,80 obtained as the results of the calculation of the proximity of both cases described above. The system would display the results in the form Diagnosis Results. The output of the system was the name of the disease that the patient suffers from the highest level of similarity. The similarity value was between 0 and 1 and in a percentage form. The diagnose results would be retained into the system.

C. Case Revision Process

The process of case revision needs to be carried out if the system is not able to diagnose the disease correctly. The system is considered unsuccessful to diagnose the disease if the value of similarity is less than 0,8 (80%). Revision process can only be conducted by the user with access right as an expert.

D. System Assessment Process

The system assessment process was conducted in two ways namely the test conducted by experts and the test using a random data sample that was 30% of 126 cases or as many as 38 data. For test purposes, 10 cases were added as test data which was not a case of disease I21. The additional data were cases of heart failure (I50). The result of the assessment conducted by experts by using data samples based on their knowledge, and then matched the results of the system with expert conclusions on the data tested.

System assessment was also performed by using medical record data. The value of experts' trust toward a case in the medical record data was 100% because it had been through a thorough observation and assessment. The assessment was conducted by using the data of case I21 and I50. The result of the system assessment using the data of case I21 showed that the highest similarity value was 0,99 and the lowest was 0,68. The average test result of the data of case I21 showed that the value was above the threshold of 95%. The test results using the data of case I50 showed that the highest similarity value was 0,77 and the lowest was 0,40 with the average value above the threshold of 100%.

The evaluation of the results of system assessment in diagnosing cardiovascular disease was conducted by calculating sensitivity, specificity, PPV, NPV, accuracy and error rate. Evaluation is important to determine whether the system built is feasible to be applied in diagnosing cardiovascular disease, especially for the type I21. The first stage to be done was to create a confusion matrix based on the value of similarity as the results of the system assessment, as shown in Table 2.

TABLE II. CONFUSION MATRIX OF SYSTEM TEST

	System Test			
		Case I.21	Case I.50	All Case
System test result	Similarity ≥ 0.8 (+)	36 (TP)	0 (FN)	36 (P)
	Similarity < 0.8 (-)	2 (FP)	10 (TN)	12 (N)
	All Case	38	10	48 (P+N)

Confusion matrix shows 36 cases of disease I21 with similarity values ≥ 0,8 and 2 cases with similarity values <0,8 that is on test 21 and 34 with similarity value of 0,79 and 0,68 respectively. The result obtained from the test using the case I50 is entirely <0,8. So that the level of a sensitivity, specificity, PPV, NPV, accuracy and error can be calculated by using equations (6), (7), (8), (9), (10) and equation (11).

$$\text{Sensitivity} = \frac{36}{(36+0)} \text{x} 100\% = 100\%$$

$$\text{Specificity} = \frac{10}{(10+2)} \text{x} 100\% = 83{,}33\%$$

$$\text{PPV} = \frac{36}{(36+2)} \text{x} 100\% = 94{,}74\%$$

$$\text{NPV} = \frac{10}{(10+0)} \text{x} 100\% = 100\%$$

$$\text{Accuracy} = 100\% \text{ x} \left(\frac{36}{(36+12)}\right) + 83{,}33\% \text{ x} \left(\frac{12}{(36+12)}\right) = 95{,}83\%$$

$$\text{Error rate} = \frac{(2+0)}{(36+12)} \text{x } 100\% = 4{,}17\%$$

The above calculation shows the percentage of the system's ability to recognize disease I21 correctly is 100% (sensitivity), the percentage of system ability to recognize disease which is not I21 correctly was 83,33% (specificity), positive predictive value was 94,74% (PPV), negative predictive value was 100% (NPV), And the accuracy was 95,83% with an error rate of 4,17%.

V. Conclusion

Based on the research and the results of the tests that have been conducted, it can be concluded that this study resulted in a case-based reasoning system with weighted Minkowski similarity calculation method used to perform early diagnosis of cardiovascular disease. This system performed a diagnostic process by taking into account the proximity between the case base and the target case based on the patient's condition (symptoms and risk factors), sex and age. The test results of the system for early diagnosis of cardiovascular disease using medical records of patients with disease I21 (based on case basis) and medical records of patients with I50 disease (not in accordance with the case basis), indicated that the system was able to recognize the disease I21 correctly (Sensitivity) of 100%, recognize non-I21 disease (specificity) of 83,33% with accuracy of 95,83% and error rate of 4,17%.

Acknowledgment

The research presented in this paper for conducted on the participation of various parties. Thank you for Dr Sardjito Public Hospital in Yogyakarta and Dr Irsad Andi Arso, M.Sc., Sp.PD., Sp.JP (K), who has been willing to be the resource for this research. We also expressed his gratitude to those who work in medical records installation of Dr Sardjito Public Hospital in Yogyakarta for the provision of medical record data as supports for this research.

References

[1] S. Wang, M. Petzold, J. Cao, Y. Zhang, and W. Wang, “Direct Medical Costs of Hospitalizations for Cardiovascular Trends and Projections,” Medicine (Baltimore)., vol. 94, no. 20, pp. 1–8, 2015.

[2] A. Labellapansa, A. Yulianti, and A. C. Reasoning, “Lambda Value Analysis on Weighted Minkowski Distance Model in CBR of Schizophrenia Type Diagnosis,” in 4th International Conference on Information and Communication Technologies, 2016, vol. 4, no. c, pp. 1–4.

[3] S. H. El-sappagh and M. Elmogy, “Case Based Reasoning : Case Representation Methodologies,” Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 11, pp. 192–208, 2015.

[4] R. Carbó-dorca, “Molecular quantum similarity measures in Minkowski metric vector semispaces,” J. Math. Chem., vol. 44, no. 3, pp. 628–636, 2008.

[5] M. A. E. Abbas, “Anticipating of Cardiovascular Heart Diseases using Computer based Poly Trees Model,” Int. J. Comput. Appl., vol. 59, no. 13, pp. 19–27, 2012.

[6] K. C. Shekar, M. Tech, D. Ph, and M. Tech, “Improved Algorithm for Prediction of Heart Disease Using Case based Reasoning Technique on Non-Binary Datasets,” Int. J. Res. Comput. Commun. Technol., vol. 1, no. 7, pp. 420–424, 2012.

[7] A. M. Salem, M. Roushdy, and R. A. Hodhod, “A Case Based ExpertSystem for Supporting Diagnosis of Heart Diseases,” AIML J., vol. 5, no. 1, pp. 33–39, 2005.

[8] P. Pal and S. Tomar, “A Medical Multimedia based Clinical Decision Support System for Operational Chronic Lung Diseases Diagnosis and Training,” Int. J. Comput. Appl., vol. 49, no. 8, pp. 1–12, 2012.

[9] A. Cardiology, “Case-Based Reasoning for Diagnosis Heart Failure in,” Austin Cardiol, vol. 1, no. 1, pp. 1–5, 2016.

[10] M. K. Rao, K. V. Swamy, K. Anithasheela, and B. C. Mohan, “Face recognition using different local features with different distance techniques,” Int. J. Comput. Sci. Eng. Inf. Technol., vol. 2, no. 1, pp. 67–74, 2012.

[11] B. M. Elbagoury, A. M. Salem, and M. Roushdy, “A Hybrid Case-Based and Content-Based Retrieval Engine for Mobile Cancer Management System - MCMS A Hybrid Case-Based and Content-Based Retrieval Engine for Mobile Cancer Management System - MCMS,” Int. J. Bio-Medical Informatics e-Health, vol. 1, no. May 2014, pp. 15–19, 2013.

[12] Y. Liu, S. L. Chen, A. M. Yen, and H. Chen, “Individual risk prediction model for incident cardiovascular disease : A Bayesian clinical reasoning approach,” Int. J. Cardiol., vol. 167, no. 5, pp. 2008–2012, 2013.

[13] R. Cordeiro, D. Amorim, A. Shestakov, B. Mirkin, and V. Makarenkov, “The Minkowski central partition as a pointer to a suitable distance exponent and consensus partitioning,” vol. 67, pp. 62–72, 2017.

[14] R. Cordeiro, D. Amorim, and B. Mirkin, “Minkowski metric , feature weighting and anomalous cluster initializing in K-Means clustering,” Pattern Recognit., vol. 45, no. 3, pp. 1061–1075, 2012.

[15] C. Yang, X. Yu, Y. Liu, Y. Nie, and Y. Wang, “Neurocomputing Collaborative fi ltering with weighted opinion aspects,” Neurocomputing, vol. 210, pp. 185–196, 2016.

[16] A. K. Akobeng, “Understanding diagnostic tests 1 : sensitivity, specificity and predictive values,” Acta Pædiatrica, vol. 96, pp. 338–341, 2007.

BYOD Implementation Factors in Schools: A Case Study in Malaysia

Yusri Hakim bin Yeop[1], Zulaiha Ali Othman[2], Siti Norul Huda Sheikh Abdullah [3], Umi Asma' Mokhtar[4], Wan Fariza Paizi Fauzi[5]
Center for Cyber Security, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*—The Bring Your Own Device (BYOD) initiative has been implemented widely in developed countries as a mechanism to prepare the students for the 4th industrial revolution. Success stories of the initiative vary depending on factors pertaining to its implementation. This study aims to identify the key factors to implement BYOD in schools for educational purposes. The research employed a mix-method approach by means of a survey. The data was collected from teachers through questionnaires and and from the school management through interviews. The respondents included 204 teachers from 5 schools in Putrajaya and Dengkil. Principals, senior assistants, ICT teachers and technicians from three schools were interviewed. They represented the school management group. A descriptive statistical analysis is conducted using SPSS statistical software. The research has identified four key factors for the successful implementation of BYOD in Malaysian schools. Two of the factors are related to the Cyber Security Policy at the schools - enforcing a secure network infrastructure and safety control requirement in the implementation of BYOD at schools. These security-related factors are important for the schools from the very beginning. They can be further categorized according to the implementation stages: pre-, during and post-adoption; cost allocation, preparation of controls and training to support BYOD's implementation at schools are the corresponding factors to each stage. On the other hand, the other two key factors are related to the schools' readiness - ensuring the successful implementation of BYOD whereby the school management group is willing and prepared to tackle any arising BYOD-related issues in the future.**

Keywords*—*Cybersecurity awareness; cybersecurity education; safety; school cybersecurity policy

I. Introduction

The Information and Communication Technology (ICT), particularly the Internet has changed our lifestyle in various ways namely in communication, education, business, and governance. One of the impacts of ICT can be seen through the rapid changes of the education system, from conventional to modern system. The conventional system is believed to be old- fashion that needs more innovations in order to keep up with the technology. It is believed that with the 21st century technology, schools have more opportunities and ability to shape the education system, and no longer dependent on the poorly funded-educational system or limitation of resources. According to [1], the educational technology is a systematic and organized process of applying modern technology to improve the quality of education, for example, learning and teaching and help with the application of modern educational teaching techniques. The domain of educational technology is apparent where technology is used as a teaching and learning tool and as a tutor. However, the educational technology is still not implemented sufficiently. One of the reasons is due to the lack of school equipment. Hence, many countries have initiated the Bring Your Own Device (BYOD) to school program to overcome the issue, and at the same time, to embrace the technology for promoting a better education system.

II. Bring your Own Device

BYOD is defined as the practice of allowing ones to bring their own personal devices to the workplace for the work purposes [2]. In this study, BYOD is referred to as a concept that allows the students to bring personal devices such as laptops or tablets for better learning interaction and experience [3].

A. The Implementation of BYOD in European Countries

In European countries, the ownership and type of devices that can be brought to school to facilitate the school to manage the usable devices and applications have been determined in the last five years [4]. Table 1 tabulates findings from the BYOD implementation in nine European countries, covering various factors such as infrastructure, security, application, knowledge, constraints, benefit, and method of use. Each country has its own mechanism on how to implement BYOD in schools, which is also dependent on the country's policy. For example, in Estonia, the focus is more on the application factor as compared to other countries. The applications used are Showbie, Socrative and Padlet, whereas, other countries like Norway, Portugal, United Kingdom, Austria, Finland, and Switzerland do not specify the applications for student's learning. All of the countries acknowledged that the factors i.e infrastructure, security and knowledge are deemed important to be considered when implementing BYOD.

The are several constraints i.e. slow Internet connection, lack of ICT knowledge, device's security, costly devices, limited access, device's specification and data package as well as health and social issues, which need to be addressed if the government wants to continue implementing BYOD to schools. Since students are explosed to the gadgets and technology at the young age, a control mechnisme must be imposed as practised in Finland, Ireland and Switzerland to prevent addiction and negative influences.

We would like to thank the Universiti Kebangsaan Malaysia for financial support through a research grant: DCP-2017-015/4.

TABLE I. SUMMARY OF BYOD IMPLEMENTATIONS IN THE EUROPEAN COUNTRIES [4]

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Infrastructure: **1. Type of Device; 2. Device Ownership/Financing; 3. Internet Network**									
1.1 Notebook	/		/		/	/	/		/
1.2 Tablet/Ipad	/	/	/	/	/	/	/	/	/
1.3 Smartphone	/	/	/	/	/	/	/		/
2.1 Parents own the device	/	/		/	/	/	/	/	/
2.2 School provides to students	/					/	/		
2.3 Government subsidized			/						
3.1 School/Government provides the WiFi	/	/		/		/	/		
3.2 Access Point (WiFi in class)		/	/			/	/	/	
3.3 Broadband Network Capacity up to 100Mbps		/	/			/			/
3.4 Improving schools' network (sponsorship/government)						/			
3.5 Technical Assistant provided					/		/		/
3.6 External vendor assistance for technical services		/	/					/	
3.7 Provides a charging station and a personal safe			/				/		
Security									
Firewall usage		/			/				
Mobile device management									/
Specifications of student personal devices are determined by the school						/		/	
No specifications for student personal devices	/	/	/	/	/		/		/
Student's devices inspection						/			
Application									
Office 365 (Cloud)		/							
Showbie (school paperless apps)	/								
Socrative (quiz and training apps)	/								
Padlet (digital whiteboard)	/								
Edmodo (school paperless apps)								/	
Knowledge									
Teachers are qualified and technology savvy	/	/	/		/	/	/		
Trained students can help other students in device problem			/		/		/	/	/
Provides ICT training and Internet safety for teachers	/		/	/	/		/	/	/
ICT use of pedagogical exercise to teachers			/		/		/	/	/
Constraints									
Slow Internet connection	/			/				/	
Lack of ICT knowledge among teachers				/					
Effect of the networks filter (i.e. FB, YouTube, and online games) causing limited access to the apps				/					
Use private data package due to poor school WiFi	/								
Parental worries about device security	/								
Students device is damaged or forgot to carry the device		/							
Low specs devices and interferes with learning Health problem									/
Expensive device	/								
Government charges Internet but it is not enough to buy the right device	/								
Various types of devices limiting helps from the teachers		/							
Teacher's time constraint to attend the training		/							
Benefit									
60% students use smartphones in schools							/		
Students received information faster than teachers							/	/	
Teacher used no books in learning							/		
School reputation increases	/								
Students are more interested in self-learning	/	/							
Save times	/								
Method of Use									
BYOD use in 1-2 times a week							/		
Used in Science and Language classes							/		/
Used in groups or both to be more effective	/							/	
Start with new students enrolling in year 1, 2 and 3 respectively								/	

1-Estonia, 2-Dublin, 3-Norway, 4-Portugal, 5-United Kingdom, 6-Austria, 7-Finland,8-Ireland, 9 Switzerland

B. BYOD in Malaysia

In Malaysia, the Ministry of Education (MOE) has realized the importance of the Internet and spent over 6 billion Ringgit Malaysia for the development of ICT in education, such as Smart School [5]. The basis of learning in schools is writing, reading and counting (3M), but it is no longer sufficient for the 21st century education, which requires the 21st century skills such as critical thinking and problem-solving skills. According to [6][7], students need to be equipped with the 21st century skills to ensure their competitiveness in the globalization era. It comprised of four domains namely digital age literacy, inventive thinking, effective communication and high productivity.

Hence, MOE had transformed the existing national education curriculum to develop students who are balanced, resilient, curious, principled, informed, caring, patriotic, creative, able to think critically, communicate and work effectively in a team [8]. MOE wanted to ensure the students are capable to compete at the international level. On January 13, 2018, Malaysian education ministers have announced that 2000 classes in selected schools will be equipped with 21st Century Learning (PAK-21) classes [9], especially using project-based approaches [8]. Technology is an important component of the PAK-21 classes and this posed a challenge for MOE because of the high cost needed to provide adequate, up-to-date and secure computer labs for the schools. In addition, it is evident that the existing school network (1BestariNet) needs to be upgraded to support this effort.

MOE has recommended BYOD for Teaching and Facilitation (T&F) and for other activities supporting the implementation of the curriculum and co-curriculum. On March 29, 2018, MOE issued a memorandum outlining the BYOD policy for schools and guidelines for the implementation of BYOD at schools [10]. The BYOD policy listed the permitted types of devices which include laptops, tablets, audio devices or any other devices except for smart phones. These devices may be owned by parents or guardians.

Devices can also be contributed by the Parents and Teachers Association (PTA), schools, organizations or individuals. The BYOD policy also described several device control and security measures such as device registration, periodic checking, reporting in case of loss, secure Internet and Wi-Fi access, safe storage and charging stations in the classrooms.

Various issues need to be considered before implementing BYOD in Malaysian education system specifically teachers' knowledge [11], social problems [12], infrastructure [13], security [13], health issues [14], financial capability [14], etc. A teacher's technical knowledge on ICT and safety-related knowledge when engaging in the cyberspace is important to help the school to successfully implement BYOD, as stated by [9] that teachers are still having problem in dealing with new technology. If teachers are technically skillful, then they can assist the schools to improve the quality of learning in Malaysia with the implementation of BYOD.

This study investigated the level of knowledge and skills of teachers in terms of ICT, cyber security and Malaysian cyber laws as well as the existing school infrastructure. This study also investigated the level of readiness to implement the BYOD and categorized the factors based on the three categories which will be further discussed in this paper.

C. *Teaching and Facilitation*

Student and teacher have a mixed learning styles and teaching styles preferences respectively, which means it is important for teachers to adopt appropriate instructional strategies [15]. This is due to technology advancements which has changed the traditional learning and teaching styles causing diversity methods in education system. One of the examples to accommodate the changes is through self-regulated learners by embracing the Education 3.0 that using technology as a tool in teaching and learning [16]. The Education 3.0 is more e-learning driven that offers open and flexible learning activities in order to create room for student creativity and social networking outside traditional boundaries. Hence, the Teaching and Facilitation approach is suitable in the 21st Century Learning.

III. Methodology

This research employed a mixed method approach [17-19] that consists of three phases.

Phase 1 is a literature study and statement of research problems. Phase 2 is the phase of studying the possible factors affecting the implementation of BYOD at schools by referring to the key elements in [5]. The data are collected through three techniques: literature review, a closed forum on the cyber issues in Malaysia and interviews of personnel from involved agencies across the ICT and education government sectors. The questionnaires are distributed to teachers in schools around Putrajaya and Dengkil. While interviews were conducted with school administrators, ICT teachers and school technical staff from schools in Putrajaya and Dengkil to further strengthen the findings from the questionnaires.

Lastly, Phase 3 is data analysis using statistical analysis to find put the factors that influence and the level of readiness to implement BYOD.

A. *Questionnaire and Interview*

The questionnaire was divided into five components namely (A) General profile of the respondent, (B) School infrastructure availability and readiness of school implement BYOD in their school, (C) Preparation of security controls safely, (D) Level of teacher knowledge and readiness applying BYOD and (E) Teacher's opinion(s) on health issues if BYOD is implemented. All of the questions use the Likert scale method with measurements of 1 to 5 where the answer scale used is 1 = Strongly Disagree, 2 = Disagree, 3 = Simple Agree, 4 = Agree, 5 = Strongly Agree.

A total of 4 interview sessions were conducted in a period of one month in 3 schools. Each interview session was between one to two hours.

B. *Closed Forum*

A closed-session discussion entitled "Cybersecurity: Towards a Secure and Sustainable Cyber Use" was held at UKM with invited participants from the government sectors i.e. Royal Malaysia Police, Ministry of Women, Family and Community Development, Ministry of Education, Ministry of Communications and Multimedia Commission, Law faculty of Universiti Kebangsaan Malaysia and the National Occupational Safety and Health Institute. The forum discussed the government's efforts to direct/foster Malaysians towards a safe and healthy/beneficial cyber culture. Each department's or agency's representative to share his/her views regarding the proposed title for 40 minutes followed by a 10-minute session for questions and answers.

C. *Respondents, Sampling, Distribution and Fieldwork*

Within a two-week time frame, a total of 300 sets of questionnaires were distributed to five government primary and secondary schools in Putrajaya and Dengkil. A total of 222 sets were returned of which 18 sets were rejected because the forms were incomplete and contained dubious answers. Hence, only 204 sets, which is 68% of the total set distributed, can be analyzed using SPSS software.

TABLE II. Infrastructure

No	Item	L	F	CSM	BTP
1	Expansion of school internet access and access control	/	/		/
2	Provision of Locker to save device		/		
3	Technical requirements	/			/
4	Class changes to suit PAK-21	/			
5	1 device / high school student	/			
6	Ease of charging in class and preparation of power sockets	/			
7	BYOD Model Type	/	/		
8	Wi-Fi capability	/			
9	Use of MDM or Content Filtering	/			/
10	Ability of broadband available	/			
11	Equipped with anti-virus applications	/			

L-Literature, F-Forum, CSM-Cyber Security Malaysia, BTP – Education Technology Division

IV. Data Analysis

This section discusses the findings of the questionnaires and interviews on the key factors of BYOD Implementation in the local government schools to establish a school preparation model for the implementation of BYOD. This section is divided by two parts: (A) Findings from the Questionnaires and (B) Findings from the Interview.

A. First Part: Findings from the Questionnaire

1) Infrastructure:-There are four factors namely Infrastructure, Health, Safety and Knowledge to implement BYOD in schools. Each factor consists of several elements. In Table II, all the selected elements that contributed to the Infrastructure factor are listed. It was generally agreed that 'expansion of school internet access and access control' is one of the important elements. The technical requirements, lockers as a place to keep the devices, 1 device 1 student, charging station, WiFi, and more are also important to implement the BYOD.

The readiness level is identified based on the mean value, which is interpreted as follows:

- 4.00 – 5.00: High
- 3.00 – 3.99: Medium
- 2.00 – 2.99: Low
- 1.00 – 1.99: Very low
- 0.00 – 0.90: Not available

Table III shows the analysis result of readiness of the school facilities of the school to implement BYOD. The needs of ICT technicians are high. Also, other criteria such as internet connection, power sockets, tables, chairs, safety locker are still insufficient. Respondents agreed that the school should provide devices/mechanism for content filtering or restrictions on internet materials that are not suitable for students in the school's Wi-Fi network. Hence, the level of respondent's readiness on the content filtering is at a medium level. Overall, it can be concluded that school infrastructure is not yet ready for implementing BYOD.

TABLE III. Readiness of School Facilities for BYOD

Topic	**Mean**	**Std. Deviation**	**Readiness Level**
Internet Connection	2.76	1.111	Low
Wi-Fi	2.55	1.28	Low
Content Filtering	3.11	1.2	Medium
Power socket in Classroom	2.44	1.167	Low
Tables and Chairs	2.83	1.252	Low
Safety Locker	2.81	1.402	Low
The need of ICT Technician	4.5	0.833	High

2) Health:- Table IV indicates that Health factor is one of the factors to be considered when implementing the BYOD. The common health issues include eye sight problems either far/near sighted, Internet addiction, loss of focus due to lack of sleep; and the need to have a control mechanisme at home to prevent problems in school.

TABLE IV. Health

No	**Item**	**L**	**F**	**CSM**	**BTP**
1	The risk of farsightedness, the time to look over the screen	/			
2	Internet addiction, attitude change	/		/	
3	Failing to complete schoolwork, lost learning focus	/			
4	Control the use of internet at home to avoid problems in school			/	

L-Literature, F-Forum, CSM-Cyber Security Malaysia, BTP – Education Technology Division

Table V shows the impact of implementing BYOD on students' health and attitude. The level of risk pertaining to the health issues i.e. gadget addiction and schoolwork disruption are high. Since the students are young users, the monitoring and controlling mechanism need to be imposed to avoid severe repercussions.

TABLE V. Health's Risks Among Students

Topic	**Mean**	**Std. Deviation**	**Level**
Blindness due to excessive use of gadgets	3.73	1.098	Medium
Tired, sleepy or lost focus in the classroom	3.78	0.978	Medium
Failed to complete the schoolwork	4.03	0.898	High
Different attitude when at home and at school due to internet addiction.	3.82	0.892	Medium
Uncontrolled gadgets and internet	4.08	0.78	High

Respondents agreed that the uncontrolled use of gadgets affected students' performance and was the cause of students failing to complete their homework. Internet addiction or excessive use of gadgets can change student attitudes. Respondents also agreed that their students behaved differently at home and at school. Students were either quiet in school but too active at home or vice versa. Internet addiction was identified to be the cause of the change of attitude and could disrupt the student's social relationships with friends or family.

3) Safety:- The elements related to safety factor are listed in the Table VI. The respondents agreed that student safety is important in both aspects: the data and devices.

TABLE VI. SAFETY

No	Item	L	F	CSM	BTP
1	Comply to National Security Policy (RAKKSSA) for safety reason including environment, technology, human, threats, and audit	/			
2	Understand the related laws i.e. Cyber law and cyber threats	/			
3	Student safety: personal data and devices		/	/	/
4	Introduce safety procedures		/		

L-Literature, F-Forum, CSM-Cyber Security Malaysia, BTP – Education Technology Division

Table VII shows the results concerning safety control at school before implementing BYOD. The table shows that all four elements under the safety factor are at medium and high levels. Respondents affirmed that the school should determine the specifications and types of devices that are allowed and these devices have to be registered and reviewed by the schools before they can be allowed to connect to the school's network. The needs for device registration, anti-virus and good school's WiFi network are high to enable the school's management to provide a safe and secure environment and to control student's online activities and use the devices for educational purposes only.

TABLE VII. THE NEED OF SAFETY CONTROL

Question	Mean	Std. Deviation	Needs level
Device spec. determined by school	3.84	1.094	Medium
Device registration	4.06	1.126	High
Guarded with Anti-virus	4.32	0.771	High
School Wi-Fi only	4.21	1.122	High

4) Knowledge:- Table VIII indicates that knowledge of cyber law and cyber threats are important and syllabus must be created for the cyber space curriculum.

TABLE VIII. KNOWLEDGE

No	Item	L	F	CSM	BTP
1	Knowledge of cyber law and cyber threats	/		/	
2	Creating a syllabus for Cyberspace curriculum for high school students	/			

L-Literature, F-Forum, CSM-Cyber Security Malaysia, BTP – Education Technology Division

Table IX shows the result of teachers' knowledge on the Government Initiatives on Cyber Security and Cyber Law. The teacher's level of awareness on cyber programs, skills in ICT, and skills in cyber security are still at the medium level. Nonetheless, the teachers realized that materials in the Internet can be used as a learning tool.

TABLE IX. LEVEL OF TEACHER'S KNOWLEDGE

Question	Mean	Std. Deviation	Level
Knows about awareness programs related to cybercrime and the cyber ethics	3.89	0.724	Medium
Materials in the Internet can be used as a learning tool	4.42	0.665	High
Skilled teacher who can handles laptops, tablets or iPad	3.84	0.961	Medium
Skilled teacher in cyber security knowledge and its threats	3.59	0.991	Medium

B. Second Part: Findings from the Interview

Three out of the seven schools that have been approached, agreed to be interviewed: School A, School B and School C. Additional interviews were conducted at the Education Technology Division, MOE to gather more information.

Based on the analysis on the feedback from school administrators on the school's readiness to implement BYOD for Teaching and Facilitation, various issues such as device security, the increased in the teacher workload, inferiority complex among students, parental financial abilities and social media contagion are factors that can undermine BYOD's implementation. All these issues needed further investigation. In terms of infrastructure and facilities, the number of power sockets in the classrooms were inadequate, some of which were not well-maintained. Overall, it can be concluded that the readiness of infrastructure at these schools are still at a low level.

The school management group were also concerned about the safekeeping of personal devices. Students did not have their own lockers. Other issues highlighted include the current tables and chairs were not suitable for personal device use because of the non-ergonomic designs that will cause neck and back pain due to incorrect sitting posture and inappropriate when using the device; the capability of existing Internet is insufficient; there is not enough ICT assistants to assist schools in technical matters – some schools do not have technical specialists and ICT teachers have to take over the task of maintaining the computer laboratory and repairing damaged computers which impacted the effectiveness of the students' learning experiences.

V. RESULTS AND DISCUSSION

This research has identified four important factors needed to be addressed to achieve the aspirations of MOE's strategic plan in providing virtual education. These four factors are the availability of infrastructure, knowledge of health impacts, safety control requirements and skill level and teacher knowledge. Figure 1 shows the current status of the four factor that the schools need to improve to reach the readiness level to implement BYOD in schools.

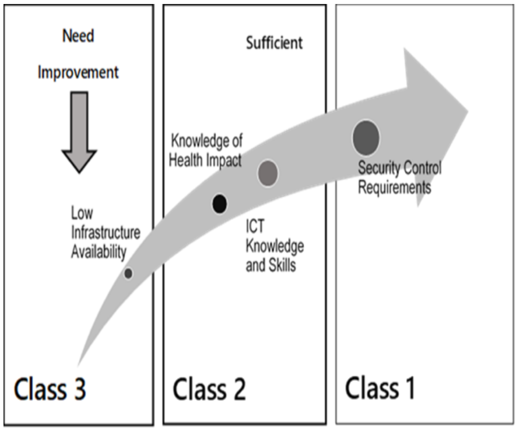


Fig. 1. Readiness Category.

Three Readiness Categories have been identified; essential, sufficient and need improvement. These readiness levels were extracted from the analysis of the respondents' responses. The category of readiness is presented to indicate which factors need more enhancement in order to implement BYOD. The categories are prioritized in ascending order from 1 to 3, which 1 as critically needed (essential), 2 as currently adequate/sufficient and 3 as need improvement.

In Class 1: Essential, majority of the respondents were concerned on the need to create security controls to protect pupils and teachers from cybercrime. Documents such as school's cyber security policy should be designed, produced, enforced and complied with to control the use of personal devices of teachers and pupils. In addition, some forms of control should also be applied to address thefts, damages and health problems.

TABLE X. PROPOSED LIST OF ELEMENTS FOR EACH FACTOR

Factor 1: Availability of the infrastructure	
A	An internet network that can accommodate all school users including students in the same time.
B	The stress test for the network is required
C	The extensive Wi-Fi network coverage also includes every class, garden and school canteen.
D	Stress test against Wi-Fi access point (access point)
E	Preparation and maintenance
F	Content filtering (content filtering) for monitoring internet access
G	Provide adequate power sockets for area charging devices in the classroom
H	Preparation of a locked-up device for storage of personal mobile devices
I	Provision of table chairs in the classroom that suit the use of personal mobile devices
J	Provision of backup devices for the use of students experiencing problems on their
Factor 2: Health effects	
A	Screen time-consuming control. 20 minutes exposed to the screen requires 20 seconds break and look at objects that are positioned 20 meters from the eye (20-20-20)
B	Train pupils proper body posture while using mobile devices
C	Train pupils to assess the bad information or images viewed on the internet
D	Conduct training sessions for parents and students in controlling the use of gadget or home internet
Factor 3: Security control requirements	
A	The permitted mobile device specification is determined by the school
B	Registration must be made on all student mobile devices before they can be used within the school.
C	The carry-in device is already equipped with anti-virus
D	Perform random investigation on devices to ensure they are safe from any intrusion.
E	Procedure to report on theft, loss and abuse.
F	Devices are only allowed access to school internet access only for monitoring
G	Create safety rules for mobile devices for teachers and students
H	Establishing safe internet usage ethics guidelines at school for students, teachers and administrators
Factor 4: Knowledge and skill	
A	Basic training on network operation and specific software
B	Training cyber security skills and threats to teachers and students
C	Duties and responsibilities of technical interpreters
D	Exposure to teachers and students towards cyber security and internet awareness programs organized by the government or private sector

In Class 2: Sufficient refers to teachers having good skills in terms of knowledge of ICT's handling, health effects, internet usage and social media, cybercrimes and cyber awareness campaigns that have been conducted by the government as well as private organizations. The analysis also showed that the school administrators support the use of personal devices to schools for Teaching and Facilitation.

In Class 3: Need Improvement category, the infrastructures and supplies are limited. Key infrastructures such as internet network and Wi-Fi capabilities provided by the government in schools are insufficient so schools have to subscribe to other internet service providers to accommodate the teachers' needs. Only schools in Putrajaya have facilities such as power sockets in the classroom and equipment storage lockers but many are not well-ventilated and most of them are damaged.

Besides the Readiness's Category, this study also proposed four factors which based on the findings. The factors are the basis for the successful implementation of BYOD as shown in Table X.

All of the factors served as a basis that need to be considered before implement the BYOD. For example, the knowledge refers to both practical and theoretical. Teachers who are technology savvy have more skills and knowledge to teach their students using learning applications effectively. Despite using the YouTube for entertainment only, the teachers can guide the students to channel their creativity and learn interactively by watching the videos. Some of the lessons need to be demonstrated rather than explain narratively, which can be fully utilized via the YouTube channel. It is also noted that the knowledgeable teachers have better understanding to use the BYOD appropriately and effectively. A conducive infrastructure would influence to the implementation and can avoid technical disruptions. Health and security factors cannot be neglected particularly when involving young students and cyber environments. Both are crucial due to many social impacts occurring nowadays. Therefore, all the proposed factors have considered the technical, social and knowledge aspects to provide a foundation for better implementation as well as student's well-being.

ACKNOWLEDGMENT

We would like to thank Universiti Kebangsaan Malaysia for financial support through grant DCP-2017-015/4.

REFERENCES

[1] Stošić, Lazar. 2015. The Importance of Educational Technology in Teaching. *International Journal of Cognitive Research in Science, Engineering and Education*. 3(1): 111-114.

[2] Oxford 2010. Oxford reference. Retrieved from http://www.oxfordreference.com

[3] Kementerian Pendidikan Malaysia. 2011. Pelan strategik interim 2011-2020. Ministry of Education. Retrieved from www.moe.gov.my/bppdp.

[4] Attewell, J. 2015. BYOD - A Guide for school leaders. Designing the Future Classroom (3): 1–64. European Schoolnet.

[5] Kementerian Pendidikan Malaysia. 2013. Pelan Pembangunan Pendidikan Malaysia 2013-2015. Kementerian Pendidikan Malaysia, hlm. Vol. 1. doi:10.1016/j.tate.2010.08.007.

[6] Punia Turiman, Jizah Omar, Adzliana Mohd Daud, and Kamisah Osman. 2012. Fostering the 21st Century Skills through Scientific Literacy and Science Process Skills. *Procedia-Social and Behavioral Science*. 59(2012): 110-116.

[7] Radha M K Nambiar, Noorizah Mohd Nor, Kemboja Ismail, and Shahirah Adam. 2017. New Learning Spaces and Transformations in Teacher Pedagogy and Student Learning Behavior in the Language Learning Classroom. *3L: The Southeast Asian Journal of English Language Studies*. 23(4): 29-40.

[8] Romarzila Omar , Zanaton H. Iksan, Sharifah Nor Puteh, 2018, A Comprehensive 21st Century Child Development through Project Based Learning 1, *Journal of Adv Research in Dynamical & Control Systems*, 10 (06-Special Issue): 1636-1642 1636.

[9] Hayati Ismail. 2018. Pembelajaran abad ke-21: Harapan, realiti dan cabaran. Utusan Online. Retrieved from http://www.utusan.com.my/rencana/utama/pembelajaran-abad-ke-21-harapan-realiti-dan-cabaran-1.590819.

[10] Kementerian Pendidikan Malaysia. 2018. GP DASAR MURID MEMBAWA PERANTI.pdf. Kementerian Pendidikan Malaysia.

[11] Pramela Krish and Noraza Ahmad Zabidi , 2007, TEACHERS AND THE NEW ICT CHALLENGES, *School of Language Studies and Linguistics, journal e-bangi*. 2(2) : Januari - Disember 2007.

[12] Fauziah Ahmad, Chang Peng Kee, Normah Mustaffa, Faridah Ibrahim, Wan Amizah Wan Mahmud and Dafrizal. 2012. Information Propagation and the Forces of Social Media in Malaysia, *Asian Social Science Journal*. 8(5):71-76.

[13] Khosraw Salamzada, Zarina Shukur and Marini Abu Bakar. 2014. A framework for cyber security strategy for developing country, Asean-Japan Workshop on Information Science and Technology 2014, FTSM, UKM.

[14] Dicky Wiwittan Toto Ngadiman, Daily Tayok, Salmy Edawaty Yacoob, and Hairunnizam Wahid. 2018. Social Relationship B40 against Purchasing Behaviour nonbasic Needs using Loans and Intention to Increase debt, *International Journal of Academic and Research in Business and Social Sciences*, 8(7): 1102 – 1117.

[15] Lee Mei Ph'ng, Thang Siew Ming, and Radha M.K.Nambiar. 2015. Matching Teaching Styles and Learning Styles: What Happens in the Case of a Mismatch? *Journal of Social Sciences and Humanities*. Special Issue (1):066-076.

[16] Riza Atiq Abdullah O.K Rahmat and Kamisah Osman. 2012. From Traditional to Self-Regulated Learners: UKM Journey towards Education 3.0. *Procedia-Social and Behavioral Science*. 59 (2012): 2-8.

[17] Board, T. (n.d.). Engaged and informed school leaders drive transformation in Ireland 33–35.

[18] Creswell, J. W. & Zhang, W. 2009. The Application of Mixed Methods Designs to Trauma Research. 22(6): 612–621.

[19] Robert V Krejcie; D Morgan. 1970. Determining Sample Size For Research Activities. 607 – 610.

A Schedule Optimization of Ant Colony Optimization to Arrange Scheduling Process at Certainty Variables

Rangga Sidik[1], Mia Fitriawati[2], Syahrul Mauluddin[3], A.Nursikuwagus[4]
Dept. Information System
Universitas Komputer Indonesia
Bandung, 40132, Indonesia

***Abstract*—This research aims to get optimal collision of schedule by using certainty variables. Courses scheduling is conducted by ant colony algorithm. Setting parameters for intensity is bigger than 0, visibility track is bigger than 0, and evaporation of ant track is 0.03. Variables are used such as a number of lecturers, courses, classes, timeslot and time. Performance of ant colony algorithms is measured by how many schedules same time and class collided. Based on executions, with a total of 175 schedules, the average of a cycle is 9 cycles (exactly is 9.2 cycles) and an average of time process is 29.98 seconds. Scheduling, in nine experiments, has an average of time process of 19.99 seconds. Performance of ant colony algorithm is given scheduling process more efficient and predicted schedule collision.**

***Keywords*—Ant colony; optimization; scheduling; process; certainty variables**

I. Introduction

Arranging schedule at university is more complex and difficult for a setting. Scheduling should be completed in time quickly and easily. Occupation of a classroom and lecturer should be avoided a collision. Meanwhile, the arrangement of the schedule should be submitted on time. Arranging a schedule at a university has to obtained by a regulation. Considering of schedule is depend on certainty variables. Regulation is stated to guide for ordering schedule [1][2]. In a digital era, all works want to be done quickly and easily. This cannot be denied considering that technology is growing and all facilities can be realized. In universities, one process that requires a touch of technology is scheduling. This variable is determined by taking into account certain boundaries and constraints. The process of making class schedules, each university has different terms and limits. While the process, accuracy is needed and must be on time. These problems are widely used as research using certainty variables [3].

Universitas Komputer Indonesia is one private university in Indonesia. There are many faculties at Universitas Komputer Indonesia. In managing of schedule, faculty of management has a problem in scheduling arrange. Time, classroom, lecturer activity, and timeslot are occupied as a factor success in managing schedule. Arranging of schedule collided between timeslot and another timeslot. Request for lecturer and lack of the classroom has caused a collision. In order to solve that problem, the research has been completed in many ways to solve. Using Ant Colony Algorithm (ACO), research has made efficient process and time reduction. ACO algorithm is used in much wide research [3]-[6]. Taking ACO in this research is made scheduling no collision and efficient in the process. In the execution of ACO, this research has been supported by many variables. The variables have included in ACO such as the number of lecturers, courses, classes, timeslot and time. Every execution in ACO, the process has always dependent by it variables. Completed research has made minimize collision in every timeslot and included lecture time submission.

II. Materials and Methods

A. Scheduling, Certainty Variables, Process

Refer to [4] has defined the schedule. Scheduling is an activity to manage the ordering job. Scheduling has many used by many fields. Research [2], [5]–[7], have presented about scheduling in ACO process. Research in robotic, traveling salesman problem, job shop scheduling, is introduced by [1], [8], [9].

In this research, we have proposed a term of certain variables. Certainty variables are parameterized in order to complete the schedule. We have used these variables in the ACO process. Considering, using these variables, is dependent on regulation at Universitas Komputer Indonesia. It also, the availability of ACO which can be processed in many variables.

The terminology of a process referred to manage the schedule. We have proposed these term, to ease of activity scheduling term. The process is also referred to as a sequence of a procedure. Process scheduling at Universitas Komputer Indonesia has regulated in optimization scheduling has always been following by regulation stated. It has made different with other scheduling. Every time, management was changed, it means, the process scheduling would be changed too.

B. An Colony Optimization (ACO)

Ant colony optimization is an evolutionary computational technique proposed by Dorigo et al. [10]. ACO is composed of three steps for a cycle: i) explore(), release ants for finding the destination, ii) pheromoneUpdate(), update pheromones the ants travel across the paths, iii) iteration(), comparing paths the ants found and find the best path if a predefined condition is met. To illustrated the ACO process, we have shown in Fig. 1. Fig. 1 is illustrated how pathfinding by ants ACO algorithm is taken from [10].

Directorate of Research and Community Service, Ministry of Research an Higher Education of Indonesia in 2018 Fiscal Year.

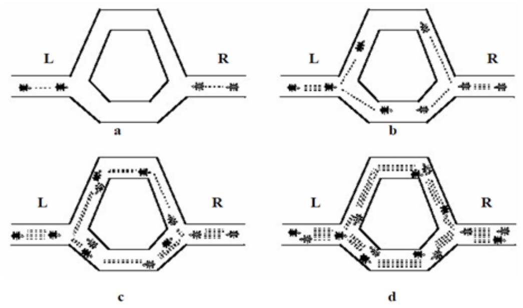


Fig. 1. (A) Ant Track is illustrated in two Ways, Right and Left Side. (B) Ants have Begun Track by two Paths. Each Side has Pictured Nature of Ant on Finding the Path. (C) Ant Sequence has followed the Track of Formerly Path. The Ant Track Previous has followed by other Ants with Pheromone Track. (D) The Figure has depicted the best Solution Path. The best Solution is the Path which has Minimize the Length of a Track. This Path has Chosen by Ant to Move between the Side and Another Side.

We have added some statements to align in scheduling problems. We have followed the step, to find the best solution in scheduling. At the algorithm below, we have adjusted the step to align in scheduling optimization at Universitas Komputer Indonesia.

Step 1

a) Parameters setting, defined as following

- Path intensity of ants and change (τ_{ij})
- City, we have defined as schedule order (*n*=175), including coordinat (*x*,*y*) or distance between timeslot and other timeslot (d_{ij})
- Schedule begin and target schedule (looping)
- Cycle ant constant (*Q*) = 1
- Ant track intensity constant (α) = 0.01, $\alpha \geq 0$
- Visibility constant (β) = 0.01 , $\beta \geq 0$
- Visibility between schedule =

$$1/[d_{ij}\ (\eta_{ij})] \tag{1}$$

- Ants (*m*)
- Evaporation of ant constant (ρ) = 0.03, $0 < \rho < 1$.
- Maximum cycle (*NCMax* = collision = 0) and never change as long as processing, for τ_{ij} always update for each *NC* =1 to *NC* = *NCMax*, or convergence reach.

b) Initialization for first schedule timeslot.

Afterward, τ_{ij} initialization done, and than ant *m* placed in the first schedule randomly.

Step 2

Put the first schedule containing the number of lecturers, courses, classes, timeslot and time in tabu list (temporary for searching schedule). Yields, for the first schedule from the first cycle, have to input as first element in tabu list. After completed in this step is defined indexing schedule by ant. It means variable tabuk1 can be contained indexing from 1 until n as defined in the first cycle

Step 3

Compilation of route for visiting schedule to each ant. An ant colony has distributed into the schedule, will be moved as many as 175 travels from origin schedule into the target schedule. Every ant has preferred schedule which did not in tabu$_k$. In final travel, tabu$_k$ has contained ant colony. If s is indexed visiting and schedule is stated N-tabu$_k$, then visiting probability would be counted as equation following [10]:

$$\mathrm{P}_{\mathrm{ij}}^{\mathrm{k}} = \left[[\tau_{\mathrm{ij}}]^{\alpha}.[\eta_{\mathrm{ij}}]^{\beta}\right]/[\textstyle\sum_{\mathrm{k}^{\mathrm{i}\epsilon\{\mathrm{N-tabu_k}\}}}[\tau_{\mathrm{ik'}}]^{\alpha}.[\eta_{\mathrm{ik'}}]^{\beta}] \tag{2}$$

$P_{ij}^{k} = 0, for\ next\ j$, where *i* is indexed for origin schedule and j is targeted schedule. τ_{ij}(t) means pheromone concentration of the schedule i and j (i,j) at time t. ηij is the heuristic factor that its value can be 1/d$_{ij}$ (d$_{ij}$ means the distance between schedule i and j) and is a constant, and we can see that the schedule of the shorter edge has a great probability to be chosen.

Step 4

a) Measure of route length for every ant.

Measure of length closed tour or L_k for every ant is done after one cycles is completed. Equation for count in one cycle completed as following [1]

$$L_k = d_{tabu_{k(n)},,tabu_{k(1)}} + \textstyle\sum_{s=1}^{n-1} tabu_{k(s),,}\, tabu_{k(s+1)} \tag{3}$$

Where d_{ij} is distance for schedule *i* and *j* that has formulated as following

$$\mathrm{d_{ij}} = \left[\left(\mathrm{x_i}-\mathrm{x_j}\right)^2 + \left(\mathrm{y_i}-\mathrm{y_j}\right)^2\right]^{\frac{1}{2}} \tag{4}$$

b) Searching minimum route.

In sequence of L_k has been counting, we have gained minimum closed length route or $L_{min}NC$ and whole closed length route or L_{min}

c) Updating pheromone in scheduling.

Ant footprint that has owned, would be stated as a track. Total of epavoration and difference ants, it has marked update of intensity. Equation for evaporation can be written as following [1]

$$\Delta\tau_{ij} = \textstyle\sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{5}$$

$\Delta\tau_{ij}^{k}$ can be measured as following [1]

$$\Delta\tau_{ij}^{k} = [Q]/[L_k] \tag{6}$$

For *i* and *j* as origin schedule and target schedule. $\Delta\tau_{ij}^{k} = 0$, for other *i* and *j*.

Step 5

a) Counting intensity of ant footprints for next cycle.

Updating of intensity ant footprint has been caused by evaporation and differences of ant total. Updating value of for update global pheromone ant footprint has formulated as following [1]:

$$\tau_{ij} = \rho.\tau_{ij} + \Delta\tau_{ij} \quad (7)$$

b) Reordering intensity for ant footprint.

On the next cycle, we need to order intensity and than getting intensity value is 0.

Step 6

Empty of tabu list, repeat step 2. Tabu list of a temporary list has occupied into a tabu list for the new cycle. Every cycle, ant colony optimization, must empty the tabu list. It can occur if the sum of a cycle does not reach or does not convergence. The algorithm will be repeated at step 2 with the new footprint intensity. We have defined an equation for collision and bound of iteration as following

$$collision= \sum \text{timeslot } (l,d,t) + \sum \text{timeslot } (c,d,t) \quad (8)$$

Timeslot(l,d,t) and *timeslot(c,d,t)* is defined a value in schedule. Variable *l* means lecturer, *c* means classroom, *d* meand day, and *t* means time.

III. Result

A. Setting Variables Contsraint in Optimization Scheduling

In engineering knowledge for processing variables in ACO, our research has applied some assumptions. These assumptions have a pattern for the need of schedule and then justify the result. We are proposed some assumptions as following:

- Ant termonology is a variable combine among lecturer, courses, and classroom.
- Food resource is a pair of classroom, day, and time.
- Pheromone is a schedule that acquire by lecturer.
- Distance is a count from difference value among classroom, day, and time.
- Setting for first scheduling is activated by staff.
- Limitation of process is just implemented for theory subject and just in classroom.
- Setting day is just for Monday until Satuday with six days in total.
- Every schedule, result from ACO, can be implemented on three hours class.
- Limitation of occupancy in one day is only implemeted four timeslot and three hours class.
- Expecially for Friday, the third timeslot does not occupied.
- On Saturday, amount of timeslot is only three.
- Population ant colony is formulated as following

$$\text{Population=c x d x t} \quad (8)$$

c,d, and *t* are defined as paramater population like classroom, day, and time.

- Constraints of scheduling process as following

a) Lecturer is not permitted acquire in two schedules at the same day and time.

b) Scheduling do not process in two slot at the same day and time.

B. Ant Colony Optimization in Execution

In this section, how ACO is used by scheduling problem. In section 2, we have proposed six steps for completing the job. At the first step, we have shown at section 2.a for setting parameter. The continuous process, the next step, we can be tailored following procedural therein.

Step 1a

The result has shown at section 2, at step 1a. We continue to next step 1b.

Step 1b

Initialization first scheduling, we have constructed in randomize manner. Every variable, it has to transform with generating codification. A single value, we have preferred to code every variable. Afterward, the process continues to generate every single value into a tabular model. Combination table can be seen in table 1.

TABLE I. EXAMPLE GENERATE CODE INTO TABLE REPRESENTATIVE

Schedule	Lecturer	Course	Section	Classroom	day	time
J1	1	1	5	2	4	1
J2	4	3	4	1	2	2
J3	3	2	3	2	4	1
J4	1	4	2	1	2	2
J5	2	5	1	3	2	2
J6	5	1	2	1	1	3

At table 1, abbreviation J1 has meant a schedule at first timeslot. Single value, e.g 1,2, and 3, as codefication from variable that shown by column header.

Step 2

The process is started with implement randomize value into a table that called tabu list. Result at tabu list is the same as table 1. Difference both of table is, a tabu list is used for the transaction as long as the ACO process and the primary table is used for saving schedule from the ACO process.

Step 3

Compilation of route for visiting schedule to each ant. We have applied (2). Every timeslot can be calculated with using a measure of distance between timeslot and other timeslots. The result of this applied is probability visitation of track on every timeslot. Equation 2 has been applied into Table 1 and retrieved to count another equation. The result can be seen in Table 2.

TABLE II. PROBABILITY VISIT (P_{ij}) TAILORED BY (2)

Schedule	*ηij J1*	*τij(t=0)*	$[\tau_{ij}]^{\alpha}.[\eta_{ij}]^{\beta}$	*Pij*	Cum.Prob
J1	0.00	0.01	0.0000	0.000	0.000
J2	0.22	0.01	0.0022	0.191	0.191
J3	0.50	0.01	0.0050	0.427	0.618
J4	0.00	0.01	0.0000	0.000	0.618
J5	0.29	0.01	0.0029	0.247	0.865
J6	0.16	0.01	0.0016	0.135	1

At Table 2 is shown about the result from applied (2). η_{ij} J1 is meant to factor heuristics at i and j times to timeslot schedule at J1. Concentration of pheromone at i and j, $\tau_{ij}(t=0)$ start from 0.01 at beginning of time (t=0). P_{ij} is symbolized for the probability of track visit.

Step 4a

Measure of route length for every ant. Arranging the route of each ant's visit to all points. At this stage, each schedule is attached to all schedules to calculate the distance. The distance between schedule can be calculated using (4). Example to calculate a distance between schedule and other schedules, we have used ordering pairwise among lecturer, classroom, day, and time. E.g J1 , lecturer is 1, classroom 2, day = 4, time = 1. If these values enter to table tabuk list, then distance will be calculated. The result on Step 4a can be seen at Table 3.

TABLE III. DISTANCE BETWEEN SCHEDULE USING (3) AND (4)

Schedule	**J1**	**J2**	**J3**	**J4**	**J5**	**J6**
J1	0.00	4.47	2.00	3.32	3.46	6.32
J2	4.47	0.00	3.36	3.00	3.46	2.00
J3	2.00	3.46	0.00	3.87	3.46	5.29
J4	3.32	3.00	3.87	0.00	3.00	4.36
J5	3.46	3.46	3.46	3.00	0.00	4.47
J6	6.32	6.32	5.29	4.36	4.47	0.00

Step 4b

This step has focused to find minimum distance. The process has tailored every value at Table 3 into a pairwise decision table. Table decision is a table that presented relation of between schedule and other schedules. We have calculated to find the minimum distance in every pairwise schedule with (3). Omitted pairwise, at Table 4 has presented with pair between schedule and other schedules as many as schedule. E.g, if the schedule has 175 timeslots, then the pairwise would be made 175 x 175 schedule. At Table 4, we have presented an example in the pairwise schedule. At Table 4, every pairwise route has used to find minimum distance L_k.

TABLE IV. AN EXAMPLE PAIRWISE SCHEDULE IN DISTANCE

Ant	Route Pairwise		Lk (Minimum Lenght)
1	J1	J5	3,46
2	J2	J4	3,00
3	J3	J5	3,46
4	J4	J2	3,00
5	J5	J3	3,46
6	J6	J4	4,36

Step 4c

Step 4c is implemented for updating pheromone in global. We have used (5) and (6) to calculate updating pheromone.

At Table 5, Q is taken from cycle constant=1. Column L_k is taken from table 4. Meanwhile $\Delta\tau_{ij}$ (5) defined as update for pheromone.

TABLE V. AN EXAMPLE UPDATIING PHEROMONE USING (5) AND (6)

Q	Lk	Q / Lk
1	3,46	0,14451
1	3,00	0,16667
1	3,46	0,14451
1	3,00	0,16667
1	3,46	0,14451
1	4,36	0,11468
$\Delta\tau_{ij}$		0,88154

Step 5a

Counting intensity of ant footprints for next cycle. At Table 5, we have presented global updating pheromone. On this step, we should counted updating value for new intensity. Using intensity as following:

τij new= (0.03 x 0.01) + 0.88154 = 0.8818

At the process, we have shown, using (7) has retrieved new intensity = 0.8818

Step 5b

In this step, running proces ACO, has only retrieved value of new intensity. New intensity is stored in memory process to use for the next process. Value of new intensity = 0.8818.

Step 6

The last step in ACO has repeated the process. After Step 6, all the table or tabu list must be clear. Iteration in ACO, we have defined that repeat would be iterated to Step 2 if there is no collision in the schedule. ACO which has designed, the iteration would be terminated with bound of collision = 0 in the schedule. As long as the collision is not 0, then ACO is iterated to the Step 2. Illustration for collision can be seen at Table 6 and Table 7.

At Table 6, at the grey cell, the table has presented lecturer who has collision with other schedule. Schedule *J2* and *J3* have known as collision. *Timeslot(l,d,t)* of *J2* = 4,4,3 and *timeslot(l,d,t)* of *J3* = 4,4,3. If ACO was found collision like Table 6, then ACO would iterate to step 2.

Another collision in *timeslot(c,d,t)*. At Table 7, we can be seen collision at classroom, day, and time.

TABLE VI. AN EXAMPLE OF COLLISION FOR LECTURER, DAY, AND TIME

Schedule	*l*	Courses	Class subject	*c*	*d*	*t*
J1	1	1	5	2	3	2
J2	4	2	4	1	4	3
J3	4	3	2	1	4	3
J4	6	5	1	3	4	3
J5	3	4	3	2	3	2

TABLE VII. AN EXAMPLE OF COLLISION IN CLASSROOM, DAY, AND TIME

Schedule	***l***	**Courses**	**Class subject**	***c***	***d***	***t***
J1	1	1	5	2	3	2
J2	4	2	4	1	4	3
J3	4	3	2	1	4	3
J4	6	5	1	3	4	3
J5	3	4	3	2	3	2

TABLE VIII. AN EXAMPLE OF SUMMARIZE COLLISION

Schedule	Collision		Total of Collision
	Timeslot(l,d,t)	*Timeslot(c,d,t)*	
J1	0	1	1
J2	1	1	2
J3	1	1	2
J4	0	0	0
J5	0	1	1

At Table 7, grey area at schedule *J2* and *J3*, it can be seen the schedule collision at classroom, day, and time. *Timeslot(c,d,t)* of *J2* = *Timeslot(c,d,t)* of *J3*. *Timeslot(c,d,t)* of *J2* = 1,4,3 and *timeslot(c,d,t)* of *J3* = 1,4,3. It means, *J2* and *J3* has collided to each other.

At Table 6 and Table 7, the ACO process has shown a collision process. ACO has made the summarized table to retrieve where the schedule still reaches collision. We have prepared a summary table as storing table for schedule collision. At Table 8 is presented how many schedules have a collision on the first cycle.

At Table 8, in the column total of collision, there are still collision. ACO process is bounded by total of collision = 0. If the ACO process was presented like Table 8, then ACO process would be repeat to Step 2. In the process iteration, we have introduced a rule. The rule, only timeslot which have a collision, would be iterated.

IV. DISCUSSION

In this section, the results of the research that have been made, we used the term ant as a varabel of the term schedule. In this reasearch, teminology of Ant colony is known as a collection of schedules that have been arranged. The ACO application in the case under study has been changed at the time of the iteration. In ACO [10], iterations are carried out by comparing the value of the initial intensity with the intensity of the change. Iteration will stop when the difference in intensity meets the value of the conditions given. Whereas in the research conducted in this scheduling case, we used the number of collision for the iteration. The number of collision that we require is collision = 0. This is in line with research [1], [3], [10] which relies on intensity as a repetition requirement.ACO that has been implemented in the study, we made a test of the schedule used. The tests carried out can be seen at Table 9.

TABLE IX. TESTING PARAMETER ACO

Parameter Testing	Total
Total of Scheduling (testing scheduling)	175
Total of Classroom	9
Total of day	6
Total of timeslot	4
Population Scheduling	216

TABLE X. RESULT TEST ON 175 SCHEDULES

Number	Collision	Cycle	Time (Sec)
1	0	8	25,225
2	0	10	33,053
3	0	10	36,185
4	0	9	25,857
5	0	7	20,608
6	0	12	45,106
7	0	8	24,327
8	0	10	30,529
9	0	7	19,999
10	0	11	38,857

In Table 9, the testing parameter is a measure for optimizing ACO. Tests carried out, out of 216 population schedules, we only took 175 schedules. The optimization, for the ACO for the research conducted, can found the schedule with the minimum collision. This research was conducted by finding the value of optimization collision = 0. We can achieve the objectives of the study in accordance with the schedule optimization schedule.

The time given during the ACO process, provides efficiency in the preparation of SCHEDULING. We have tested 10 expereiment. Experiments conducted with a number of different cycles provide a significant time estimate [11]. More cycles give longer time. The experiments carried out also, have received collision = 0. Changes in parameters and intensities assumed by the number of collision provide fast time performance. In table 10, shows the experiment with the system that we have made.

In Table 10, we presented a conclusion that the application of the ACO algorithm can make scheduling optimal [12-14]. Furthermore, this research can make the scheduling process not occur in collision [15-17]. To reach the number of collision 0, on 10 trials the average requires as many as 9 cycles (rounding out of 9.2) with an average time of 29.98 seconds. The fastest time occurred in the 9th experiment, which was 19.99 seconds.

V. CONCLUSION

ACO implementation in research on scheduling optimization has an effect on the process of scheduling. Addition of calculations on intensity, provide experiment solutions by reaching the number of collision schedules to 0. Experiments carried out have reached an average time of 29.98 second. The time is reached on 9th cycles. We have concluded that improvement in ACO parameter would be improvement in scheduling. Primarily, on counting bound of iteration when process needs to loop. Proposed collision variable can caused looping reach collision = 0. Improvement of ACO for intensity have an influence to success achieve the schedule. This is properly aligned with problems that the aim of research.

REFERENCES

[1] Marques, “Ant Colony Optimisation for Job Shop Scheduling,” pp. 1–8.

[2] “SOLVING JOB SHOP SCHEDULING PROBLEM using AN ANT COLONY ALGORITHM Contribution / Originality,” vol. 5, no. 5, pp. 261–268, 2015.

[3] E. Science, “Production scheduling with ant colony optimization Production scheduling with ant colony optimization,” 2017.

[4] M. Tawfeek, A. El-sisi, A. Keshk, and F. Torkey, “Cloud Task Scheduling Based on Ant Colony Optimization,” vol. 12, no. 2, pp. 129–137, 2015.

[5] W. Amir, F. Wajdi, A. Azman, and A. Wahab, “Solving Vehicle Routing Problem using Ant Colony Optimisation (ACO) Algorithm,” vol. 5, no. 9, pp. 500–507, 2018.

[6] D. C. Ant, “Differential-Evolution-Based Coevolution Ant Colony Optimization Algorithm for Bayesian Network Structure Learning,” 2018.

[7] Z. Wei, “The Research of Genetic Ant Colony Algorithm and Its Application,” vol. 37, pp. 101–106, 2012.

[8] K. Akka and F. Khaber, “Mobile robot path planning using an improved ant colony optimization,” no. June, pp. 1–7, 2018.

[9] J. A. C. Math and R. Bv, “Solving Traveling Salesmen Problem using Ant Colony Optimization Algorithm,” vol. 4, no. 6, pp. 4–11, 2015.

[10] M. You-xin, Z. Jie, C. Zhuo, and A. B. Idea, “Production Scheduling,” no. 1.

[11] WANG, Gang; GONG, Wenrui; KASTNER, Ryan. Operation scheduling: algorithms and applications. In: High-Level Synthesis. Springer, Dordrecht, p. 231-255, 2008.

[12] CHEN, Wei-Neng; ZHANG, Jun. An ant colony optimization approach to a grid workflow scheduling problem with various QoS requirements. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol.39 no.1: 29-43, 2009.

[13] HUANG, Kuo-Ling; LIAO, Ching-Jong. Ant colony optimization combined with taboo search for the job shop scheduling problem. Computers & operations research, vol.35 no.4: 1030-1046, 2008.

[14] Atighehchian, A., Bijari, M., & Tarkesh, H. A novel hybrid algorithm for scheduling steel-making continuous casting production. Computers & Operations Research, vol.36 no.8, 2450-2461, 2009.

[15] Xing, L. N., Chen, Y. W., Wang, P., Zhao, Q. S., & Xiong, J. A knowledge-based ant colony optimization for flexible job shop scheduling problems. Applied Soft Computing, 10(3), 888-896, 2010.

[16] Arnaout, J. P., Rabadi, G., & Musa, R. A two-stage ant colony optimization algorithm to minimize the makespan on unrelated parallel machines with sequence-dependent setup times. Journal of Intelligent Manufacturing, 21(6), 693-701, 2010.

[17] Deng, G. F., & Lin, W. T. Ant colony optimization-based algorithm for airline crew scheduling problem. Expert Systems with Applications, 38(5), 5787-5793, 2011.

Towards Secure Risk-Adaptable Access Control in Cloud Computing

Salasiah Abdullah[1]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Khairul Azmi Abu Bakar[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—The emergence of pervasive cloud computing has supported the transition of physical data and machine into virtualization environment. However, security threat and privacy have been identified as a challenge to support the widespread adoption of cloud among user. Moreover, user awareness on the importance of cloud computing has increase the needs to safeguard the cloud by implementing access control that works on dynamic environment. Therefore, the emergence of Risk-Adaptable Access Control (RAdAC) as a flexible medium in handling exceptional access request is a great countermeasure to deal with security and privacy challenges. However, the rising problem in safeguarding users' privacy in RAdAC model has not been discussed in depth by other researcher. This paper explores the architecture of cloud computing and defines the existing solutions influencing the adoption of cloud among user. At the same time, the obscurity factor in protecting privacy of user is found within RAdAC framework. Similarly, the two-tier authentication scheme in RAdAC has been proposed in responding to security and privacy challenges as shown through informal security analysis.**

***Keywords*—Security; privacy; cloud computing; risk-adaptable access control; authentication**

I. Introduction

Cloud computing has shown a great impact in improvising information sharing between users from a different geographical location. It has led to a greater impact of knowledge sharing as people from different geographical location have the opportunity to access the cloud and share files without limited boundaries and network restriction. Alternatively, cloud computing is known as one of the best platform [1], [2] to meet the needs of consumers as it provides the platform of unlimited storage, managing and accessing data via network of remote server hosted on the internet. Thus, cloud computing has been the major savior to serve the rising needs of user who demands for storage capabilities and security as more and more documents are being created daily.

However, cloud security and privacy have been the major challenges in cloud computing when users lost their control in data storage due to the resources migration from physical to virtual storage [3], [4]. The escalation of trust relationship [5] between user and cloud provider is crucial in managing secured storage in cloud to cater the demand of user and resources that keeps growing.

This situation has been the rise factor of the introduction of access control which is a promised mechanism to ensure the enforcement of security policies in cloud. In the early stages of computing, security experts are eager in designing new security mechanism to handle massive changes in controlling access via cloud computing. Researches [6], [7] on the evolution of access control model such as Identification Based Access Control (IBAC) and Role Based Access Control (RBAC) showed the dependencies on predefined user identity and roles as it is working great in a non-distributed environment.

Besides that, IBAC has a problem with synchronization of remote user authentication and massive increase in administration overhead [8]. Later, RBAC was introduced which is based on role identification to gain access into the system. However, researchers found discrepancies in determining the privilege of user beyond administrative domains using RBAC [9]. Besides that, both IBAC and RBAC are known as conventional access control that only support static, rigid and limited support of access policies [6]. Thus, the concept of Attribute Based Encryption (ABE) has been intoduced to cater the difficulties in maintaining Access Control List (ACL) in a dynamic cloud environment [10]. In addition, [11] applied the concept of Ciphertext-Policy ABE (CP-ABE) to secure the resources and prevent unauthorized access but the implementation is limited to the data center. At the same time, we can see the transition of access control model development that relies on the high security needs and dynamic environment.

Furthermore, the emergence of access control authentication from a conventional secure password establishment to an attribute-based access control has led to the development of an efficient RAdAC to secure data in cloud. The advantages of RAdAC are the ability to cater the dynamic environment in handling exceptional access request and the flexibility in accessing resources [12]. This can resolve the issue using conventional password authentication scheme which depends on static access control policies and vulnerable to the password relevancy. Moreover, password authentication could not support rapid changing environment that involve massive user and resources in bulk.

Although both IBAC and ABAC are still widely used, RAdAC seems to be the latest evolution of access control model as not much research has been done yet. RAdAC applies the concept of analyzing each request dynamically as these request may be granted if the metric of risk is complied. However, there is a need to expand in line with the evolution of access control model. Most researchers who are involved in the development of RAdAC model only focus on the access authentication and resource encryption but neglect the need to preserve users' privacy.

Subsequently, the challenges in cloud security and the privacy-concern issue in RAdAC development has led us to propose a reliable and secure authentication scheme in two-tier architecture. Mutual authentication takes place as only authorized user get the privilege to access the resources in cloud. Besides that, user authenticity is verified using two-factor authentication which is user ID with password and signed token.

The structure of this paper consists of Section II which highlights on the related work in preparing secured cloud and discusses on the preliminaries of cloud computing and RAdAC. At the same time,the authentication scheme with two-tier security architecture has been proposed by expanding the capability of RAdAC model. It follows by informal security analysis presented in Section III shows that the proposed scheme offers privacy preserving access control through anonymous data transaction and mutual authentication. Furthermore, secured fine-grained access control is shown by the capability of the scheme in handling user revocation and password guessing attack. This is followed by conclusion in Section IV.

II. Material and Method

In this section, we discuss on the related work in handling cloud issues and security risk. It is follows by the overview of cloud computing, its security issues and proposed solution. In addition, we also discuss on RAdAC model and analyse existing framework.

A. *Related Work*

Various researchers have conducted researches focusing on security issues and challenges in controlling cloud technology. There are several organizations that play their role to initiate programs such as FT7, SWIFT and POSITIF in order to study and improve the dimensions of future cloud architecture [13].

Discussion in [14] revealed that the security issues could pose a threat to cloud computing and proposed security measures to handle the problem. However, the study focuses only on the current security issues and measures without considering on long-term cloud perspectives. Later, an expectations of future cloud research has successfully proposed in [15] by analyzing the strengths and weaknesses of security resolution to maintain a safe cloud environment. In addition, a study conducted in [16] towards security issues in services model of cloud computing is valuable but the security solution is applicable only on Cloud Service Provider (CSP).

Furthermore, reliability is believed to be one of the important aspect in decision support system to convince users that the resources obtained from the cloud are safe and accurate [17]. Nevertheless, awareness on the paramount secured factor in the cloud service environment has motivated significant researches towards reliable authentication framework in cloud computing [18].

In the nutshell, understanding how cloud works as well as identifying issues and security risks in cloud technology would be an important aspect to improve the possibility of users in adopting the technology. Moreover, determining the level of cloud capabilities and challenges may lead to the effective development of access control.

B. *Cloud Computing*

Cloud computing is the internet-based technology that includes a storage service and communication, efficient resource management and incurs minimal cost. In addition, cloud computing imposed on virtualization technology in providing computing resources based on user’s requirement [19]. Based on standard definition by National Institute of Standards and Technology (NIST), cloud computing is a model that allows network access to resources on configured computing (network, servers, applications, storage hub and services) with minimal administration or interaction [20].

Cloud computing architecture as Fig. 1 consists of four different layers which are standard definition, key features, service and deployment model. The standard definition of cloud acts as the first layer that shape the key features of cloud computing. Next, the second layer consists of five key characteristics of the cloud that drives consumer engagement in service model and deployment model.

On demand self-service is the ability of users to handle computing functions without service provider interaction. Pervasive network is a wide accessibility network from different user platform. Next, the resource pooling such as storage and network bandwidth is locally managed by the service provider in accordance to request from different users. Rapid elasticity is the ability of resource management and user to be scalable at any time. Lastly, measured service is the cloud's ability to automatically control and use resources in an optimal mode with the metering capability (pay-per-use basis).

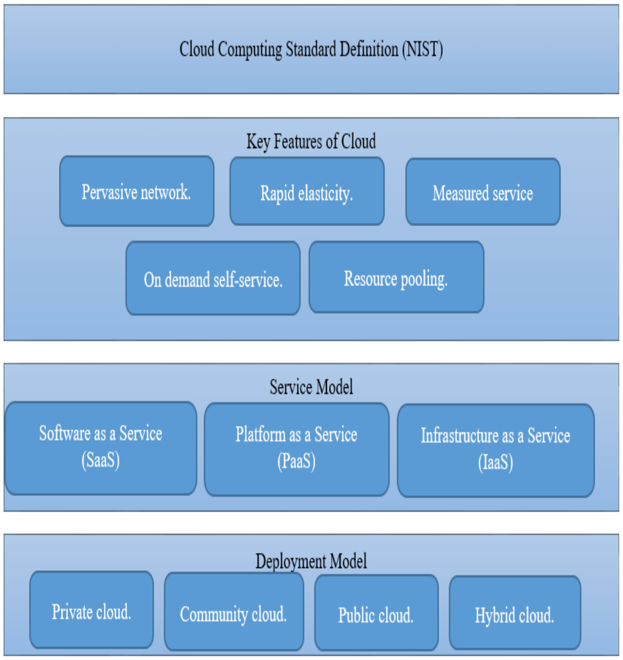


Fig. 1. Adaptation of Cloud Computing Architecture [14],[16],[20].

The third layer includes three service models in cloud computing which are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [21]-[22]. Service provider provides application software based on user demands in SaaS which has been the most common model in the organisation [23]. On the other hand, PaaS allows service provider to totally support the computing environment while in IaaS, service provider offers virtual infrastructure components to users.

In the deployment model, private cloud is an environment that is going through in-house development with specific resources to certain organizations while the public cloud is developed for general use. In addition, the community cloud is targeted to specific customer groups that share the same interest while hybrid cloud is a combination of private, public and community clouds. All of the mentioned type of clouds act as the fourth layer of cloud computing architecture [24].

Intelligibility of user on cloud computing architecture and its importance, as well as the use of cloud has increased the potential of the cloud to grow in the information technology industry. Failure to facilitate the cloud computing with high security control will lead to malicious attack as it may increase the possibility of information leakage. Healthcare industry may benefits most from the utilization of cloud services to cater the needs in protecting of sensitive information [25].

TABLE I. SECURITY ISSUE AND EXISTING SOLUTION IN CLOUD COMPUTING

Author	Security Issues	Existing Solution
Subashini and Kavitha [14]	• Security, integrity, confidentiality and data access. • Vulnerability in virtualization. • Availability, authentication and identity management.	• Intensify the Service Level Agreement (SLA). • Develop security framework. • Apply encryption and access control.
Zissis and Lekkas [17]	• Confidentiality and privacy. • Integrity. • Availability.	• Effective security management. • Develop security framework. • Apply cryptography encryption. • Implement access control.
Shahzad [27]	• Denial of Service (DoS). • Cloud storage security. • Integrity,confidentiality and data availability.	• Secured access control. • Effective identity management. • Encryption during authentication.
Y. Liu et al. [15]	• Data and security control. • Storage virtualization. • Authentication.	• Apply encryption. • Strengthen access control. • Effective security management.
Suzic et al. [13]	• Identity management. • Authentication and trust.	• Implement access control. • Apply cryptography for encryption.
Hepsiba and J.G.R. Sathiasee-lan [16]	• Malicious attack. • Denial of Service (DoS). • Security, integrity, confidentiality and data availability.	• Strong encryption and access control. • Management of information security. • Authentication protocol.

In addition, the level of privacy in the cloud environment could help to preserve the confidentiality of the data while protecting user identity. Whereas, level of reliability relies on effective cloud management by providing storage and communications to cater user needs. Thus, the level of privacy and reliability are the dependent factor to support the development of cloud technology in an organization.

1) Security issues and existing solutions: Previous study conducted on the issues and challenges of cloud security is a catalyst and serve as a benchmark in developing a comprehensive cloud environment. Organization is advised to analyze the security risk of cloud computing before jumping into an agreement to fully utilize the technology [21]. Table 1 summarizes a conducted past study in identifying security issues, challenges and proposed solutions in the cloud environment.

Existing solution as the Table 1 complies with the Information Security Principle which has been outlined as confidentiality, integrity and availability [21]. User needs to ensure security requirements in cloud such as reliability, authentication and identity management has been applied to protect the robustness of virtualization environment. However, the biggest challenge to rule out the principle is ensuring the capability of cloud in processing its resources with zero knowledge on the resources nor the identity of user [26].

C. Risk-Adaptable Access Control (RAdAC)

The increasing capacity of resources in access control system embarks the dynamic features in the architecture of RAdAC model. There are four components in existing RAdAC architecture which are Policy Enforcement Point (PEP), Policy Decision Point (PDP), Subject and Risk Engine as Fig. 2. The decision process starts when subject issues access request for specific resources. PEP handles access request from subject and sends it to PDP for access decision. If the request complies with the risk policy, Risk Engine performs risk quantification and analysis based on agreed metrics and response back to PDP. PEP will enforce obligation immediately after receiving decision from PDP. However, the existing architecture is vulnerable to security attacks by the curious component.

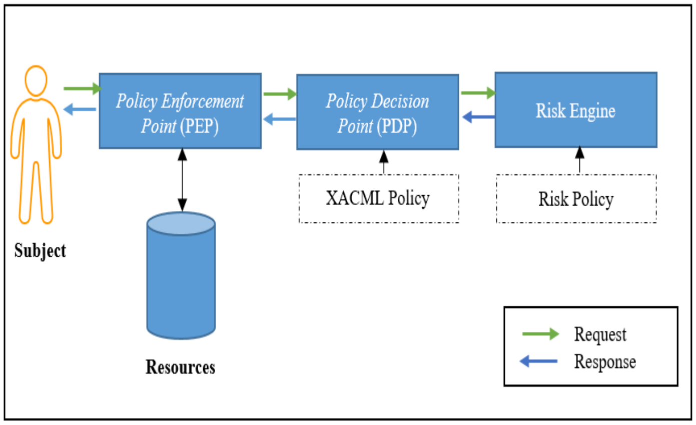


Fig. 2. Existing Architecture of Risk based Access Control [12].

The capability of RAdAC in managing ad-hoc request has become more prominent in access control environment compared to conventional predefined policies. RAdAC also works well in rapid-change environment to cater larger range of increase in users and resources. Nevertheless, failure in managing user identity and access structure in RAdAC has led to poor cloud management [12]. Thus, the implementation of RAdAC model is advantageous if the management of user identity and hidden access structure can be carried out effectively.

There are several works by other researchers in RAdAC development. Risk based access control which has been developed by [28] proposes the application of Policy Decision Point (PDP) with classifiers to quantify risk. Furthermore, RAdAC model has been implemented in healthcare information system to protect sensitive data and support dynamic health environment [29]. However, the implementation of RAdAC model in cloud computing is still in its infancy.

Thus, the fidelity of Risk Adaptive Authorization Mechanism (RAdAM) implementation in cloud has been proposed by [30] to determine access decision and introduce adaptable algorithm in cloud computing. However, the capability of RadAM in managing cloud federation has not been studied extensively.

Ontology in risk based access control is the extended work of [12] to define the independent risk policies of RAdAC model in specific hierarchical process. Subsequently, the ontology approach in RAdAC allows risk quantification without the need of cloud federation. At the same time, the indicative structure of the proposed model has been demonstrated by the privilege of Cloud Service Provider and its inference capability to support dynamism in access decision.

Subsequently, failure in managing identity of user in access control model may disrupt the effective implementation [12]. Thus, this research has been the benchmark for this paper in designing the architecture of proposed authentication scheme. Afterwards, most of previous works bypass the need to protect privacy of user in developing risk based access control model [28], [31], [32].

The analysis of existing RAdAC Model involved determining the related framework and refining the elements into the corresponding characteristics. Table II is a summary of publications related to RAdAC framework and published in journals and conferences. However, risk assessment is not within the scope of this paper as it highlights only on the privacy preserving in RAdAC.

As a result, three of the existing model shows the relationship involving the adaptation of risk metrics into RBAC model [31], [33], [34]. This concept supports the statement regarding access control evolution to adapt with flexible and dynamic features in cloud computing. Therefore, RAdAC is a continuous model that has been built using existing access control model as a basis. However, the development of RAdAC involves extensive improvement on additional function with element of risk and current context to cater the dynamicity of cloud environment.

TABLE II. COMPARISON TABLE OF RADAC FRAMEWORK

Elements	**Framework**				
	Ricardo dos Santos et al. [12]	***Khambhammettu et al.*** [35]	***Baracaldo and Joshi*** [33]	***Bijon et al.*** [34]	***Choi et al.*** [31]
Domain	Cloud computing	Not available	Not available	Not available	Real-time system.
Reliability	Make inferences by modifying weightage values based on risk metric.	Protect object in term of sensitivity and evaluate the user reliability on specified value.	Set threshold to prevent unauthorized access and abusive of data.	Deactivate user when system detects anomalies in user and run-time.	Identify specific method of medical analysis based on classified context.
Approach	RAdAC + Ontology	Threat assessment.	RBAC + risk and reliability assessment.	RBAC + risk assessment.	RBAC + risk assessment.
Advantage	Flexibility and dynamicity.	Protect from outsider attack.	Protect from insider attack.	Provide dynamic access.	Support medical information system.
Protection of users' privacy	Not available	Not available	Trust-based	Not available	Not available.
Access Decision	Depends on Aggregated Risk Score.	Depends on threat assessment score.	Depends on reliability threshold.	Depends on risk threshold.	Depends on risk level of patient.

Based on Table II, [34] is focusing only on the implementation of dynamic user authentication access while other studies [24]-[25], [27]-[28] discussed on the protection aspect of objects and resources with encryption methods or proven algorithm. At the same time, four from the five framework in the table did not mention about privacy protection of user. Hence, the need in safeguarding users' privacy in the RAdAC model has not been discussed in depth. Newer enhancement on the security and privacy landscape is compulsory to accelerate widespread adoption of cloud utilization among user.

D. Proposed Authentication Scheme

Authentication verifies user's identity and enables authorization to dictate different access of user. It is the way security system challenges user to prove identity credential based on something you know (e.g. password), what you have (e.g. digital certificate) and what you are (e.g. fingerprint) [36]. The architecture of risk based access control that has been defined in this paper is the extension from existing one-tier architecture of RAdAC Model that has been discussed by [12]. However, this scheme is the extended version of previous work which applied two-tier architecture as it offers protection of users' privacy by guaranteeing anonymity of information transfer using secure asymmetric cryptography method.

This method uses encrypting mechanisms by ensuring encapsulation of message to remain anonymous. Additionally, this method uses dual-keys which are public key for message encryption and private key for decryption of message. kc is assumed to represent the function of access decision value to support data transfer process as follows

$$kc : \mathbb{U}(U) \times S \times \mathbb{U}(P) \rightarrow \{0,1\} \qquad (1)$$

kc determines whether user with a set of identity attributes $u = \{u_1, \ldots, u_n\} \subset U$ get the permission to access resources $s \in S$ based on access policy $p = \{p_1, \ldots, p_n\} \subset P$. When subject/user U sends an access request for resources S in cloud, they need to register at Identity Provider (IdP) to comply with authentication process.

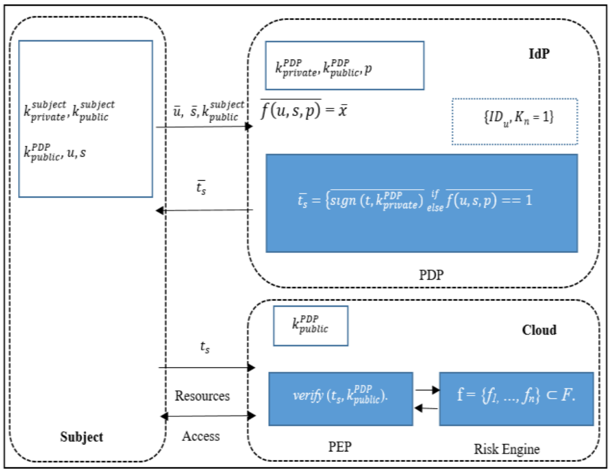

$k^{subject}_{private}, k^{subject}_{public}$
k^{PDP}_{public}, u, s
$\bar{u},\ \bar{s}, k^{subject}_{public}$
$\bar{t_s}$
t_s
Resources
Access
Subject
$k^{PDP}_{private}, k^{PDP}_{public}, p$
IdP
$\overline{f(u,s,p)} = \bar{x}$
$\{ID_u, K_n = 1\}$
$\bar{t_s} = \{\overline{sign\ (t, k^{PDP}_{private})}\ ^{if}_{else}\ \overline{f(u,s,p) == 1}$
PDP
k^{PDP}_{public}
Cloud
$verify\ (t_s, k^{PDP}_{public}).$
$f = \{f_1, \ldots, f_n\} \subset F.$
PEP
Risk Engine

Fig. 3. Proposed Authentication Scheme using Asymmetric Cryptography.

Based on Fig. 3, IdP with dual key ($k^{PDP}_{public}, k^{PDP}_{private}$) acts as PDP which received access request from subject and generate user ID, ID_u and temporary password PW_u randomly by $RPW_u = h\ (PW_u \mid\mid R_u)$ in the sign-up phase. Next, IdP will store $\{ID_u, K_n = 1\}$ in ID management table as $K_n = 1$ refers to active user who signs up once. K_n represented the number of registration that has been done by user. Login phase continues when user U send login request message $< ID_u, RPW_u >$ to IdP.

Encryption mechanism will takes place as PDP will sign the token and send the encrypted format to be verified by user and cloud manager (act as PEP) using PDP public key (k^{PDP}_{public}). Session key k^{PDP}_{public} is to be used by user and cloud manager as it is assumed to be delivered during the access request. Furthermore, user is occupied with dual key ($k^{subject}_{private}, k^{subject}_{public}$) to support the encryption mechanism. Next, authorization process begins as risk engine that has been invoked by the PEP started to analyze the risk policies based on risk metrics initiated by the cloud service provider or resource owner. User access to cloud is granted based on the predefined threshold that has been set at the first place.

III. Results and Discussion

In this section, informal security analysis shows the capability of authentication scheme in managing secure transaction.

Proposition 1: Proposed scheme offers secure anonymous transaction and mutual authentication.

Proof: In the recent scheme, identity of user is transmitted during the access request thus revealed the sensitive information of user to the cloud. In our proposed scheme, cloud cannot misuse user information as it only holds encrypted data of user. Anonymous transaction takes place as user send his public key $k^{subject}_{public}$ with identity attributes and requested resources in an encrypted format, $\bar{u}$ dan $\bar{s}$. PDP will user the session key $k^{subject}_{public}$ to encrypt p to $\bar{p}$ to compute kc. Next, $\overline{f(u,s,p)}$ generated encapsulated value $\bar{x}$ to ensure fine-grained access control. At the same time, this scheme offers privacy preserving of user identity, requested resources and the basic policy structure in the IdP. PDP will issue encrypted identity token $\bar{t_s}$, by signing the token t using $k^{PDP}_{private}$ as follows:

$$\bar{t_s} = \{\overline{sign\ (t, k^{PDP}_{private})}\ ^{if}_{else}\ \overline{f(u,s,p) == 1} \qquad (2)$$

Next, user will decide whether $\bar{t_s}$ fulfil his access request before decrypted the $\bar{t_s}$ using $k^{subject}_{private}$. Furthermore, user will use k^{PDP}_{public} to generate t_s as the verification process will be assigned by PEP using *verify* (t_s, k^{PDP}_{public}). Authorization process takes places as access decision is based on risk metrics $f = \{f_1, \ldots, f_n\} \subset F$ and risk threshold. Risk metrics such as user or device characteristic and situational, heuristic or environmental factors might influence the access decision.

Nevertheless, mutual authentication in this scheme is proved by two factor user authentication which is user ID with temporary password $< ID_u, RPW_u >$ and identity token that has been issued by IdP. It shows that access to cloud is granted only to an authorized user with valid credential.

Proposition 2: Proposed scheme support revocation or re-registration phase.

Proof: To initiate the revocation process, user U will send revoke request message to IdP. Next, IdP will store $\{ID_u, K_n = 0\}$ in its database where $K_n = 0$ shows user has been revoked and deactivated. If user need to re-register the services again, user need to prove his last valid ID ID_u and IdP will update back its data into $\{ID_u, K_n = 1\}$.

Proposition 3: Proposed scheme is secure against password guessing attack.

Proof: Password guessing attack by malicious user or untrusted cloud is impossible as they cannot initiate the value of parameter R_u. IdP will generate the temporary password randomly $RPW_u = \text{h}\ (PW_u \,||\, R_u)$ during the sign-up phase.

The informal analysis has been conducted as the justification to identify correct implementation and proof of concept of the authentication scheme. Thus, the scheme is viable in managing secure transaction, handling revocation and password guessing attack.

IV. Conclusions

Access control is one of the fundamental requirements in managing security risk. However, the rising needs in preserving users' privacy has been seen as the imperative obligation to protect identity of user. Therefore, security risks and privacy challenges in cloud should be taken into serious considerations. Furthermore, it plays a fundamental role in ensuring the wide adoption of cloud computing technology.

Similarly, the implementation of access control is crucial as RAdAC model offers dynamic characteristic in addressing vulnerabilities as it is able to deter the capabilities of conventional access control. In this paper, we have identified the general cloud architecture that serve as a benchmark in educating user on the paramount secured factor in the cloud computing service environment. Furthermore, analyzing security issues in cloud computing has diverse existing solution in defining a secure and reliable strategy against threats and vulnerabilities.

Subsequently, RAdAC model has been discussed by summarizing the existing framework to formulate a strategy in a systemic point of view. Thus, this discussion has envisioned the future roadmap in cloud by the introduction of two-tier security architecture in the authentication scheme. Informal security analysis demonstrated that the proposed scheme serves as a promising solution to cater security and privacy issues in cloud.

In the future, we plan to develop a framework of risk based access control with hidden access policy and apply the concept in real cloud platform.

References

[1] Meva and C. K. Kumbharana, "Issues and challenges of security in cloud computing environment.," Int. J. Adv. Netw. Appl., pp. 108–111, 2015.

[2] S. Abolfazli, Z. Sanaei, A. Tabassi, S. Rosen, A. Gani, and S. U. Khan, "Cloud adoption in Malaysia: Trends, opportunities, and challenges," IEEE Cloud Comput., vol. 2, no. 1, pp. 60–68, 2015.

[3] L. Wei et al., "Security and privacy for storage and computation in cloud computing," Inf. Sci. (Ny)., vol. 258, pp. 371–386, 2014.

[4] H. Takabi, J. B. D. Joshi, and G. J. Ahn, "Security and privacy challenges in cloud computing environments," IEEE Secur. Priv., vol. 8, no. 6, pp. 24–31, 2010.

[5] N. Fotiou, A. Machas, G. C. Polyzos, and G. Xylomenos, "Access control as a service for the cloud," J. Internet Serv. Appl., vol. 6, no. 1, 2015.

[6] A. H. Karp, H. Haury, and M. H. Davis, "From ABAC to ZBAC : The evolution of access control models," ISSA J., no. April, pp. 22–30, 2010.

[7] M. Mulimani and R. Rachh, "Analysis of access control methods in cloud computing," no. July, 2016.

[8] V. Boyko, P. Mackenzie, and S. Patel, "Provably secure password-authenticated key exchange using Diffie-Hellman," Eurocrypt, vol. 2, pp. 156–171, 2000.

[9] Y. Zhu, D. Huang, C.-J. Hu, and X. Wang, "From RBAC to ABAC: constructing flexible data access control for cloud storage services," IEEE Trans. Serv. Comput., vol. 8, no. 4, pp. 601–616, 2015.

[10] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," Proc. 13th ACM Conf. Comput. Commun. Secur. - CCS '06, p. 89, 2006.

[11] A. Sudarsono, M. U. Harun, and A. Rasyid, "Secure data sensor in environmental monitoring system using attribute-based encryption with revocation," vol. 7, no. 2, pp. 609–624, 2017.

[12] D. Ricardo dos Santos, R. Marinho, G. Roecker Schmitt, C. Merkle Westphall, and C. Becker Westphall, "A framework and risk assessment approaches for risk-based access control in the cloud," 2016.

[13] B. Suzic, A. Reiter, F. Reimair, D. Venturi, and B. Kubo, "Secure data sharing and processing in heterogeneous clouds," Procedia Comput. Sci., vol. 68, no. 316, pp. 116–126, 2015.

[14] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," J. Netw. Comput. Appl., vol. 34, no. 1, pp. 1–11, 2011.

[15] Y. Liu, Y. Sun, J. Ryoo, S. Rizvi, and A. V. Vasilakos, "A survey of security and privacy challenges in cloud computing: Solutions and future directions," J. Comput. Sci. Eng., vol. 9, no. 3, pp. 119–133, 2015.

[16] C. L. Hepsiba and J.G.R.Sathiaseelan, "Security issues in service models of cloud computing," Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 3, pp. 610–615, 2016.

[17] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," Futur. Gener. Comput. Syst., vol. 28, no. 3, pp. 583–592, 2012.

[18] A. J. Choudhury, P. Kumar, M. Sain, H. Lim, and J. L. Hoon, "A strong user authentication framework for cloud computing," in Proceedings - 2011 IEEE Asia-Pacific Services Computing Conference, APSCC 2011, 2011, pp. 110–115.

[19] C. G. Song, N. Y. Hwang, H. C. Yu, and J. B. Lim, "A dynamic resource manager with effective resource isolation based on workload types in virtualized cloud computing environments," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 5, pp. 1771–1776, 2017.

[20] P. Mell and T. Grance, "The NIST definition of cloud computing recommendations of the National Institute of Standards and Technology," Nist Spec. Publ., vol. 145, p. 7, 2011.

[21] B. Hari Krishna, S. Kiran, G. Murali, and R. Pradeep Kumar Reddy, "Security issues in service model of cloud computing environment," in Procedia Computer Science, 2016, vol. 87, pp. 246–251.

[22] Nurul Elliza Jasmin and Mohammad Khatim Hasan, "Framework for the implementation of E-Government system based on cloud computing for

Malaysian public sector," Ejournal.Ukm.My, vol. 7, no. 1, pp. 1–18, 2018.

[23] N. Jafri and M. M. Yusof, "Managing data security risk in model Software As a Service (Saas). Pengurusan risiko keselamatan data dalam model perisian sebagai perkhidmatan (Software As a Service)(Saas)," vol. 7, no. 1, pp. 99–117, 2018.

[24] M. H. Kayali, N. Safie, and M. Mukhtar, "Literature review of cloud based E-learning adoption by students: State of the Art and Direction for Future Work," IOP Conf. Ser. Mater. Sci. Eng., vol. 160, no. 1, 2016.

[25] A. Meri, M. K. Hasan, and N. Safie, "Success factors affecting the healthcare professionals to utilize cloud computing services," Asia-Pacific J. Inf. Technol. Multimed., vol. 6, no. 2, pp. 31–42, 2017.

[26] M. D. Ryan, "Cloud computing security: The scientific challenge, and a survey of solutions," J. Syst. Softw., vol. 86, no. 9, pp. 2263–2268, 2013.

[27] F. Shahzad, "State-of-the-art survey on cloud computing security challenges, approaches and solutions," 5th Int. Conf. Emerg. Ubiquitous Syst. Pervasive Networks (EUSPN-2014)/ 4th Int. Conf. Curr. Futur. Trends Inf. Commun. Technol. Healthc. (ICTH 2014)/ Affil. Work., vol. 37, no. 0, pp. 357–362, 2014.

[28] I. Molloy, L. Dickens, C. Morisset, P.-C. Cheng, J. Lobo, and A. Russo, "Risk-based security decisions under uncertainty," Proc. Second ACM Conf. Data Appl. Secur. Priv., no. February, pp. 157–168, 2012.

[29] K. Lee, S. G. Choi, D. H. Lee, J. H. Park, and M. Yung, "Self-Updatable encryption: Time constrained access control with hidden attributes and better efficiency," Adv. Cryptol. - ASIACRYPT 2013, vol. 8269, pp. 235–254, 2013.

[30] D. Fall, T. Okuda, Y. Kadobayashi, and S. Yamaguchi, "Risk Adaptive Authorization Mechanism (RAdAM) for Cloud Computing," J. Inf. Process., vol. 24, no. 2, pp. 371–380, 2016.

[31] D. Choi, D. Kim, and S. Park, "A Framework for context sensitive risk-based access control in medical information systems," Comput. Math. Methods Med., vol. 2015, 2015.

[32] D. Díaz-López, G. Dólera-Tormo, F. Gómez-Mármol, and G. Martínez-Pérez, "Dynamic counter-measures for risk-based access control systems: An evolutive approach," Futur. Gener. Comput. Syst., vol. 55, pp. 321–335, 2016.

[33] N. Baracaldo and J. Joshi, "An adaptive risk management and access control framework to mitigate insider threats," Comput. Secur., vol. 39, no. Part B, pp. 237–254, 2013.

[34] K. Z. Bijon, R. Krishnan, and R. Sandhu, "A framework for risk-aware role based access control," in 2013 IEEE Conference on Communications and Network Security (CNS), 2013, pp. 462–469.

[35] H. Khambhammettu, S. Boulares, K. Adi, and L. Logrippo, "A framework for risk assessment in access control systems," Comput. Secur., vol. 39, pp. 86–103, 2013.

[36] K. A. Abu Bakar and G. R. Haron, "Context-Aware analysis for adaptive unified authentication platform," in Proceedings of the 5th International Conference on Computing & Informatics, 2015, pp. 417–422.

Shape based Image Retrieval Utilising Colour Moments and Enhanced Boundary Object Detection Technique

Jehad Q. Alnihoud
Department of Computer Science, Al al-Bayt University, Al-Mafraq, Jordan

***Abstract*—The need for automatic object recognition and retrieval have increased rapidly in the last decade. In content-based image retrieval (CBIR) visual cues such as colour, texture, and shape are the most prominent features used. Texture features are not considered as a significant discriminator unless it is integrated with colour features. Colour-based image retrieval uses global and/or local features has proved its ability to retrieve images with a high degree of accuracy. In contrast, shape-based retrieval is still suffering from numerous unsolved problems such as precise edge detection, overlapping objects, and high cost of feature extraction. In this paper, global colour features are utilized to discriminate unrelated images. Furthermore, a novel hybrid approach is proposed, consisting of a combination of boundary-based shape descriptor (BBSD) and region-based shape descriptor (RBSD), image retrieval. An enhanced object boundary detection (EBOD) is proposed, which uses canny edge detector to detect shape boundaries, with morphological opening to remove isolated nodes. Subsequently, morphological closing is utilized to solidify objects within the target image to enhance shape-based features representation. Finally, shape features are extracted and Euclidean distance measure with different threshold values to measure the similarity between feature vectors is adopted. Five semantic categories of WANG image database are selected to test the proposed approach. The results of experiments are promising, when compared with most common related approaches.**

Keywords*—Boundary Based Shape Descriptor (BBSD); Region Based Shape Descriptor (RBSD); CBIR, EBOD; edge detectors

I. Introduction

CBIR is one of the most enthusiastic research area since 1970. It enables us to retrieve images based on visual content rather than textual description. Image databases with thousands or even millions of images are easily to create, maintain, and manipulate with less cost and high level of efficiency. It is obvious that many fields such as biometric security and medicine need better image database retrieval systems with degree of precision. As stated in [1], colour is one of the most important features to be extracted in any CBIR system. Many researchers deployed colour histogram approach. Colour histogram is easy to compute with acceptable level of retrieval accuracy. But it lacks to spatial distribution and less efficient in handling noise. To overcome limitations of the colour histogram colour moments is applied [2]. Human can easily recognize objects within an image. Therefore, shape descriptor is considered as one of the most significant descriptors that may enhance image based retrieval. There are two major methods of feature extraction in shape based image retrieval, namely boundary or contour based descriptor and region based descriptor.

Boundary based feature extraction is relied on outer boundary, while with region based feature extraction the whole region is considered[3].The boundary of objects within an image may identified by determining sharp discontinuities (changes in pixel intensity) within image. Sharp discontinuities in an image are very well known as edge detection [4]. Sobel, Prewitt, Canny, Laplacian and Roberts are considered as examples of traditional edge detection operators [5].

As stated in [6] boundary based shape descriptors (BBSD's) are needed when boundary (contour) has importance over the interior content of the shape while region based descriptors (RBSD's) are needed when the interior content of the shape is significant to the retrieval process. BBSD's and RBSD's are further classified as local descriptor and global descriptor. When image is segmented to different regions and features are computed and based on these regions then it is considered as local descriptor, while if the whole shape is considered as one region then it is considered as global descriptor. Most researchers considered either BBSD's [7] or RBSD's [8]. Furthermore, simple shape descriptors such as major axis length, eccentricity, and circularity are not perform as good discriminators if there is no big differences between shapes [6]. There is a lack of researches to explore the integration of boundary based and region based image retrieval techniques. Whenever, it comes to shape based retrieval a lot of concern is given to global features extraction and boundary based retrieval. That because shape based features extraction is a time consuming process, so most of researchers compromises between efficiency and accuracy. In this paper, a hybrid approach to combine boundary based shape descriptors and region based shape descriptors is proposed. To enhance the accuracy of retrieval and maintain high level of feature extraction efficiency, a systematic approach is proposed to isolate the shape region from the background, enhance object recognition by proposing a new algorithm named as enhanced boundary object detection (EBOD), utilising morphological opening to remove isolated nodes and morphological closing to solidify objects within image, and extract global shape based descriptors. The time consuming processes are done as off-line process, while simple global features extraction of image query is done as on-line process. The rest of the paper is organized as follows.

Section 2 illustrates the proposed approach in details. Section 3 presents similarity measures deployed in this research. Section 4 discusses experimental results. Finally, Section 5 highlights conclusion and future work.

II. Proposed Approach

The proposed approach allows colour features extraction based on colour moments, while object based recognition is done using boundary based and region based techniques. Features extracted via colour and shape is combined together in one feature vector to represent target image. Then similarity based on Euclidean distance measure is used to rank and retrieve images. Figure 1 shows the system diagram of the proposed approach.

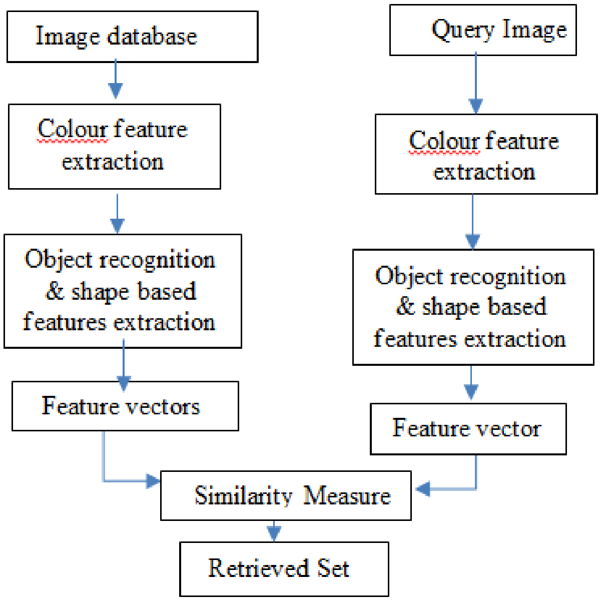


Image database
Query Image
Colour feature extraction
Colour feature extraction
Object recognition & shape based features extraction
Object recognition & shape based features extraction
Feature vectors
Feature vector
Similarity Measure
Retrieved Set

Fig. 1. System Flow Diagram.

Figure 2 presents the interface of the proposed system. Query by example (QBE) is adopted in this work.

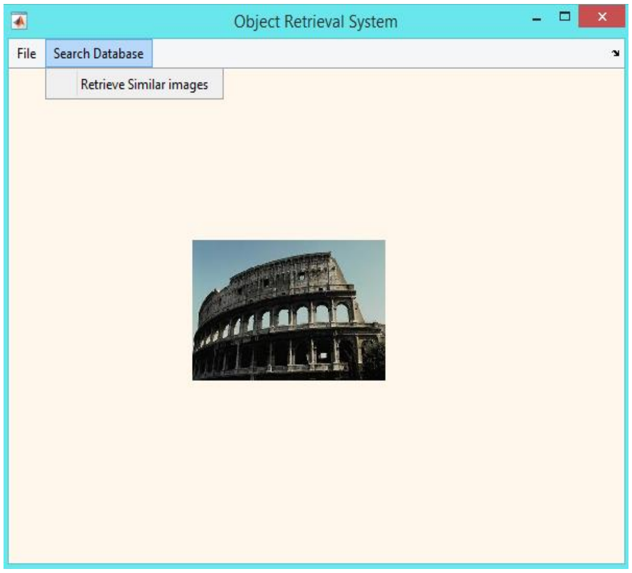


Fig. 2. Proposed System Interface.

A. Colour Features Extraction

Global and local colour features are widely used in CBIR. Many research attempts is made to localize colour features by dividing images into equal sub images or dividing images to overlap sub images [9]. Colour-based image retrieval utilising local features overcomes the limitations of global features like; the depiction of spatial distribution of colours. In this research, spatial distribution of colours is not significant because shape rather than colour is the main concern. Consequently, global colour features were adopted to reduce the cost of computation. Colour feature extraction involves two steps:

- Separate each channel of the RGB images to R, G, and B.
- Extract features (colour moments) shown in table 1 for each channel.

TABLE I. Colour Features Description

Colour Features	**Description**	**MATLAB function**
Mean	The mean value for each colour channel R, G, and B.	mean2
Standard deviation	The standard deviation value for each colour channel R, G, and B.	std2
Entropy	The entropy value for each colour channel R, G, and B. As stated in [10] entropy is a measure of randomness used to depict the texture of the input image. Entropy is defined as : -sum(p.*log2(p))	Entropy

B. Object Recognition

One of the most challenging topics in shape-based image retrieval is the accuracy object identification and recognition. In order to, retrieve images based on shape we need to identify objects, isolate them from background, and extract shape-based features. Figure 3 shows the proposed system architecture to identify, enhance, and solidify object.

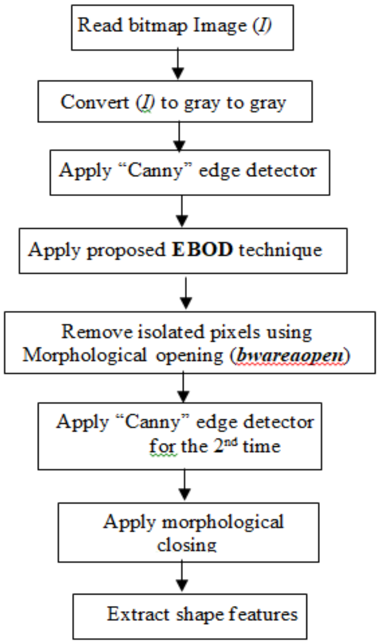

Read bitmap Image (*I*)
Convert (*I*) to gray to gray
Apply "Canny" edge detector
Apply proposed **EBOD** technique
Remove isolated pixels using Morphological opening (***bwareaopen***)
Apply "Canny" edge detector for the 2nd time
Apply morphological closing
Extract shape features

Fig. 3. Proposed Object Recognition Model.

1) Enhanced Boundary Object Detection (EBOD): EBOD technique is a novel technique proposed to enhance object boundary (contour) detection and fix-up. It consists of two major steps as sated below.

Step1: Pre-processing

- Apply canny edge detector operator
- Apply Morphological opening to remove isolated pixels using *bwareaopen(I, 50)*

Step 2: Identify start and target points to fill in between

Aim: To solve the problem disconnected edges, which are considered as part of object boundary (contour).

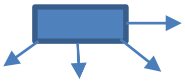
Definition:

- Start point S_p: Last node connected to other nodes, $S_p = (x_1, x_2)$
- Target Point (T_p): First point to connect with,

  $T_p = (y_1, y_2)$

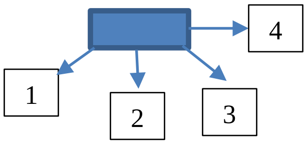
Algorithm I: Identify Start and Target point (IDST)

a) Scan the image from left to right row by row until node with no one of the 4-connectivities shown in the illustration below is reached. Then consider it as $S_p$with x_1 and x_2.



b) To identify the $T_p$continue scanning after $S_p$until a node with 1's is found then consider that point as T_p and apply the suitable rule as shownin Algorithm II.

- After filling the nodes in between Sp and Tp, check the connectivity's of Tp based on the following priority rules.



```
If 1 is True then move to 1
        Else if 2 is True move to 2
        Else if 3 is True move to 3
        Else if 4 is True move to 4
  Else
      Let Tp = New Sp then go to b
```

Algorithm II: Join Disconnected Edges (JDE)

$S_p = (x_1, x2)$
$T_p = (y_1, y2)$
$\Delta r = y_1 - x_1$: difference in rows
$\Delta c = y_2 - x_2$: difference in columns

1st case:

If ($\Delta r = 1$ && $\Delta c \geq 1$) then do:

- Let $x_1 = x_1 + 1$ and $x_2 = x_2 + 1$
- Fill in between (x_1, x_2) and (y_1, y_2) with 1

2nd case:

If ($\Delta r \geq 1$ && $\Delta c < 1$)
That indicate to left connectivity then do:

- Based on Δr value add one to x_1 gradually and fill the new points with 1's.
- Fill in between last point reaches $(x_1, x2)$ and target point (y_1, y_2) with 1's.

3rd case:

If ($\Delta r \geq 1$ && $\Delta c \geq 1$ && $\Delta r == \Delta c$)
That indicate to diagonally right connectivity then do:

- Add one to x_1 and x_2 till we reach target point and fill the new points with 1's.

4th case:

If ($\Delta r == 0$ && $\Delta c \geq 1$)
Then, the start point and the target point at the same row. Do:

- Add the Δc gradually to the start point and fill new points with 1's till it reaches the target point.

5th case:

If ($\Delta r > 0$ &&& $\Delta c < 0$) then that indicate to left diagonally connectivity then do:

- Add +1 to x1 and -1 to x2
- Fill in between with 1's till it reaches target point

The following figures (4-11) show an example to explain the different cases of **JDE** filling. Consider figure 4 as the original image after edge detection is applied.

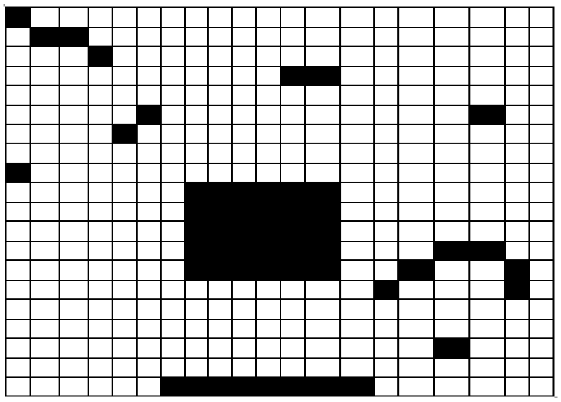


Fig. 4. Original Image.

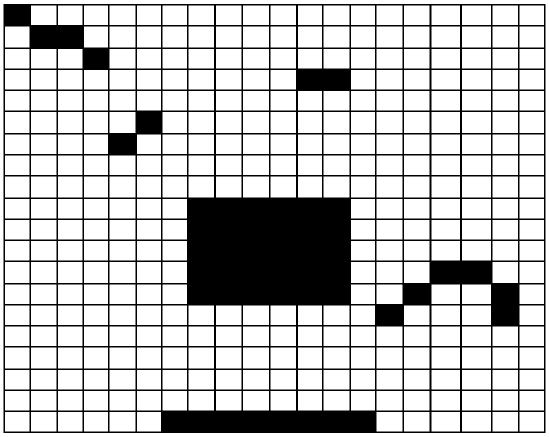


Fig. 5. Remove Isolated Nodes.

Figure 5 shows an example of removing isolated nodes.

Figure 6 shows an example of the 1st case. Index of last point is (3, 4), target point is (4, 12), $\Delta r = 1$, $\Delta c = 8$. Add one to the row and column of the first point (4, 5) then fill in between (4, 5) and (4, 12) with 1's.

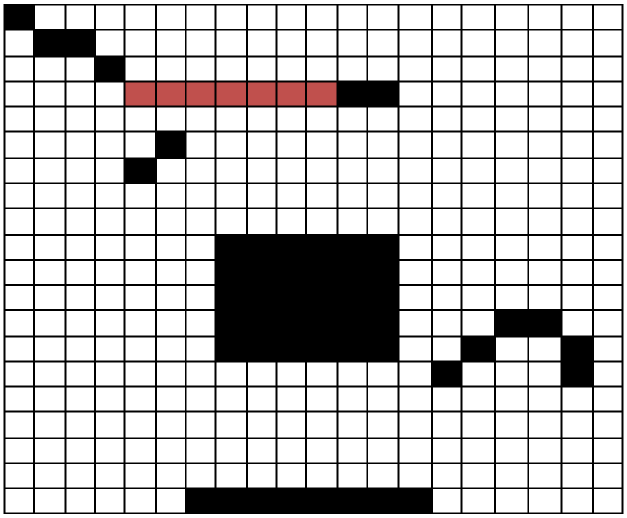


Fig. 6. 1st Case.

Figure 7 shows an example of the 2nd case. Index of last point is (4, 13), target point is (6, 6), $\Delta r = 2$, $\Delta c = -7$. Add one to the row of S_p (5, 13) and fill with 1, add 1 to S_p again (6, 13) and fill with 1. Finally, fill in between (6, 6) and (6, 13) with 1's.

Figure 8 shows an example of the 3rd case. Index of last point is (7, 5), target point is (10, 8), $\Delta r = 3$, $\Delta c = 3$. Add one to x_1 and x_2 gradually to reach target point and fill the in between points with 1's.

Figure 9 shows an example of the 4th case. Index of last point is (14, 13), target point is (14, 16), $\Delta r = 0$, $\Delta c = 3$. The two points at the same row, so fill in between with 1's. In between points (15, 15) and (15, 19), 4th case was applied.

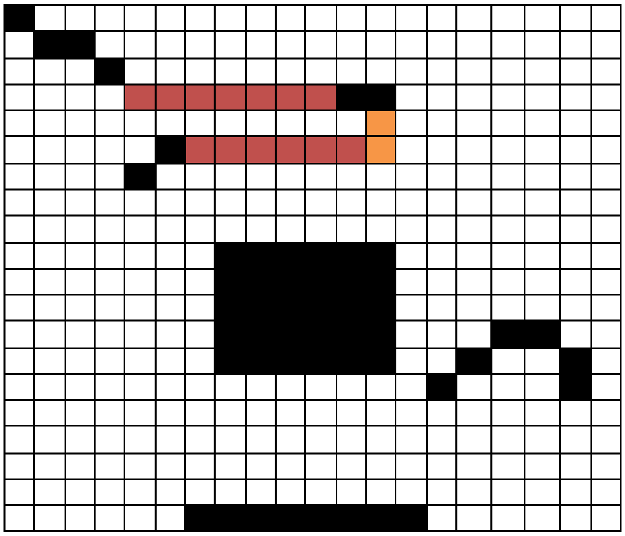


Fig. 7. 2nd Case.

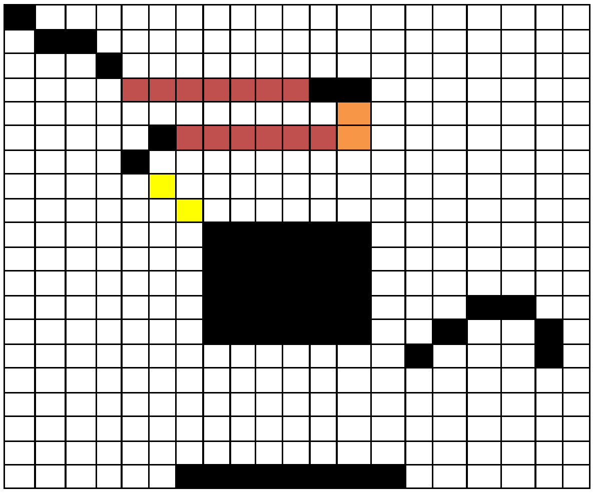


Fig. 8. 3rd Case.

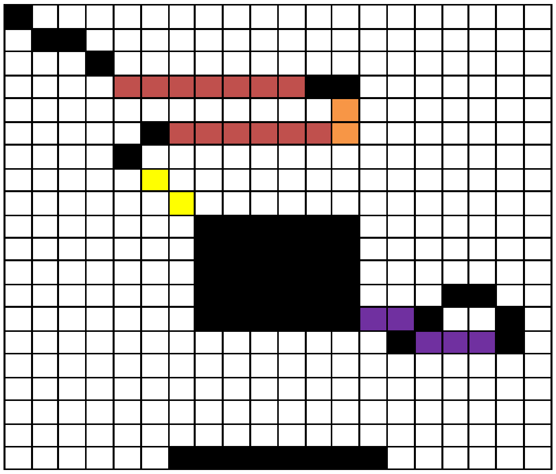


Fig. 9. 4th Case.

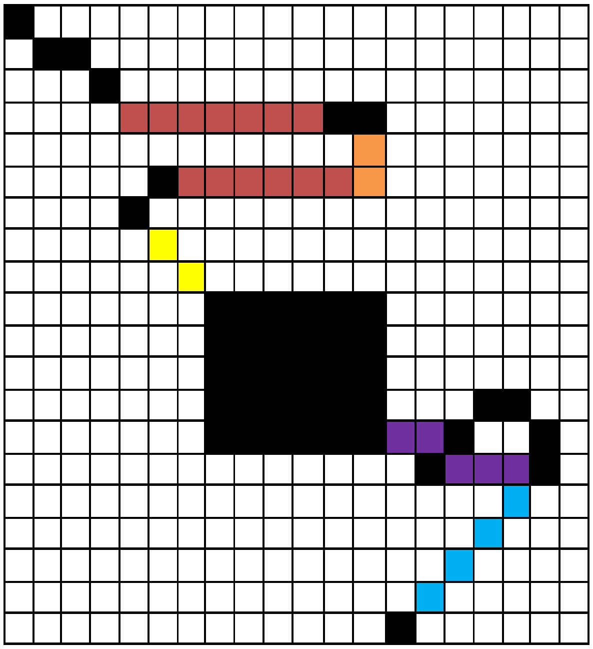


Fig. 10. 5th Case.

Figure 10 shows an example of the 5th case. Index of last point is (15, 19), target point is (20, 14), Δr = 5, Δc = -12. Add 1 to x_1 and -1 to x_2 gradually to reach target point and fill the in between points with 1's.

Figure 11 shows the final result.

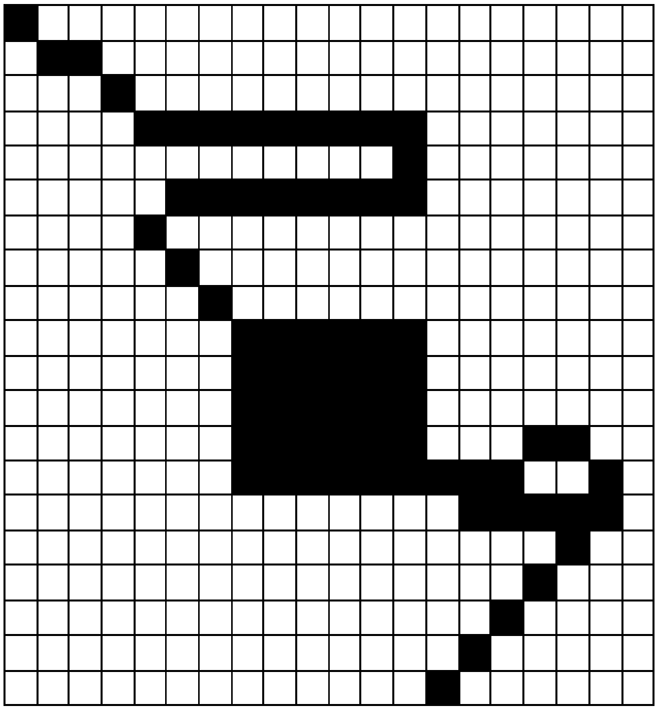


Fig. 11. Final Result of EDEA.

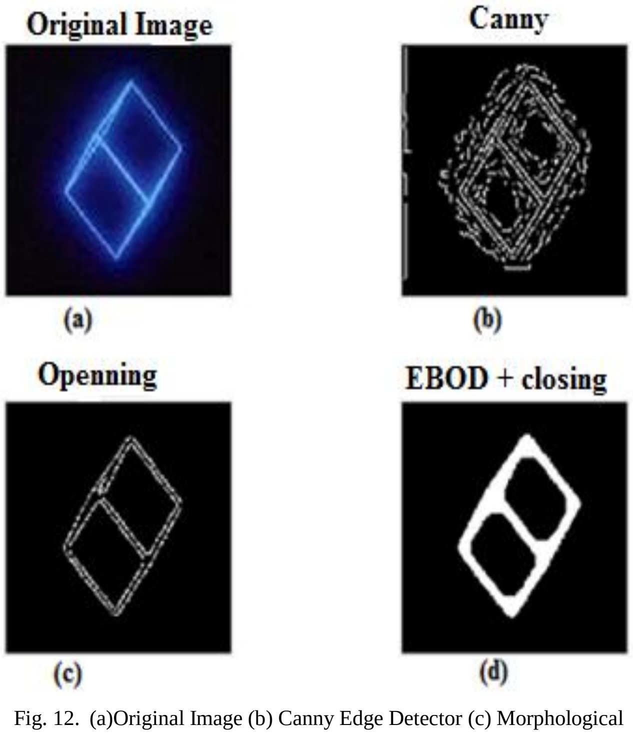


Fig. 12. (a)Original Image (b) Canny Edge Detector (c) Morphological Opening (d) EBOD with Closing Morphological Operator.

Figure 12 shows the result of applying object recognition model to a real image shown at the upper left corner. In order to achieve better results, the scanning process is done twice from left to right and from right to left. The same approach EBOD is applied with slight modification.

2) Shape-based features: Features selection is crucial to any CBIR system. In this research, many shape-based features were tested and the most significant features are selected to represent objects within image and to compare query image with images database. Table 2 shows shape features that are selected from list of features available in MATLAB R2015a documentation. **Regionprops** function in MATALB is used with the following features shown in Table 2.

TABLE II. SHAPE FEATURES DESCRIPTION

Shape feature	**Description**
Area	It is a scalar that specifies the actual number of pixels in the region.
Centroid	It is a vector that represents the center of mass of the region. In this research, the first two elements of the vector are considered. These elements specifies the horizontal and vertical coordinate of the center of mass.
Major Axis Length	It is a scalar that identifies the length of the major axis of the ellipse. The ellipse and region have similar normalized second central moments.
Minor Axis Length	It is a scalar that specifies the length of the minor axis of the region based on ellipse that has normalized second central moments same as the region.
Eccentricity	The eccentricity is the ratio of the distance between the foci of the ellipse -that has the same second moment as the region- and its major axis length.
Equiv. Diameter	It is a scalar that identifies the diameter of a circle that has the same area as the region. Calculated as sort(4*Area/pi).

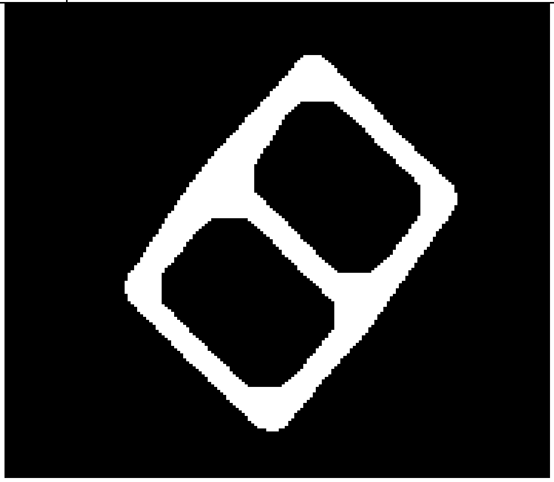


Fig. 13. Example Image.

TABLE III. SHAPE BASED FEATURES

Area	**Centroid**	**Major Axis Length**	**Minor Axis Length**	**Eccentricity**	**Equiv. Diameter**
3891	103.56 98.045	150.53	105.03	0.71	70.386

Table 3 shows an example of shape features extracted, which represent the image shown in figure 13.

III. Similarity and Performance Measures

Distance measure or distance functions are used to calculate the similarity between feature vectors. As stated in [11], Euclidian distance measure is more precise as compared to the most common distance measures available. Euclidean distance measures the similarity by calculating the distance between two vectors using square root of the sum of the squared difference between two vectors v1 and v2.

$$\text{Euc} = \sqrt{\sum_{i=1}^{n} (v1_i - v2_i)^2} \qquad (1)$$

To measure the performance of any CBIR systems precision and recall are the most common measures available. Precision measure the effectiveness of retrieval as it measures accuracy of the retrieved set as compared to the query image [12]. Precision is calculated by dividing the number of relevant (accurate) retrieved images over the number of all retrieved images, while recall measures the system ability to return as much as possible of relevant images available in the database [13].As stated in [13] the following formulas define precision and recall.

$$P_{recision} = \frac{Relevant\ Hits}{All\ Hits} \qquad (2)$$

$$R_{ecall} = \frac{Relevant\ Hits}{Expected\ Hits} \qquad (3)$$

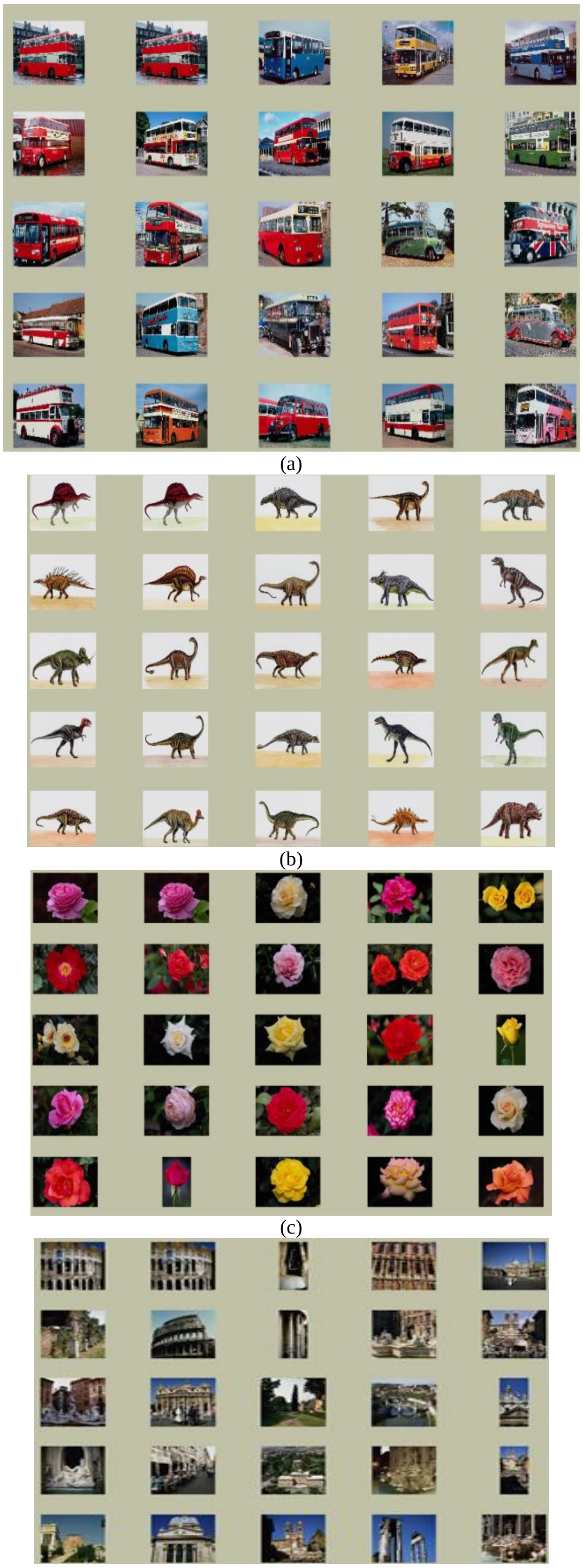
When precision and recall is measured to the same image using different threshold and many hits (images) are used to evaluate the precision and recall of the same data set then average precision, average recall, and overall average precision are needed to measure the performance of the proposed system.

IV. Experimental Results

The proposed approach is implemented using MATLAB R2015a with WANG database [14]. WANG database is extensively used in CBIR to test the effectiveness of any CBIR system because of its clear categorization and reasonable size in each category [11]. From this database, five semantic categories of WANG database are chosen as shown in table 4. If the distance between image query and image database features is less than or equal a given threshold (α) then it is considered as similar image. The distance measure is calculated based on (eq. 1). Then precision, recall, average precision, and average recall are used to evaluate the system performance.

TABLE IV. Categories of Test Images

	Image index	
Image category	***From***	***To***
Buildings	0	99
Buses	100	199
Dinosaurs	200	299
Roses	400	499
Horses	500	599



(a)

(b)

(c)

(d)

Fig. 14. (a)Retrieved Samples. Busses Retrieved Set (b) Dinosaurs Retrieved Set (c) Flowers Retrieved Set (d) Buildings Retrieved Set.

Figure 14 shows an example of retrieved sets for different image query from different image categories available in WANG database.

Image queries are shown in the upper left corner of each retrieved set. Due to space limitation it is not possible to show all samples of retrieval results.

From each category, **five** different images are selected as image queries. At different threshold (α) and for each query image number of retrieved images (NRI), precision (PR), and recall (RE) are calculated. The following tables (5-9) show the results of testing for the different semantic categories.

TABLE V. TEST RESULTS PRECISION VS. RECALL (BUILDINGS)

Query	$\alpha \leq 0.1$			$\alpha \leq 0.2$			$\alpha \leq 0.3$		
	NRI	PR	RE	NRI	PR	RE	NRI	PR	RE
Q1	20	0.90	0.18	30	0.86	0.26	42	0.83	0.35
Q2	25	0.89	0.22	33	0.87	0.29	45	0.84	0.38
Q3	18	0.78	0.14	29	0.82	0.24	33	0.69	0.23
Q4	32	0.93	0.30	38	0.92	0.35	50	0.90	0.45
Q5	28	0.78	0.22	36	0.77	0.28	43	0.76	0.33
Average		**0.86**	**0.21**		**0.84**	**0.28**		**0.81**	**0.35**

TABLE VI. TEST RESULTS PRECISION VS. RECALL (BUSES)

Query	$\alpha \leq 0.1$			$\alpha \leq 0.2$			$\alpha \leq 0.3$		
	NRI	PR	RE	NRI	PR	RE	NRI	PR	RE
Q1	30	0.97	0.29	39	0.95	0.37	50	0.94	0.47
Q2	32	0.97	0.31	38	0.94	0.31	49	0.93	0.46
Q3	40	0.98	0.39	45	0.97	0.44	53	0.96	0.51
Q4	45	0.93	0.42	47	0.89	0.42	55	0.87	0.48
Q5	36	0.80	0.29	43	0.77	0.33	49	0.73	0.36
Average		**0.93**	**0.34**		**0.91**	**0.37**		**0.89**	**0.46**

TABLE VII. TEST RESULTS PRECISION VS. RECALL (DINOSAURS)

Query	$\alpha \leq 0.1$			$\alpha \leq 0.2$			$\alpha \leq 0.3$		
	NRI	PR	RE	NRI	PR	RE	NRI	PR	RE
Q1	35	0.97	0.34	46	0.97	0.45	55	0.96	0.53
Q2	38	1.0	0.38	45	1.0	0.45	54	0.98	0.53
Q3	36	1.0	0.36	42	0.98	0.41	50	0.96	0.48
Q4	40	0.98	0.39	46	0.96	0.44	58	0.96	0.56
Q5	42	1.0	0.42	52	1.0	0.52	60	1.0	0.60
Average		**0.99**	**0.38**		**0.98**	**0.45**		**0.97**	**0.54**

TABLE VIII. TEST RESULTS PRECISION VS. RECALL (ROSES)

Query	$\alpha \leq 0.1$			$\alpha \leq 0.2$			$\alpha \leq 0.3$		
	NRI	PR	RE	NRI	PR	RE	NRI	PR	RE
Q1	38	0.95	0.36	47	0.94	0.44	55	0.91	0.50
Q2	39	0.95	0.37	49	0.93	0.46	57	0.92	0.53
Q3	40	0.97	0.39	50	0.96	0.48	59	0.94	0.56
Q4	52	0.96	0.50	56	0.94	0.53	60	0.93	0.56
Q5	43	0.88	0.38	48	0.87	0.42	54	0.87	0.47
Average		**0.94**	**0.40**		**0.93**	**0.47**		**0.91**	**0.52**

TABLE IX. TEST RESULTS PRECISION VS. RECALL (HORSES)

Query	$\alpha \leq 0.1$			$\alpha \leq 0.2$			$\alpha \leq 0.3$		
	NRI	PR	RE	NRI	PR	RE	NRI	PR	RE
Q1	35	0.85	0.35	45	0.84	0.38	57	0.84	0.48
Q2	40	0.90	0.36	50	0.86	0.43	63	0.82	0.52
Q3	45	0.88	0.40	58	0.86	0.50	68	0.85	0.58
Q4	39	0.89	0.35	43	0.88	0.38	49	0.81	0.40
Q5	38	0.94	0.36	47	0.93	0.44	53	0.92	0.49
Average		**0.89**	**0.36**		**0.87**	**0.43**		**0.85**	**0.49**

TABLE X. RESULTS SUMMARY

Image Category	**Overall Average precision**	**Overall Average recall**
Buildings	0.84	0.28
Buses	0.91	0.39
Dinosaurs	0.98	0.46
Roses	0.93	0.46
Horses	0.87	0.43
Average	**0.91**	**0.41**

Previous tables show that there is slight decrease in precision as recall increases. Recall is controlled through the tolerance (threshold) value. The following table 10 summarizes the testing results.

As shown in table 10 the best precision is achieved with dinosaur's category and buses, while roses achieved better results as compared with buildings and horses. Nature of image has dramatic effect on any CBIR system. Moreover, when the object is in contrast with the image background, the shape based image retrieval achieved better results. Dealing with real world images as in WANG database is challenging as compared with synthesized image databases such as MPEG 7. Table 11 shows two of the most relevant approaches to compare with.

TABLE XI. RELEVANT APPROACHES

Ref. No.	**Segmentation technique**	**Features extracted**	**CBIR method**	**Data sets and Performance**
15	Canny edge detector	Affine invariant signature, length, and curve descriptors	ACID descriptor with curve	Fish shape, MCD and MPEG-7 CE data sets. The precision performance as 91%, 97% and 84% respectively.
16	Canny operator and morphology methods	Space symmetry length sequence	Global space symmetry matching.	MPEG-7 shape CE Part B dataset. Precision rate of 88.01%

The overall precision of the proposed approach is 91%. The image database used to test the proposed system is WANG database containing real world (natural) images. Despite of the nature of the image database the results obtained is better as compared with two of the most common approaches presented in [15], and [16]. In [15], the overall precision is 90%, while in [16] its 88%. Consequently, the proposed approach is efficient, robust, and realistic.

V. Conclusion and Future Work

A new and efficient approach to shape based image retrieval is proposed in this paper. Edge detection as a boundary based technique, and region-based technique utilising morphology are used in addition to colour moments. Combining colour and shape is essential to discriminate the unrelated images being retrieved. Furthermore, shape features extracted using boundary-based image retrieval are not necessarily suitable for region-based image retrieval. In order to, improve the precision of retrieval a new edge detection enhancement algorithm is proposed. This algorithm is used to remove isolated nodes and solve the problem of disconnected edges raised by the most common edge detectors such as Canny, Sobel, and Prewitt. Global colour features and region-based local features are integrated to enhance the accuracy of retrieval. Furthermore, morphological operators opening and closing are used in systematic way to solidify object's region within the target image. As a result, the shape discriminators are able to discriminate images precisely. The experimental results obtained are promising when compared with the most common approaches related to the method proposed in this paper. Despite using a real world images from the WANG database, the precision rate is still high. As future work, user feedback to bridge the semantic gap will be explored, in order to enhance the system performance based on supervised learning.

References

[1] S. H. Shirazi, A. I. Umar, S. Naz, N. A. Khan, M. I. Razzak, B. Alhaqbani, “Content-based image retrieval using texture colour shape and region”, International Journal of Advanced computer science and applications (IJACSA), Vol. 7, No. 1, 2016.

[2] A. J. K. Iqbal, M. O. Odetayo, “Content-based image retrieval approach for biometric security using colour, texture and shape features controlledby fuzzy heuristics,” Journal of Computer and System Sciences, Vol. 78,pp. 1258-1277, 2012.

[3] T. S. B.S. Manjunath, P. Salembier, Introduction to mpeg-7: Multimedia content description interface, Wiley, Chichester, England, ISBN 978-0-471-48678-7, 2002.

[4] R. C. Gonzalez, and Richard E. Woods, Digital Image Processing, 2ndedition, Beijing: Publishing House of Electronics Industry, 2007.

[5] G. Kumar Srivastava, R. Verma, R.Mahrishi, S. Rajesh, "A novel wavelet edge detection algorithm for noisy images", International conference on Ultra-Modern Telecommunications & Workshops (ICUMT), pp. 1-8**,**2009. **DOI:** 10.1109/ICUMT.2009.5345404.

[6] P. Singh, V. K. Gupta, P. N. Hrisheekesha, “A review on shape based Descriptors for Image Retrieval”, International Journal of Computer Applications, Vol. 125, No. 10, pp. 27-32, 2015.

[7] Z. Huang, JinsongLeng, “Analysis of Hu's Moment Invariants on Image Scaling and Rotation,” IEEE 2ndInternational Conference on Computer Engineering and Technology, Vol. 7, pp. 476-480, 2010.

[8] Rong-Xiang Hu, Wei Jia, Haibin Ling, Yang Zhao, JieGui, “Angular Pattern and Binary Angular Pattern for Shape Retrieval”, IEEE Transactions on Image Processing, Vol. 23, No. 3, pp. 1118-1127, March 2014.

[9] J. Q. Alnihoud, “Image Retrieval System based on Colour Global and Local Features Combined with GLCM for Texture Features”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 9, 2018.

[10] R. C. Gonzalez,R.E. Woods, S.L. Eddins, Digital Image Processing Using MATLAB, New Jersey, Prentice Hall, 2003, Chapter 11.

[11] Y. Mistry, D. T. Ingole, M. D. Ingole, “Content based image retrieval using hybrid features and various distance metric”, Journal of Electrical Systems and Information Technology, 2017, http//dx.doi.org/10.1016/j.jesit.2016.12.009.

[12] H. Muller, W. Muller, S. M. Maillet,T. Pun, D. McG. Squire, "A framework for benchmarking in visual information retrieval," International Journal on Multimedia Tools and Applications, Vol. 21, No. 1, pp. 55-73, 2003.

[13] S. Manjus, N. Raj, "Content based image retrieval using wavelet transform and feedback algorithm", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Special Issue 5, July 2014.

[14] C. Man Pun, and C. F. Wong, “Fast and robust, colour feature extraction for content- based image retrieval,” International Journal of Advancements in Computing Technology, Vol. 3, No. 6, 2011.

[15] H. Fu, Z. Tian, M. Ran, M. Fan, “Novel affine-invariant curve descriptor for curve matching and occluded object recognition” IET Computer Vision, Vol. 7, No. 4, pp. 279–292, 2013.

[16] Yu Shi, Guoyou Wang, Ran Wang, Anna Zhu,” Contour descriptor based on space symmetry and its matching technique”, Institute for Pattern Recognition and Artificial Intelligence, Optik, Vol. 124, No. 23, pp. 6149– 6153, 2013.

Evaluating the Applicability of a Social Content Management Framework: A Case Analysis

Wan Azlin Zurita Wan Ahmad[1,] Muriati Mukhta[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Yazrina Yahya[3]
Faculty of Economy and Management
Universiti Kebangsaan Malaysia
Bangi, Malaysia

***Abstract*—Social media platform plays an important role in engaging customers. The social content resulting from social media interactions between the organisations and the customers need a proper management. Therefore, in this work, a framework for social content management is introduced to support the management of social content. This framework is developed based on two main concepts. The first is the existing concepts that are present in content management, whilst the second concept is derived from the theory of service science. This approach is adopted to cater for existing concepts in enterprise content management, that are relevant to social content management and also to cater for the concept of value co-creation which forms the basis of engagement between the organisations and the customers. The applicability of the proposed social content management framework needs to be evaluated in order to determine the extent of its applicability in practical situations. Therefore, the main focus of this article is to report the usability of the proposed framework against the practices of the government agencies of Malaysia in managing the social content. The evaluation method used is based on the score of system usability scale. The results from the evaluation revealed that the proposed framework is usable and is deemed practical to be used in organisations.**

***Keywords*—Service science; social content; social content management; social media; value co-creation**

I. Introduction

Social media as a new platform, is fast becoming an important tool in assisting organisations to formulate strategies and engagement initiatives between the organisations and their customers. Social media has also become a channel for the source of information for engagement [1], [2]. A study by [3] proves that social media is able to produce open innovation in policies and services offered by organisations, by taking into account the direct needs of the customers. Besides that, a research by [4] stated that the digital infrastructure such as a social media platform would foster digital innovation in the organisations. Social media interactions also could impact on the risk perception [5]. Based on this significance, content from the interactions between the organisations and the customers on social media platforms needs to be taken care of. Therefore, social content management is introduced to manage the social content resulting from social media interactions.

Social media facilitates value co-creation between the organisations and its customers which benefit both parties. Thus, value co-creation is an aspect that needs to be emphasized in the management of social content. Value co-creation is one of the main concepts in service science [6]–[9]. Hence, the social content management framework is formed based on a service science perspective (see Fig. 1). From the service science perspective, this framework (see Fig. 1) adopted the theory of service dominant logic (SD-L) and the DART model. With the application of SD-L, it highlighted the actors, identification of service, the application of resource, and the exchange of service [6]. It also give attention to institutions and institutional arrangement to govern the process and perform on holistic ecosystem [6]. Apart from that, the DART model, which emphasises on the components of engagement platform, experience domain and also a capable ecosystem heading towards co-creation of value and outcomes of all actors involve [10]–[12]. The DART model has similarity with the process in content lifecycle which is the factor under the service exchange [13].

Detailed explanations of the framework in Fig.1 is stated in earlier research by [13] and [14]. Briefly, Fig. 1 shows the relationships between the elements and factors that formed a social content management based on science service perspective. A social content management is the dynamic process of collecting social content, resulting from active participation between organisations and customers on social media platforms. These interactions are supported by resource integrations and service exchanges. This dynamic process should be controlled by related institutions and institutional arrangement that occur within a good service ecosystem. The actors involved in social content management could be divided into two groups, namely, the organisations and the customers. For the purpose of managing content, three levels of organisational management are involved, namely, top management level, tactical level management (namely, middle management), and operational level management. These three levels have different roles in managing social content. The top management level plays the role of determining the policies and direction of social content management. The tactical level management is involved in managing social content and conducting analytical analysis. The operational level management is directly involved in the interaction process with the customers to collect and manage the social contents at the early stage. Social content is important for enabling service innovations because the customers' needs are obtained directly during the process of capturing contents through the active interactions between the organisations and the customers. Hence, social content should be managed to ensure that the organisations could produce accurate results based on input from social interactions, which could lead to innovations in the services offered.

Such framework needs to be evaluated to test it applicability in a real working environment. Therefore, this article is focusing on the evaluation of the framework. Four organisations from the government of Malaysia are chosen as case studies in this research to show the practicality of the proposed framework. The system usability scale (SUS) score is used to show the level of usability of the developed framework in the selected government agencies of Malaysia.

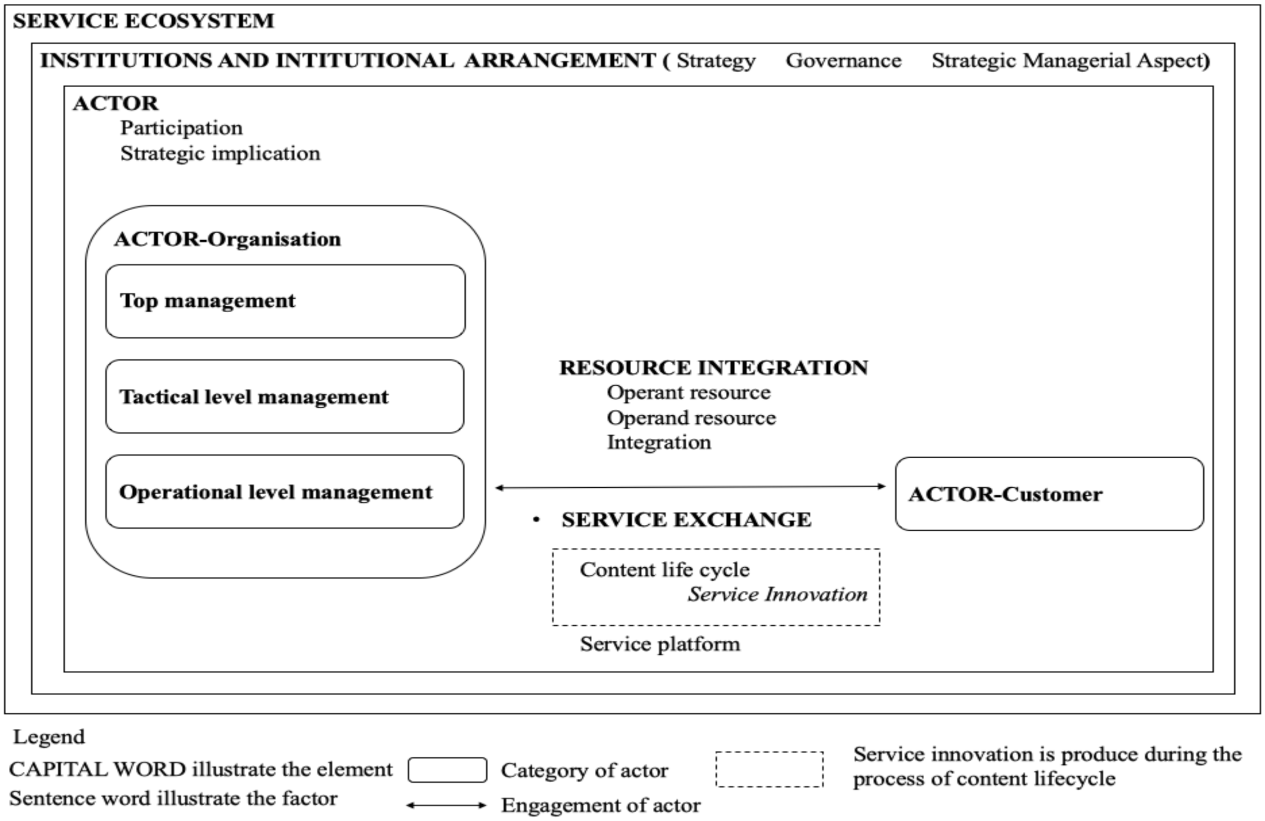

SERVICE ECOSYSTEM
INSTITUTIONS AND INTITUTIONAL ARRANGEMENT (Strategy Governance Strategic Managerial Aspect)
ACTOR
Participation
Strategic implication
ACTOR-Organisation
Top management
Tactical level management
Operational level management
RESOURCE INTEGRATION
Operant resource
Operand resource
Integration
ACTOR-Customer
• SERVICE EXCHANGE
Content life cycle
Service Innovation
Service platform
Legend
CAPITAL WORD illustrate the element
Sentence word illustrate the factor
Category of actor
Engagement of actor
Service innovation is produce during the process of content lifecycle

Fig. 1. Social Content Management Framework based on Service Science Perspective [13].

II. Material and Method

In this study, a case study is used to evaluate the framework. One of the questions asked in the case study protocol is "Based on the explanation on the framework and the understanding of the guideline, is the proposed framework useable for the organisation?". Therefore, the SUS score is adopted in this study to evaluate the usability of the framework. For this article, the research method that is adopted is illustrated in Fig. 2.

Further explaination of the research method is in the next sub-section

A. Protocol Development

In this study, there are three main activities in protocol development as follows:

1) Guideline development based on the framework:-A high-level guideline is developed based on the proposed framework in order to explain the framework in detail. The guideline guides the organisations to manage the social content based on the elements and factors that have been identified. Besides that, the guideline means to ensure the organisations are aware of how the framework works in assisting the organisations to manage the social content. Write-up of the guideline consist of the four chapters namely,

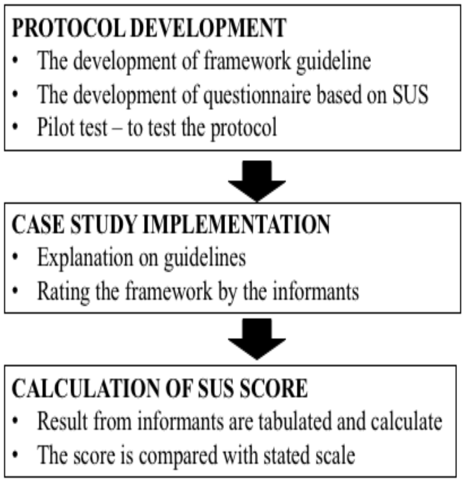

PROTOCOL DEVELOPMENT
• The development of framework guideline
• The development of questionnaire based on SUS
• Pilot test – to test the protocol
CASE STUDY IMPLEMENTATION
• Explanation on guidelines
• Rating the framework by the informants
CALCULATION OF SUS SCORE
• Result from informants are tabulated and calculate
• The score is compared with stated scale

Fig. 2. Research Method.

- First chapter is “Introduction”. This chapter explains on what is social content, social content management, the importance of managing social content. This chapter also explains on how to use the guideline.
- Second chapter is “Social content management framework based on service science perspective”. This chapter briefly explain on the framework as well as elements and factors that contribute to the development of the framework.
- Third chapter is “Guideline for organisation.” This chapter describe the high-level guideline on how the organisation should manage the social content. The explanation is given on all elements and factors that affect the management of social content as stated in second chapter. In addition, the checklist also provided in the guidelines according to the factors. Summary of checklist for the guideline is in Table I.
- Fourth chapter is “Conclusion” that conclude the guideline.

TABLE I. SUMMARY OF CHECKLIST FOR GUIDELINE

Factor	**Role of the organisation**
Participation Participation is about involvement and engagement of human resources (namely the actor) while managing social content.	1) Organisation need to ensure that the top management is involved in setting goals, formulating policies, setting resources, and making decisions on innovations based on the managed social content. 2) Organisation need to ensure that the tactical level management involves in controlling governance, analysing content, and maintaining the managed social content. 3) Organisation need to ensure that the operational level management involves in capturing the social content through social media interaction and managing the social content in the early stages. 4) Organisation need to ensure that the operational level management engage with the customers during the collaboration process to create opportunities for service innovation. 5) Organisation need to ensure that all actors, namely the organisations and the customers, be involved in the use of technology and the change management program as predetermined.
Strategic implication Strategic implications are the main effects that result from the participation of actors in social content management.	6) Organisation need to consider the impact of social content management to the organisations. 7) Organisation need to consider the impact of social content management to the customers. 8) Organisation need to give the customers the freedom of expression, namely the customers could determine the actual value based on their needs. 9) Organisation need to consider the impact of engagement between the organisation and the customers in creating opportunities for service innovation to the organisations.
Operant resource Operant resource is a dynamic resource in social content management that is meant to increase the competitiveness of organisations (referring to the skills, capabilities, knowledge)	10) Organisation need to have the skills and knowledge in: a) Formulating corporate strategy for social content management. b) Managing the social content within the content lifecycle. c) Managing the designated technology for social content management. d) Creating a collaborative environment while socialising with the customers on social media platform. *e)* Decisions making process on service innovation based on the managed and analysed social content.
Operand resource Operand resource is a static resource involved in social content management (such as tools, technology, budget, manpower).	11) Organisation need to be equipped with: a) Appropriate technology to manage the social content. b) Hardware with appropriate configuration such as backup configuration to manage the social content. c) Software with suitable capabilities such as smart search, support versioning, and customisation to manage the social content. d) Dedicated repository for storing the social content. e) Sufficient budget to manage the social content. f) Optimum number of manpower to manage the social content.
Integration Integration is the relationship between various sources namely human resources and content assets in social content management.	12) Organisation need to integrate: a) Operand resource with appropriate operant resources. For example, the use of technology (which reflect the operad resource) to manage social content with appropriate skills (which reflect the operat resource). b) Repositories that store the social content with other applications in the organisation to provide relevant inputs. c) Software for social content management with other applications. For example, integration of the software for social content management with MS Word and email to facilitate the operational task.
Content lifecycle Content lifecycle is the process of monitoring social content starting from capturing up to the maintaining, to support value co-creation process between organisations and customers.	13) Organisation need to capture the social content through interaction, namely from the dialogue during the socialisation process between the organisations and the customers on social media platform. 14) Organisation need to manage the social content that could be accessed by stakeholders and other applications using suitable resources. 15) Organisation need to analyse the social content for service innovation to obtain appropriate input based on the captured social content that suit with the organisation’s objective. 16) Organisation need to ensure that the result of the analysed social content being featured in an understandable format to the stakeholders. 17) Organisation need to maintain transparent state of the social content as easy to access, reusable, and always up-

	to-date.
Service platform Service platform is the space of interaction in managing social content to increase the efficiency of service exchange (such as systems).	18) Organisation need to ensure the service platform is: a) Flexible. b) Has a user-friendly interface. c) Facilitates relationships between different content categories. For example, to improve the organisation policy with the input from the customers, there is a need to provide the content from different categories that are tied together, namely from the category of interaction (ie. comment section on social media platform) and the policy documents.
Strategy Strategy is the planning, measures and methods in the management of social content to meet the needs of the organisations and the customers.	19) Organisation need to ensure that social content management strategies are designed with the aspects of, a) Actor-driven. b) Content-driven. c) Process-driven. d) Technology-driven e) The mechanism for managing the social content. f) The planning for change management program. 20) Organisation need to ensure that the management of social content is at the optimum level. 21) Organisation need to ensure that the social content is captured and analysed conforming to the organisational objectives. 22) Organisation need to ensure that challenges in managing social content are reduced to highlighting value in the designated strategy.
Governance Governance is an administrative routine for controlling the social content management in order to ensure the integrity of the content, which involves various human resources and content assets.	23) Organisation need to comply with existing policies based on organisational objectives. For example, the management of social content need to comply with the Organisation's ICT Security Policy to ensure the security aspect of managing the social content. 24) Organisation need to create new policies that are appropriate for managing social content. For example, the administrative instructions to facilitate the process of managing the social content. 25) Organisation need to have a sound governance structure such as committees or specific unit in the organisation to oversee the management of social content. 26) Organisation need to set a clear role to all actors involved in managing the social content.
Strategic managerial aspect Strategic managerial aspect is the acceptance of the actors to changes in technology, administration and content management methods and the increase in the level of competence of actors.	27) Organisation need to get the commitment of all actors involved. 28) Organisation need to have a robust change management program namely, a) Appropriate training programs for all actors involved such as analytical and soft skills training. b) Awareness programs at every level of management in the organisation. c) Awareness and engagement programs with the customers.
Service ecosystem Service ecosystem is a holistic environment that allows the value co-creation process to take place in the social content management.	29) Organisation need to ensure that social content management considers the organisational workflows. 30) Organisation need to consider a good project management mechanism to facilitate the development of innovative services based on the analysed social content. 31) Organisation need to consider the need for risk mitigation. 32) Organisation need to consider the management of social content perform in a good and conducive environment. 33) Organisation need to emphasise on active collaboration between various actors towards providing service innovation. 34) Organisation need to promote knowledge sharing based on the managed social content.

2) Questionnaire development based on system usability scale:- SUS is a scale that was introduced by [15]. It could be applied to measure the usability of the system in an organisation [15]–[17]. The SUS provides a simple and easy administrative questionnaire as well as produce reliable results with a small sample size of informants [15]. Furthermore, in a previous study, [18] adapted the SUS score in order to assess the usability of the framework in the organisation. Based on [18], the original questionnaire is amended, namely by changing of term “system” to the “framework”. Hence, this study applied the SUS score to evaluate the usability of the proposed framework. The template of questionnaire is based on SUS as stated in Table II. The SUS questionnaire consists of subjective evaluation with ten items measured. It is based on a Likert scale of five, namely from “1-strongly disagree” to “5-strongly agree”.

3) Pilot test:- The objective of the pilot test is to review the suitability and accuracy of the developed protocol. A pilot test was conducted at Agency X, a central agency in the government of Malaysia that plans and manages ICT implementation. This agency also manages and implement the Project X, a project that provides convenient access to government services through a single point of contact. Project X captures citizen feedbacks and comments on all platforms including the social media. The protocol is improved based on comments from the informants in Agency X.

TABLE II. TEMPLATE OF QUESTIONNAIRE [18]

Please tick (/) in the column provided	1	2	3	4	5
1. I think I would like to use this framework frequently					
2. I found the framework unnecessarily complex					
3. I thought the framework was easy to use					
4. I think that I would need the support of a technical person to be able to use the framework					
5. I found various elements and factors in this framework well integrated					
6. I thought there was too much inconsistency in this framework					
7. I would imagine that most people would learn to use this framework very quickly					
8. I found this framework very cumbersome to use					
9. I felt very confident using this framework					
10. I needed to learn a lot of things before I could get going with this framework					

B. Implementation of Case Study

According to [19], there are four types of case study, namely (1) single (holistic), (2) single (embedded), (3) multiple (holistic) and (4) multiple (embedded). As this study is not a unique study, hence multiple case study is more appropriate than single case study as it could give a diversity in data findings. Besides that, analytical units are important in case study design because they are fundamental to the study conducted [19]. This study emphasizes the management of the social content in the organisations by considering the needs of the customers, therefore, the appropriate unit of analysis is the “Organisation”. Hence, this study applies to the third type of case study, namely the multiple (holistic) case study with the unit of analysis “Organisation”.

A selection of case study is based on certain criteria. In the Malaysian environment, the government's focus is to enhance the service delivery by giving priority to the citizen by taking into consideration the views of the citizen in enhancing the quality of services offered by the government [20], [21]. Therefore, the organisations in the public sector begin to consider social content as input to service innovation. This is based on the Internet User Survey 2016 by [22], 69.6% of Internet users in Malaysia use social media to obtain information in the public sector compared to other communication channels. In addition, referring to [23], the most popular social media platform in Malaysia is Facebook with the percentage of 41%.

Besides that, based on [19], the selection of case studies relies on two criteria, (1) access to data, and (2) access to more than one informant. In that regard, out of the ten agencies that got the highest “like” on Facebook, four public sector agencies of Malaysia were agreed to participate in the evaluation process using the SUS score. For the case study, the sessions were set with 11 informants from four different public sector agencies of Malaysia. The list of informants and their characteristics are listed in Table III.

TABLE III. INFORMANTS AND CHARACTERISTIC

Informant ID	Characteristic of informant	Session
Organisation A provides the education services to the citizen and safeguarding the welfare of teachers		
A1	• The head of ICT Department. • Top management level who involve in the decision-making process include decision on service innovation from the managed social content.	1 hour
A2	• Manager in ICT Department. • Tactical level of management in social content management. • Involves in the development of innovative service initiatives, as a result from the analysed social content.	1 hour
A3	• The head of social media unit who responsible in managing official social media platform. • Tactical and operational level of management.	1 hour
Organisation B leads the management and enforcement of road transport		
B1	• Executive in the human resource department. • Operational level of management in social content management.	45 minutes
B2	• The head of social media unit who responsible in managing official social media platform. • Tactical and operational level of management in social content management.	1 hour
B3	• Senior manager of ICT department. • Operational and tactical level of management in social content management.	1 hour
Organisation C responsible for the management and development of human resources in public services		
C1	• The head of social media unit who responsible in managing official social media platform. • Tactical and operational level of management in social content management.	45 minutes
C2	• Executive in social media unit. • Operational level of management in social content management.	45 minutes
C3	• Executive in ICT department. • Operational level of management in social content management. • Involves in the development of innovative service initiatives, as a result of the analysed social content.	1 hour
Organisation D responsible for fire and rescue		
D1	• Manager in the human resource department. • Operational level of management in social content management.	45 minutes
D2	• The head of social media unit who responsible for managing official social media platform. • Tactical and operational level of management in social content management.	45 minutes

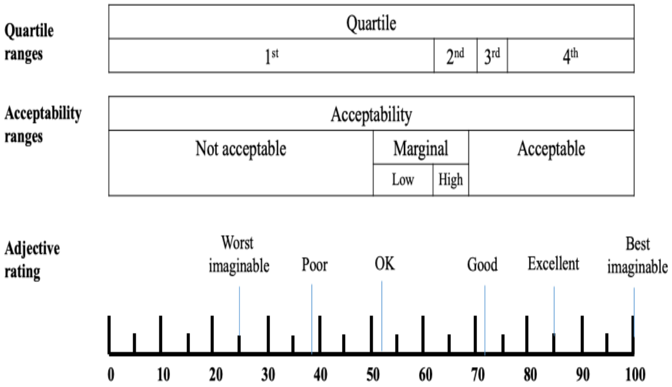


Fig. 3. SUS Score Value [15].

There were three main activities in the implementation of case study namely,

- individual informant was provided with the information on the framework and guidelines,
- individual informant was asked to evaluate the usability of the framework based on the SUS's questionnaire as stated in Table II, and
- the response was collected and analysed according to the predefined scale. The Bangor's SUS adjective scale based on quartiles as stated in Fig. 3 was referred as the scale in this study [15].

C. Culculation of SUS Score

For the purpose of calculating the SUS score, the steps set by [15] are as follows

- For odd numbered questions (Question number 1, 3, 5, 7, 9), the score of each item is calculated individually as,

Odd score = individual value for odd question–"1"

- For even numbered questions (Question number 2, 4, 6, 8, 10), the individual item score is calculated individually as,

Even score = "5"- individual value for even question

- Formula to calculate the value of SUS (range between 0 to 100) is:

SUS score value = Total score ("odd score" and "even score") * 2.5

- Mapping of SUS score values referring to the Bangor's SUS adjective scale (Refer Fig. 3).

III. Result and Discussion

The results of the analysis in the evaluation of the social content management framework, based on service science perspective, is given in Table IV. It is referring to the SUS score value rated by 11 informants from four different public sector agencies of Malaysia.

TABLE IV. Result of Analysis based on SUS Score Value

Organisation	Informant ID	SUS Score	Mean value of SUS Score per organisation
Organisation A	A1	90.00	87.50
	A2	87.50	
	A3	85.00	
Organisation B	B1	85.00	86.67
	B2	87.50	
	B3	87.50	
Organisation C	C1	87.50	89.17
	C2	95.00	
	C3	85.00	
Organisation D	D1	87.50	87.50
	D2	87.50	
Overall mean value of SUS score		**87.73**	

Referring to Fig. 3, the proposed framework is acceptable if the value of SUS score is obtained in the "third" and "fourth" quartiles as well as received the score of "70 and above". For individual organisation, all four different public sector agencies of Malaysia received the mean of SUS score value "above 85.0". Therefore, based on the Bangor's SUS adjective scale based on quartiles (Refer Fig. 3), the result for individual organisation shows,

- the quartile range for each public sector agency of Malaysia falls under the "4th quartile range",
- the acceptability ranges of the proposed framework for each public sector agency of Malaysia are at the "acceptable state", and
- complying to the Bangor's SUS adjective scale in Fig. 3, the framework could be stated as "excellence usability" for each public sector agency of Malaysia. This means the framework could be adopted in assisting each public sector agency of Malaysia to manage the social content.

Besides that, for the overall exercise, based on the calculation of SUS score value in Table IV, the overall mean value of SUS score is "87.73". Therefore, based on the Bangor's SUS adjective scale based on quartiles (Refer Fig. 3), the result for overall exercise shows,

- the quartile range for the assessment falls under the "4th quartile range",
- the acceptability ranges of the proposed framework are at the "acceptable state", and
- complying to the Bangor's SUS adjective scale in Fig. 3, the framework could be stated as "excellence usability".

Based on the finding of the means value of SUS score on each public sector agency of Malaysia and the overall exercise, it could be concluded that the proposed social content management framework based on service science perspective is acceptable to be applied in the organisations. It is also showing that the framework is able to guide the organisations to manage their social content.

IV. Conclusion

This article provides an overview of the evaluation of the social content management framework based on service science perspective on the working environment through the implementation of the case study. Agencies under the government of Malaysia are selected as case study because the current direction of the government is to consider the needs of the citizen directly in formulating service innovations offered by the organisation. The SUS score is adapted to show the acceptability state and the usability of the proposed frameworks in real working environment. From the findings, the proposed social content management framework based on service science perspective is acceptable to be applied in the organisations. For future work, the framework needs to be mapped with the actual execution of social content management processes in the respective organisations in order to identify room for improvements for individual organisations and also to perform cross-case analysis to find the similarities and differences between cases.

Acknowledgment

The study is supported by Research Grant FRGS/2/2014/ICT01/UKM/02/1, Ministry of Education Malaysia, Universiti Kebangsaan Malaysia and Public Service Department of Malaysia.

References

[1] Salman, M. A. M. Salleh, M. A. Yusoff, and M. Y. H. Abdullah, "Political Engagement on Social Media as Antecedent for Political Support among Voters in Malaysia," J. Komunikasi, Malaysian J., vol. 34, no. 2, pp. 152–165, 2018.

[2] M. A. Mohd Sani, S. Hassan, M. K. Ahmad, and Kartini Aboo Talib Khalid, "Generation Y ' S Political Participation and Social Media in Malaysia," Malaysian J. Commun., vol. 32 , no. 1, pp. 125–143, 2016.

[3] E. Loukis, Y. Charalabidis, and A. Androutsopoulou, "Promoting open innovation in the public sector through social media monitoring," Gov. Inf. Q., vol. 34, no. 1, pp. 99–109, 2017.

[4] S. Nambisan, K. Lyytinen, A. Majchrzak, and M. Song, "Digital Innovation Management: Reinventing Innovation Management Research in a Digital World," MIS Q., vol. 41, no. 1, pp. 223–238, 2017.

[5] F. Hilverda, M. Kuttschreuter, and E. Giebels, "Social media mediated interaction with peers, experts and anonymous authors: Conversation partner and message framing effects on risk perception and sense-making of organic food," Food Qual. Prefer., vol. 56, pp. 107–118, May 2017.

[6] S. L. Vargo and R. F. Lusch, "Institutions and axioms : an extension and update of service-dominant logic," J. Acad. Mark. Sci., vol. 44, pp. 5–23, 2016.

[7] C. K. Prahalad and V. Ramaswamy, "Co-creating unique value with customers," Strateg. Leadersh., vol. 32, no. 3, pp. 4–9, 2004.

[8] C. K. Prahalad and V. Ramaswamy, "The future of competition: co-creating unique value with customers," Penguin Books India, 2004.

[9] M. Mukhtar, M. N. Ismail, and Y. Yahya, "A hierarchical classification of co-creation models and techniques to aid in product or service design," Comput. Ind., vol. 63, no. 4, pp. 289–297, 2012.

[10] V. Ramaswamy and K. Ozcan, The co-creation paradigm. Stanford University Press, 2014.

[11] B. Leavy, "Venkat Ramaswamy – a ten-year perspective on how the value co-creation revolution is transforming competition," Strateg. Leadersh., vol. 41, no. 6, pp. 11–17, 2013.

[12] H. Mohamed, N. F. Elias, M. Mukhtar, Y. Yahya, S. A. Hanawi, R. Jenal, and W. A. Z. W. Ahmad, "Model Nilai Cipta-Sama dalam Sistem Pengukuran Prestasi," J. Pengur., vol. 45, no. 2015, pp. 155–163, June, 2015.

[13] W. A. Z. W. Ahmad, M. Mukhtar, and Y. Yahya, "Developing the Dimensions of a Social Content Management Framework," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 2018, October 2018.

[14] W. A. Z. W. Ahmad, M. Mukhtar, and Y. Yahya, "Managing Social Content: A Delphi Study," J. Pengur., vol. 53, no. 2018, pp. 1–20, October 2018.

[15] J. Brooke, "SUS - A quick and dirty usability scale," Usability Eval. Ind., vol. 189, no. 194, pp. 4–7, 1996.

[16] S. C. Peres, T. Pham, and R. Phillips, "Validation of the system usability scale (sus): Sus in the wild," in Proceedings of the Human Factors and Ergonomics Society, pp. 192–196, 2013.

[17] S. Borsci, S. Federici, and M. Lauriola, "On the dimensionality of the System Usability Scale: A test of alternative measurement models," Cogn. Process., vol. 10, no. 3, pp. 193–197, 2009.

[18] N. A. A. Bakar, "Enterprise architecture implementation assessment model for malaysian public sector," Universiti Teknologi Malaysia, 2017.

[19] R. K. Yin, Case Study Research Design and Methods, Fifth. SAGE Publications, Inc, 2014.

[20] EPU, "Rancangan Malaysia Ke-Sebelas 2016-2020," 2016.

[21] MAMPU, "Pelan Strategik ICT Sektor Awam Malaysia 2016-2020, 2016.

[22] SKMM, "Internet Users Survey 2016," 2016.

[23] SKMM, "Pengurusan Media Sosial Dalam Sektor Kerajaan," 2016.

An Efficient Rule based Decision Support System using Semantic Web Technology

Jawed Naseem[1]
Statistical Support and IT Unit
Pakistan Agricultural Research
Council Karachi, Pakistan

S. M. Aqil Burney[2]
College of Computer Science
IOBM University
Karachi, Pakistan

Nadeem Mehmood[3]
Department of Computer Science
University of Karachi
Karachi, Pakistan

***Abstract*—The Semantic Web technology is an efficient mechanism to query or infer knowledge on a global scale using the internet by providing logical reasoning through rule based system. In this paper application of semantic web technology is discussed in context of agriculture knowledge management and delivery. In agriculture, adoption of newly developed technology is essential to enhance crop production. However, timely dissemination of authenticated agriculture information for decision making at larger scale to diversified end user has always been a challenge due to several reasons. One of the reasons is storing and delivering agriculture knowledge in machine readable form. In this paper a frame work based on semantic web is presented for collection, storing and updating of agricultural information at centralized location and delivering knowledge through intelligent decision support system through semantic web. The frame work utilizes rule based system for querying information from agriculture knowledge base.**

Keywords—Agriculture information; semantic web; rule based system; intelligent DSS

I. Introduction

Agricultural research ensures food security of a nation. Developing agricultural technology and its dissemination to end user is the only option for sustainable agriculture development. The timely communication of newly developed research based technology to end-user for adoption is a challenge. The wider gap is always existed between development of technology and its adoption at farm level due to absence of efficient mechanism for delivery of information to end-user [9]. The conventional method of agricultural extension supported by use of information technology can fill the wider gap for access to agricultural information.

Crop production management is an essential component in agriculture. Monitoring of several activities are required during plant growth to ensure appropriate development of plant and maximize yield. Decision involved are planning pertaining to pre cultivation (selection of varieties, land preparation etc) monitoring during cultivation & growth (irrigation management, fertilizer application) and post harvest management. Feasible decision making requires access to latest relevant information to use appropriate technology, methods. Secondly, processing information with expert perspective and propose solution or action for end-user.

Presently computer science and information technology provide enormous opportunities for dissemination of information in every field of life [1]. From a standalone computer application or database on a machine to an organizational network or World Wide Web information can be process and disseminated quickly and efficiently. Several types of DSS has been developed in agriculture [7] and [11]. An ontology based domain rule development has been discussed by [8]. An intelligent plant disease diagnoses system has been presented by [10] and [11].

The ontology based knowledge engineering in agriculture is discussed by [2]. Burney has proposed a Bayesian network for wheat disease diagnoses in production management [12].

The utilization of semantic web for decision support has relevance particularly in agriculture. The Semantic Web shares many goals with Decision Support Systems (DSS) mainly to interpret information precisely to deliver relevant reliable and accurate information to a end user irrespective of time and location. While DSS have more specific goals, since the information need is targeted towards making a particular decision, e.g., selecting a particular crop variety for improved yield

In this research a case study of knowledge management of wheat production technology in Pakistan is under taken. A frame work and rule based system is presented for decision making in wheat production management through semantic web. Section I contain introduction, Section II discuss challenges and issues in wheat production management, Section III discuss semantic web technology, Section IV present an efficient rule based systems for framing queries and retrieval information through semantic web. Conclusion is discussed in Section V

II. Dynamics in Dissemination of Agriculture Technology

Sustainable agricultural development requires adaptation and incorporation of new technologies developed to enhance agricultural production. The technology may involve development of new varieties, agricultural production management, water management or crop protection. Adaptation of these technologies involves continuous updating of knowledge regarding a particular technology and processing of information by expert to deliver appropriate solution to problem in agriculture. High level human expertise in agriculture, like other disciplines of science, are not only scares but also costly. Beside this expert knowledge in agriculture require mass dissemination of information to large audience of end-users including policy makers, researcher, extension people and ultimately to farmer. Conventional means of communication of agricultural information have limited scope of knowledge acquisition, processing and instant delivery to end-user. Computer based systems [7] and soft computing models can provide an effective and efficient knowledge management [3]. Knowledge engineering in domain of agriculture comprises three basic component, Knowledge acquisition & representation, information processing and delivering of possible solution.

In a system where a number of experts contribute to knowledge base it is challenging to develop a consensus on various options. Therefore developing a mechanism for consensus of expert is essential in knowledge base system. [1] has proposed a system to increase group consensus which is based on consistency as well as measuring consensus together with maintaining the individual opinion of each expert.

A. Spatio Temporal Variation

Spatio temporal variation is one of the challenge in development of DSS in agricultural. Wheat production technology significantly varies across the country based on agricultural zones which are classified on the basis of environmental parameter. Research system has developed wheat varieties which are suitable to particular zone. The variation not only affects wheat quality and yield but risk associated with the attack of diseases and pest of wheat crop. Particular disease and pest are more vulnerable in particular environment. Like intensity of pest reduces by increase of temperature. Beside this within agricultural zone (spatial variation) temporal variation also come into focus. Yield of wheat is affected by the planting date of wheat variety. With late sowing of crop yield is reduced over time. Developing a computing model which captures this spatio temporal variation is essential.

B. Uncertainty

Crop growth in agriculture encompasses several uncertain situation, weather condition, timely access to agricultural input, spread of disease & pest and scheduling of competing crop. Predicting the uncertain outcome is essential for timely decision making in wheat growth. Whether timely harvest of previous crop will affect planting date of wheat? How temperature and humidity over next week will favor growth of pest? Whether required fertilizer will be available in market for timely application? These are the question which requires accurate answer (as much as possible) for sustainable growth of wheat. To some extend wheat yield is a game of chance also. The combination any of these uncertain situations can lead to make or break scenario of wheat crop. It is not very uncommon that uncertain weather condition may destroy wheat crop completely. No soft computing model for management of crop can succeeds without capturing inherent uncertainty in affecting factors. Luckily many methods in AI are capable of predicting uncertainty.

C. Knowledge Representation

Agriculture like any other discipline is huge canvas of knowledge comprising thousands of concepts and hundreds of relationships between those concepts. Representing knowledge in machine readable form is the basic element in developing computing system. However, agriculture is different in the sense that it contain common simple to understand concepts like suitable varieties along with highly technical information like agronomy, pathology and entomology etc. Agriculture had very varied audience; on one end are the farmers mostly accustomed with conventional terminology in local language and on the other end is sophisticated scientific community of researcher who are supposed to develop or update agriculture knowledge. Developing a common frame work which is capable of this varied semantic level of the audience is real challenge which is successfully addressed in this research. Development of Owl based ontology [13] is an option for knowledge representation in agriculture. The advantage of OWL base ontology is the fact that owl and RDF are effective tools of semantic web [6].

III. Semantic Web

The Semantic Web is the latest development for real time sharing of information on internet. Semantic web facilitate a common mechanism which allows sharing of data across different application, enterprise, and community. Semantic web expresses information in machine-readable form. It is the web that can be processed by machines in such a way that meaning of different concepts is commonly expressed. Technically semantic web is a giant data graph defined in RDF.

The common technology used in semantic web is OWL (Web Ontology Language) ontology expressed in RDF graph. OWL is a Semantic Web language designed to express knowledge about concepts and relations between concepts [13]. OWL is a logic-based computing language which may be used by computer systems. OWL ontology can be published in the World Wide Web and may refer to other ontology or other ontology use it [15]. The semantic web includes a stack of WC3 technology comprising OWL RDF, RDFS, SPARQL, etc. Basically Web Ontology Language (OWL) is a modeling language encompasses fast and flexible data modeling and efficient automated reasoning. OWL extends RDF (Resource Description Framework) and RDFS.

The semantic web use graph data model to store information and RDF is the format in which this information is written. The information contained in a Web resource is expressed in XML-based language using RDF for description. The resource can be knowledge base, a Web page, an entire Web site, or any item on the Web that contains information in some form. RDF can be used to define any of the resources. However in case of knowledge base or representation through ontology the resource may be a URI (uniform resources identifier). RDF is a tool of meta data description used for encoding, exchange, and reuse of structured data. The use of RDF is domain independent it has not generalized application without assumption about any particular domain, nor semantics is depended on domain.

The RDF [14] graph can be queried for knowledge search. SPARQL Protocol and RDF Query Language (SPARQL) is the language to query a RDF graph. Its syntax is almost similar to SQL the main difference is SQL can query relational database while SPARQL can query NOSQL or database graph. In fact a SPARQL query can also be executed on any database that can be viewed as RDF via middleware. For instance, a relational database can be queried with SPARQL by converting Relational Database to RDF. SPARQL is capable of querying diverse data source no matter whether data is expressed in RDF or in a form which can be converted in to RDF both can be queried through SPARQL. SPARQL have capability to query several graph pattern using conjunction or disjunction attribute of the graph. The output of SPARQL queries can be results sets or another RDF graphs.

IV. Frame Work for Semantic Web Wheat Production Technology

In agriculture knowledge generation is scattered over many sources and by many roles. Traditionally in Pakistan agricultural knowledge is disseminated through conventional means through extension department. Agriculture Knowledge generator (researcher) has limited access to disseminator (extension worker) or vice versa. Secondly, communication between extension worker and ultimate user (farmer) is fragile. This wider gap can be filled only through unconventional means. A central knowledge base accessible to all stake holders can be achieved through web based technologies where authorized user with defined roles are provided with tools to update the required information. Centralized web based knowledge management and updating ensure timely availability of authenticated information.

We have suggested a frame work (Fig 1) for development of collaborative updating of knowledge to central repository by a panel of authorized person from research and non research establishment using common electronic tools of semantic web. This frame work not only ensures data authentication but remove data redundancy.

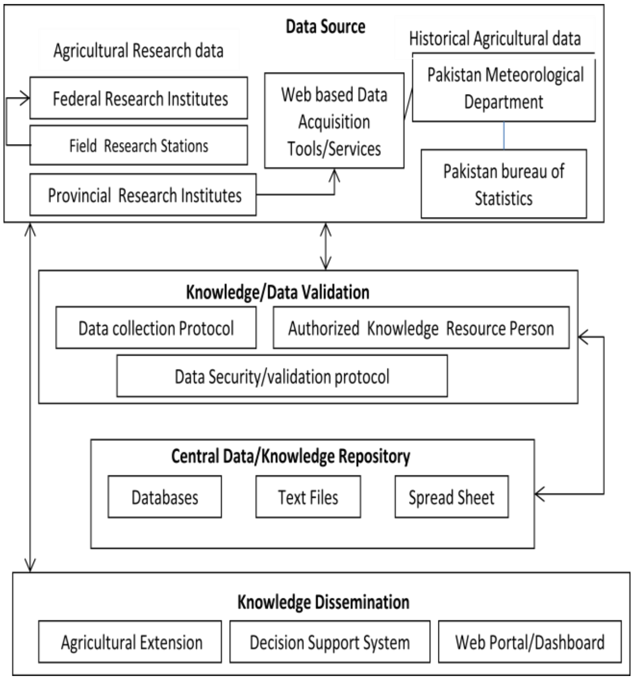

Data Source
Agricultural Research data
Historical Agricultural data
Federal Research Institutes
Field Research Stations
Provincial Research Institutes
Web based Data Acquisition Tools/Services
Pakistan Meteorological Department
Pakistan bureau of Statistics
Knowledge/Data Validation
Data collection Protocol
Authorized Knowledge Resource Person
Data Security/validation protocol
Central Data/Knowledge Repository
Databases
Text Files
Spread Sheet
Knowledge Dissemination
Agricultural Extension
Decision Support System
Web Portal/Dashboard

Fig. 1. Frame Work for Semantic Web of Wheat Production Technology.

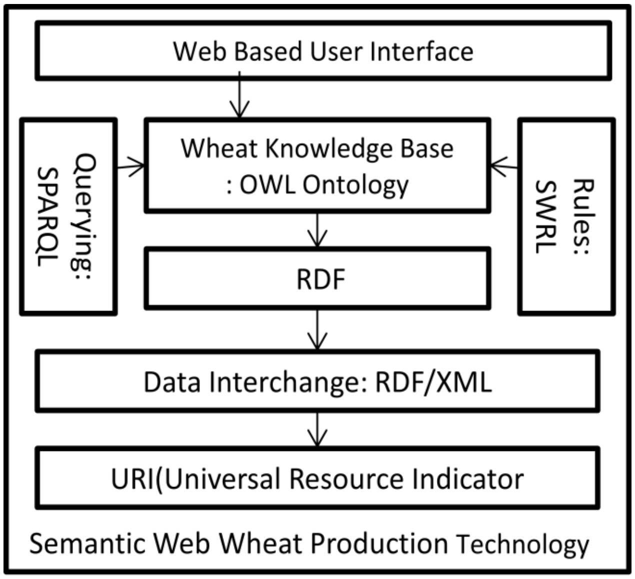

Web Based User Interface
Querying: SPARQL
Wheat Knowledge Base : OWL Ontology
Rules: SWRL
RDF
Data Interchange: RDF/XML
URI(Universal Resource Indicator
Semantic Web Wheat Production Technology

Fig. 2. Semantic Web Wheat Production Technology.

The second step of this process is mapping of the suggested frame work into a machine readable form to create a central knowledge repository which is achieved through semantic web (Fig 2). The frame work is used to collect information from relevant research institutes and department which may in the shape of databases, flat text file or other medium.

The semantic web comprises knowledge base development using owl ontology and mapping it through RDF [4]. Reasoning the KB using WSRL [5] rules or SPARQL query [14] and delivering the information through URI. On the top of it is the web based user interface to interact with the end user.

The user formulate query using interface and solution to problem is communicated by identifying relevant URI in the knowledgebase.

V. Rules Based System for Querying Semantic Web

Rules based expert system and rules are structured way of reasoning and classifying information. Production rules in the form of if & then clause are used to define or apply constrain in declarative manner. Domain rules are structured in informal, semi informal and formal ways. However, in knowledge base rules are expressed in formal system. Informal statements in natural language are transformed into formal language or rule execution language.

In semantic web rules can be developed by using Semantic Web Rule Language (SWRL)[5]. SWRL represent rules as well as define logic of extraction by combining OWL DL with the Rule Markup Language.

SWRL develop Horn-like rules by extending the set of OWL axioms which represent logical extraction from knowledge base. SWRL follow traditional rule construction in the form of an implication between an antecedent (body) and consequent (head). So SWRL rule can be read as: whenever the conditions declared in the antecedent are true, then the conditions specified in the consequent must also be true. In OWL ontology abstract syntax contains a sequence of axioms and facts. Each Axiom may be of different kinds, e.g., subClass axioms or equivalent Class axioms. Then axioms are extended to rule.

axiom::= rule

A rule based on axiom consists of an antecedent (body) and a consequent (head), each of which consists of a set of atoms which may be possibly empty. Therefore, a URI reference can be assigned in axiom which could serve to identify the rule.

Rule:: = 'Implies(' [URIreference] { annotation } antecedent consequent ')'

antecedent::= 'Antecedent(' { atom } ')

consequent::= 'Consequent(' { atom } ')

Where both antecedent and consequent can be conjunctions of atoms written in the form

a1 ∧ ... a2 ∧

While variables are indicated by using a question mark (e.g., ?x). The advantage of SWRL rule is it can be represented in human readable form. In the abstract syntax the rule would be written like

Implies(Antecedent(Property1(I-variable(x1)I-Variable(x2))

Property2 (I-variable(x2) I-variable(x3)))

Consequent(Property3(I-variable(x1)I-variable(x3))))

The following rule describes selection of agricultural zone on the basis of district and province. When district is y and province is x agricultural zone is z

District (?y) ^ province(?x) -------> zone(?z)

Where LHS represent antecedent and RHS represent consequent. It may be noted that antecedent is conjunction of atom using and properties while consequent is atomic.

In AI several methods can be used for formal expression or rules like SQL (Structured Query Language), ECA, predicate logic and propositional logic. In wheat production expert system ontology based predicate logic and axioms are used to define rules for extraction and updating of information from wheat knowledge base. Developing ontology based rule is three steps process [8] involving

- Express rule in informal natural language
- Express rule using ontology concepts and relationship
- Express rule in formal predicate logic

In wheat production management expert system conditions and actions are proposed by agriculture expert in natural language which is expressed into ontological expression and finally into SWRL syntax

Informal: if soil is loamy and soil condition is weak then apply 3 bag of NPK or 2 bag of DAP at the time of sowing.

Using ontological expression rule can be expressed as

<Soil> < has_type> <loamy> and
<Soil> <has_condition> <Weak>
then <Fertilizer> <has_name> = ”NPK” <hasQuantity> = ”2 bags”, Or
Fertilizer> <has_name> = ”DAP” <hasQuantity> = ”3 bags”,
and <Fertilizer> <hastimeof application> = ”at sowing”

The ontological expression is transformed into rule using SWRL syntax. So formal expression in predicate logic will be

SoilType(?x) ^ SoilCondition(?y) ---→ Fertilizer(?z)
Fertilizer(?f) ^ hastimeofapplication(?t) ---→ has_quantity(?q)

The terms in bracket represent variable which are replaced by named individual. The SWRL rules based knowledge can be queried using DL or SPARQL query. SPARQL query has sql like syntax as follows

SELECT ?subject ?object

WHERE { ?subject rdfs:subClassOf ?object }

So using the query syntax quantity of fertilizer for weak soil can be retrieved using following query

SELECT ?Fertilizer ?quantity

WHERE { soil ? has_condition? weak }

The three step process seamlessly transform parameter based question into to appropriate machine readable rule which extract required information from knowledgebase and delivered it through relevant semantic web URI

VI. Conclusion and Discussion

This paper presents an intelligent DSS based on semantic web for facilitating the decision making in wheat production management. A frame work is proposed for deployment of DSS through semantic web. Effective decision support is provided with the development of ontology driven rule based system by considering both the characteristic of the problem and the requirements of the decision maker.

The applicability of the frame work is demonstrated through the example of decision support for wheat production management. The example shows that the proposed DSS framework has a number of advantages over data driven desktop based DSS. The proposed DSS has three specific advantage first its ensure wider dissemination through internet, secondly it represent knowledge in logical way by using ontology, thirdly it is intelligent. The DSS is intelligent in the sense that it not only effectively retrieves required information by efficient query but also infer the knowledge which is not physically present in the knowledge base by extending logical reasoning of an expert through build in rule based system. In domain of agriculture updating of knowledge at wider scale through many sources is equally essential as retrieving of information. The ontology based semantic web resolve this issue effectively by providing a mechanism following logical updating and resolving conflicting information at real time. The knowledge is not updated until it meets required protocol.

Application of semantic web in agriculture is particularly become more relevant as generation of knowledge is scattered over many sources so updating of knowledge become more feasible with minimum IT resources available which is one of the main constraint in under developing countries. This paper proposes a system which can be implemented economically at wider scale. The proposed frame work ensures scalability and interoperability as it can be modified for other crops also.

The proposed system has one limitation that, as the new knowledge is generated obsolete information should be discarded automatically as removal of outdated information is essentially important. The next research topic related to this study is to develop an automated process which can periodically identify obsolete information, validate in context of new information and update it accordingly.

REFRENCES

[1] Alonso S., E. Herrera-Viedmab, F. Chiclanac, F. Herrerab, 2010, "A web based consensus support system for group decision making problems and incomplete preferences" Information Sciences, Volume 180, Issue 23,

[2] Burney, S. M Aqil, Jawed Naseem, 2018," Knowledge Engineering in Agriculture: A Case Study of Soft Computing Model for Wheat Production Management", International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 2, pp 56-62.

[3] Burney, S. M Aqil, Nadeem Mehmood, 2012 "Generic Temporal and Fuzzy Ontological Framework, (GTFOF) for Developing Temporal-Fuzzy Database Model for Managing Patient's Data, Journal of Universal Computer Science, vol. 18, no. 2 (2012), 177-193.

[4] Canda, K.Selcuk, "Resource Description Frame work: Metadata and Its Applications", Arizona State University, http://citeseerx.ist.psu.edu

[5] Connor, Martin O et al , "Querying the Semantic Web with SWRL", Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA 94305 martin. oconnor @stanford.edu.

[6] Darai, D.S S Singh, S Biswas, 2010, "Knowledge Engineering an overview", International Journal of Computer Science and Information Technologies, Vol. 1 (4) , 230-234.

[7] Hoogenboom, G P.W. Wilkens, P.K. Thornton, et al., 1999. "Decision support system for agro technology transfer" v3.5. In: DSSAT version 3, vol. 4 University of Hawaii, Honolulu, HI, pp. 1-36.

[8] Kalibatiene, Diana et al , 2010, "Ontology-Based Application for Domain Rules Development" Scientific Paper University of Latvia, Vol 756, Computer Science and Information Technologies pp 9-32.

[9] Khan, Ghanzafar, Ali, 2010, "Present and prospective role of electronic media in the dissemination of agricultural technologies among farmers of the Punjab, Pakistan, Ph.d theses, university of Agriculture Faisalabad, Pakistan.

[10] Kolhe Savita, Raj Kamal, Harvinder S. Saini, G.K. Gupta, 2011, "A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops", Computers and Electronics in Agriculture Volume 76, Issue 1, Pages 16-27.

[11] Magarey, R.D.; Travis, J.W.; Russo, J.M.; Seem, R.C. & Magarey, P.A. 2002. Decision Support Systems: Quenching the Thirst. Plant Disease, Vol. 86, No. 1, pp. 4-14,

[12] Naseem, Jawed , S. M Aqil Burney, 2018, "Decision Making in Uncertainty: A Bayesian Network for Plant Disease Diagnoses" International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 2, February 2018.

[13] Natalya F. Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology" Stanford University.

[14] Zheng, Weiguo et al, "Semantic SPARQL Similarity Search Over RDF Knowledge Graphs", http://www.vldb.org/pvldb/vol9/p840-zheng.pdf, pp 840-851.

[15] Zhi Ping Ding, 2011, "The Development of Ontology Information System Based on Bayesian Network and Learning," Advances in Intelligent and Soft Computing", Volume 129, , Pages 401-406.

A Type-2 Fuzzy in Image Extraction for DICOM Image

D. Nagarajan[1], M.Lathamaheswari[2]
Department of Mathematics
Hindustan Institute of Technology & Science
Chennai-603 103, India

J.Kavikumar[3], Hamzha[4]
Department of Mathematics and Statistics
Universiti Tun Hussein Onn
Malaysia

***Abstract*—Eradication of a desired portion of an image is a very important role in image processing and is also called feature extraction. This is mainly concern about reducing the number of possessions required to portray a large set of data and also reduce memory space requirement and power of data processing. Perfectly optimized feature extraction is an essential process for an effective design construction. Though there are many tools are available for extracting a feature, Type-2 Fuzzy Logic plays a vital role in producing good results. In this paper, weighted arithmetic operator is proposed using Yager triangular norms and proved the properties of the triangular norms using proposed operator. Also, the paper relates the properties to feature extraction. Also Brain has been extracted from patient MRI DICOM image using MATLAB based on Type-2 Fuzzy setting.**

Keywords—Feature extraction; MRI image; type-2 fuzzy; MATLAB; triangular norms; mathematical properties

I. Introduction

Image processing is a mapping from image points to a new value by dealing a single point from original image and thereafter it will continue for group of points. Feature extraction is nothing but shape eradication which analyzes the contest of low level data to a known design of a desired shape Also it indicates verdict their location, direction and size. Minor particulars of the image such as dots and lines are called low level features whereas high level features are assembled on the elite of low level features to catch objects and bigger shapes in the image [1-4]. This shows the important steps involved in feature extraction.

In image processing number of points and their range makes a specific effect and the main components are grayscale, resolution, color, dynamic range and storage. Image processing and feature extraction is equitable to the illumination of the matching point in the location and generally its value is determined from the output. Image is a matrix of picture elements in square shape and is described by $M \times M$ n-bit picture elements i.e., pixels. Here M is the number of points through the axes and n manages the number of values for brightness. Adopting n bits provides a range of $2n$ values, ordering from 0 to $2n-1$ which is the brightness level and is generally presented as white and black. Smaller value of n represents the decrement in image contrast [5, 9-10]. This part expresses the color contrast of the image.

Feature extraction is also meant for increasing pixel brightness to find the desired part of the image. In extracting a feature invariance property is to be strictly followed as this process should not vary according to the specified conditions it also implies the reliability and stability of the shape which is extracted. Hence it is necessary to choose the technology for feature extraction where there is a control over the parameters [11-12]. Hence choosing the technology for image extraction is very important to get better results.

The passion histogram exhibits how particular brightness levels are involved in an image and contrast of the image is calculated by the range of brightness levels. This technology plans the number of pixels with a specific level of brightness contrary to the level of brightness. All these contrary things can be dealt by fuzzy logic as it deals with uncertainty well. Generally images have uncertainty in the desired part and clear shape of it. Type-2 fuzzy logic plays an efficient role as it deals with upper and lower membership functions and foot print of uncertainty (the area between upper and lower membership functions) and its length represents uncertainty level of the image or part which is to be extracted [13]. This is the role of fuzzy logic under type-1 and type-2 fuzzy environments.

These technologies are the mathematical systems which provide an outstanding direction for better understanding about the process. Mathematica, Mathcad, MATLAB and Maple are some of the popular tools. Where in MATLAB, by giving instructions at the mandate line, the procedure will be operated and the outputs can be displayed as surfaces, graphs and images. Hence MATLAB under Type-2 fuzzy setting will provide a desired result. In this work, using Type-2 Fuzzy MATLAB, brain is extracted from MRI image. Even though the convenient measures the performance for the credential design analysis are not recognizable and difficult to derive, it helps to design a system which improves its performance measures for training data set and forecast the performance of the system for testing data set [1,14].

Feature extraction of the image to an observable level is an essential key in advance of Content Based Image Retrieval systems. From texture, color and structure of the image, the low level optical features of an image can be separated and can be used while recovery to correlate concern image and other images in the database [12, 18]. This is the stage of acquiring information about image objects which is to be analyzed. On the basis of both quantitative and qualitative reasoning, the features may be determined where the ideas might be hypothetical from the expert which is modified into quantitative values. Color image is one of the universally used features due to its stability, efficiency and computational simplicity [12, 15-16]. Hence color images have potential of clarity of the features which are to be extracted.

A color image based on mathematics is called a digital image and it consists of color information for every picture element. It is also a binary image which has only two possible values for every pixel and is reserved as a single bit 0 or 1. To design a binary image, a threshold intensity value needs to be selected. Pixels with greater intensity than the threshold value are switch to 0 (black) whereas when the intensity value less than the threshold level are switch to 1 (White) and hence the image is converted as a binary image. Gray scale images are having a range of darkness without possible color and used as a fewer information needs to be contributed for each and every pixel. The calculation of mathematical captions could contribute more information about the parameters of morphology but they are not interpreted easily [17-20, 30]. Therefore the procedure for morphology is supposed to be taken care for getting a good feature.

Medical images are generally uncertain in nature. Though there are many methodologies are available to extract the feature from the image, Fuzzy logic (Type-1) helps to extract the feature in an efficient manner but the membership functions are crisp which lies between 0 and 1. MATLAB on Type-2 fuzzy setting provides an optimized result as it handles more uncertainties based on the Footprint of uncertainty, the area between upper and lower membership functions.

From the previous proposed solutions image extraction has not been done using interval type-2 fuzzy logic for feature extraction of the brain from patient MRI. This is the shortcoming of the previous studies and the motivation of the present study. Throughout the paper type-2 fuzzy has been considered as interval type-2 fuzzy environment.

In this paper, the mathematical properties of triangular norms has been proved as it represents the essential qualities of image processing and brain is extracted from patient MRI which is taken from our experimental data using MATLAB and provided the 3D image of an extracted brain with the key components such as DICOM image of a patient MRI, interval type-2 mat lab coding for feature extraction.

II. Review of Literature

The authors of, [1] proposed the idea of linguistic variable and its application in the field of approximate reasoning. The researcher in [2] proposed different classes of fuzzy operators. The researcher in [3] proposed novel aggregations operators under probabilistic fuzzy environment. The researcher in [4] reviewed aggregation connectives under fuzzy environment. The researcher in [5] introduced theory of t-norms and inference methods under fuzzy setting. The researcher in [6] designed fuzzy systems and derived aggregation operators. The researcher in [7] discussed about imprecise reasoning for interval based data with the support of fuzzy and rough sets. The researcher in [8] proposed aggregation operators in detail and applied them video querying.

The researcher in [9] introduced fuzzy image processing using Dubois and Prade aggregation operators. The researcher in [10] proposed a computer based system on hand written records to hold forensic studies. The researcher in [11] proposed means with weight using triangular co norms. The researcher in [12] introduced elementary minimum and maximum operational laws for fuzzy numbers. The researcher in [13], the author aggregated the information collected using aggregation operations. The researcher in [14] applied fuzzy relational equations for Lossy image compression and reconstruction. The researcher in [15] proposed OWA operators with imprecise weights based on type-1 fuzzy. The researcher in [16] analyzed aggregation functions. The researcher in [17] introduced exact computations of protracted logical operations based on uncertain truth values. The researcher in [18] examined and compared different approaches over edge detection.

The researcher in [19] applied morphological operators in image analysis on uninorms. The researcher in [20] proposed novel aggregation operators for the method of active learning. The researcher in [21] explained and derived aggregation operators in detail. The researcher in [22] proposed triangular interval type-2 aggregation operators using Frank triangular norms and applied them in a decision making method. The researcher in [23] proposed fuzzy metric spaces. The researcher in [24], the author applied idea of fuzzy methodology in medical image processing. The researcher in [25] examined collection of information and the related aggregation operators. The researcher in [26] reviewed recent year applications of image processing under type-2 fuzzy.

In [27] the imitation of edge detection of an image using MATLAB with fuzzy logic is done. The researcher in [28] reviewed the role of type-2 fuzzy in the field of Bio medicine. The researcher in [29] proposed edge detection method for a digital image using fuzzy logic. The researcher in [30] proposed a methodology for edge detection for a DICOM image using aggregation operators under type-2 fuzzy. The researcher in [31] implemented a methodology for image fusion using intuitionistic fuzzy logic. The researcher in [32] reviewed about fuzzy controllers to sustain the stability of the system using type-2 fuzzy. The researcher in [33] analyzed surface of the material on curve features of digital images using fuzzy logic. The researcher in [34] proposed 3D version of brain visualization using machine learning. The researcher in [35] proposed block processing and edge detection on DICOM image. The researcher in [36] introduced denoising of the image using LU decomposition method and feature extraction using GLCM.

From this review it is found that there is no contribution of research for image extraction from a DICOM image using interval Type-2 fuzzy logic. This is the motivation of the present work.

III. Basic Definitions

The following basic concepts are given for the better understanding of the paper.

A. *Aggregation Operator [22]*

Let $(M_{\alpha})_{\alpha \in [0,1]}$ be a group of aggregation operators (AOs) which is non-decreasing. If A is an AO then

$$M_A : \bigcup_{n \in N} [0,1]^n \to [0,1].$$

Triangular Interval Type-2 Fuzzy Set (TIT2FS) [22]

The membership function (MFs) are developed using triangular fuzzy number in IT2FS called TIT2FS. In IT2FS, upper and lower MFs represented by a triangular fuzzy number $\overline{M} = \left\langle [\underline{l_M}, \overline{l_M}], c_M, [\underline{r_M}, \overline{r_M}] \right\rangle$ called TIT2FS and are defined by

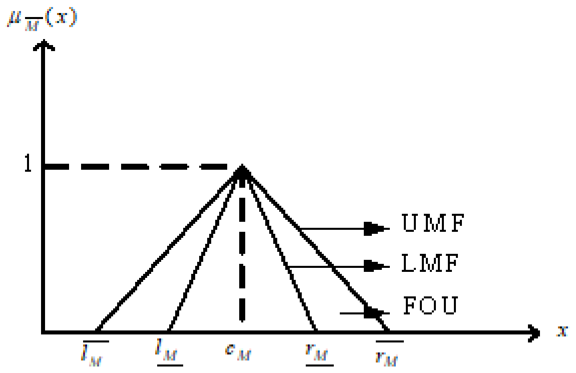


Fig. 1. TIT2FS.

$$LMF_{\overline{M}}(x) = \begin{cases} \dfrac{x - \overline{l_M}}{c_M - \overline{l_M}} & , \quad \overline{l_M} \le x < c_M \\ 1 & , \quad x = c_M \\ \dfrac{x - \underline{r_M}}{c_M - \underline{r_M}} & , \quad c_M \le x < \underline{r_M} \\ 0 & , \quad otherwise \end{cases} \tag{1}$$

$$UMF_{\overline{M}}(x) = \begin{cases} \dfrac{x - \underline{l_M}}{c_M - \underline{l_M}} & , \quad \underline{l_M} \le x < c_M \\ 1 & , \quad x = c_M \\ \dfrac{x - \overline{r_M}}{c_M - \overline{r_M}} & , \quad c_M \le x < \overline{r_M} \\ 0 & , \quad otherwise \end{cases} \tag{2}$$

where $\underline{l_M}, \overline{l_M}, c_M, \underline{r_M}, \overline{r_M}$ are the measuring points on TIT2FS satisfying $0 \le \underline{l_M} \le \overline{l_M} \le c_M \le \underline{r_M} \le \overline{r_M} \le 1$. If we consider X as a set of real numbers, a TIT2FS in X is called TIT2FN. The FOU is the area between lower and upper membership functions in figure 1. If $\underline{l_M} = \overline{l_M}, \underline{r_M} = \overline{r_M}$, then $UMF_{\overline{M}}(x)$ = $LMF_{\overline{M}}(x)$ for all the values of x in X, then the TIT2FS will become Type-1 case. Here FOU is the footprint of Uncertainty.

B. Ranking Formula for TIT2FN [22]

Let $\overline{M} = \langle [A, B], C, [D, E] \rangle$

where $A = \underline{l_M}, B = \overline{l_M}, C = c_M, D = \underline{r_M}, E = \overline{r_M}$ be the TIT2FN. The ranking value is defined by

$$Rank(\overline{M}) = \left(\frac{A+E}{2} + 1 \right) \times \frac{A+B+D+E+4C}{8} \tag{3}$$

C. Yager Triangular Norms [5]

$\underset{Y}{\otimes}$ is Yager product (T Norm) and $\underset{Y}{\oplus}$ is a Yager sum (T conorm) and are defined as follows.

$$r \underset{Y}{\otimes} s = \max\left(1 - [(1-r)^{\eta} + (1-s)^{\eta}]^{\frac{1}{\eta}}, 0 \right), \eta > 0, \ for\, all\ r, s \in [0,1]^2$$

(4)

$$r \underset{Y}{\oplus} s = \min\left(\left(r^{\eta} + s^{\eta} \right)^{\frac{1}{\eta}}, 1 \right), \eta > 0, \ for\, all\ r, s \in [0,1]^2 \tag{5}$$

D. Triangular Interval Type-2 Fuzzy Yager Weighted Arithmetic (TIT2FYWA) Operator [22]

Consider a set of TIT2FNs and the operator $TIT2FYWA_{\varepsilon} : \Omega^n \to \Omega$ is defined by

$$TIT2FYWA_{\varepsilon} \left\langle \overline{M_1}, \overline{M_2}, ..., \overline{M_n} \right\rangle = \varepsilon_1 \underset{Y}{\bullet} \overline{M_1} \underset{Y}{\oplus} \varepsilon_2 \underset{Y}{\bullet} \overline{M_2} \underset{Y}{\oplus} ... \underset{Y}{\oplus} \varepsilon_n \underset{Y}{\bullet} \overline{M_n}$$

and its weight vector is $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ and the sum of the weight vectors is equal to 1, when $\varepsilon = \left(1/n, 1/n, ..., 1/n \right)^T$, triangular interval type-2 fuzzy Yager weighted arithmetic operator will become triangular interval type-2 fuzzy Yager arithmetic averaging operator of dimension n and is defined by.

$$TIT2FYAA_{\varepsilon}(\overline{M_1}, \overline{M_2}, ..., \overline{M_n}) = \frac{1}{n} \underset{Y}{\bullet} \left(\overline{M_1} \underset{Y}{\oplus} \overline{M_2}, ..., \underset{Y}{\oplus} \overline{M_n} \right) \tag{6}$$

E. Triangular Interval Type-2 Fuzzy Yager Weighted Geometric (TIT2FYWG) Operator [22]

Let $\overline{M} = \left([\underline{l_{M_p}}, \overline{l_{M_p}}], c_{M_p}, [\underline{r_{M_p}}, \overline{r_{M_p}}] \right)$, $p = 1, 2, ..., n$ be a set of TIT2FNs. Triangular Interval Type-2 fuzzy Yager Weighted Geometric Mean Operator (TIT2FYWA), $\mathrm{TIT2FYWG}{:}\varepsilon^{\mathrm{n}} \to \varepsilon$ is $\mathrm{TIT2FYWG}_{\varepsilon}\left(\overline{M_1}, \overline{M_2}, ..., \overline{M_n} \right) = \overline{M_1}^{\underset{Y}{\bullet}\varepsilon_1} \underset{Y}{\otimes} \overline{M_2}^{\underset{Y}{\bullet}\varepsilon_2} \underset{Y}{\otimes} ... \underset{Y}{\otimes} \overline{M_n}^{\underset{Y}{\bullet}\varepsilon_n}$ .Here also sum of all weight vectors is equal to 1, when $\varepsilon = \left(1/n, 1/n, ..., 1/n \right)^T$, triangular interval type-2 fuzzy Yager weighted arithmetic operator will become triangular interval type-2 fuzzy Yager geometric averaging operator of dimension n and is defined by $TIT2FYGA_{\varepsilon}(\overline{M_1}, \overline{M_2}, ..., \overline{M_2}) =$

$$\frac{1}{n}\left(\overline{M_1} \underset{Y}{\otimes} \overline{M_2} \underset{Y}{\otimes}, ..., \underset{Y}{\otimes} \overline{M_n}\right)^{\dot{Y}\, 1/n} \tag{7}$$

F. Feature Extraction [18]

The task of feature extraction converts affluent content of images into different content features and a process of producing features to be used in the classification and selection effort. It is the second accent of classification of the image and helps to reduce the number of features supplied to the categorization task. After the selection of desired feature, that will be used in the task and failing are discarded. Even though a reduction is desirable in dimensionality, increment in inaccuracy has to be along with selective power of classifiers. This technique is useful when the size of the image is large and extracted feature is helpful for quick completion of the task like matching of the images and recovery.

G. Class of Features [19]

An operation of more than one dimension which designates few perceptible properties of the image/ body is called a feature and it measures some important components of the object. The following are the different features.

1) General features: Texture, color and shape are the purpose free features and called general features. According to the conceptual level, further they are divided into the following.

- Pixel-level features, which are estimated at every pixel such as location and color.
- Local Features, features estimated over the decision of class of the image line on image bisection and detection of edge.
- Global Features, features steered over the complete image impartially.

2) Domain explicit features: These features are application dependent such as finger prints, human faces and visionary features which are generally a fusion of low level features for a particular domain.

H. Color Feature [19]

Color feature is the visual feature and is extensively used in image recovery. Stability, efficiency, simplicity in implementation and computationally as well and low storage capacity are the advantages of color features.

I. Morphological Operators [19]

Morphological operators convert the original image into another one by interacting with the other images of the structural elements like absolute size and shape and it provides a systematic approach for characterization in many applications like object segmentation, edge detection and suppression of noise. The goal of Fuzzy mathematical morphology is to develop the binary morphological operators to gray level images.

J. Mathematical Morphology (MM) [19]

MM is a group of operations. Erosion, dilation, opening and closing are the essential morphological operations on the image pixels. To segregate bright and dark structures than the neighboring features, opening and closing transforms can be used. General methodology has been introduced for fuzzy dilation, erosion, opening and closing.

Dilation is the mechanism of increasing the maximum value in the window thereafter the brightness of the image will be increased and the image objects will be extended as well by modifying the value from 0 to 1. Erosion is the reverse process of dilation, here the image becomes darker than the original and deals the image by converting the pixel value from 1 to 0.

For getting stabilized gray scale images, morphological openings and closings are playing a vital role in image processing. These mathematical morphological operations are similar to the design of set theory and its operations as well. At this junction, fuzzy set theory plays an efficient role in mathematical morphology as the images are uncertain in nature.

K. Morphological Caption [19]

Area, volume, perimeter, gray levels, density, maximum, minimum average, standard deviation of major and minor axes, location, unusualness, point of restriction and centroid are the morphological descriptors or captions in image processing.

L. Projections [19]

Generally in image processing, there are two types projections are available on binary images such as horizontal and vertical projections. This process scan from left side of each line and records the pixel changes from 0 to 1and again to 0 where the number of changes is independent of the pixels. Stability can be expected even in noisy condition in this process. After obtaining associated components, progress of the pixel values from 0 to 1 or vice versa, need to be checked horizontally whereas background area have less progress or transitions.

If the allotted amount of changes of each row lies between two thresholds such as low and high then that row will be considered as a desired area. Next vertical transitions will be considered to find the exact location of the feature which is to be extracted by inquiring the length and height of the feature and their ratio and adequate number of pixels in that area [13].

IV. Proposed Operational Laws

Let $\overline{F}, \overline{F}_1, \overline{F}_2$ be three TIT2FNs and $\chi > 0$, then we define their operational laws as follows.

A. Addition Consider

$$U_1 = \underset{t=1}{\overset{2}{aot}}\left(\underline{l_{F_t}}\right), V_1 = \underset{t=1}{\overset{2}{aot}}\left(\overline{l_{F_t}}\right), W_1 = \underset{t=1}{\overset{2}{aot}}\left(c_{M_t}\right), X_1 = \underset{t=1}{\overset{2}{aot}}\left(\underline{r_{F_t}}\right), Y_1 = \underset{t=1}{\overset{2}{aot}}\left(\overline{r_{F_t}}\right).$$

where aot= sum of the terms

$$\overline{F_1} \oplus \overline{F_2} = \left(\left[\min\left(a_1^{\frac{1}{\chi}},1\right), \min\left(b_1^{\frac{1}{\chi}},1\right)\right], \min\left(c_1^{\frac{1}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(d_1^{\frac{1}{\chi}},1\right), \min\left(e_1^{\frac{1}{\chi}},1\right)\right]\right. \quad (8)$$

B. Multiplication Consider

$$U_2 = \overset{2}{\underset{t=1}{aot}}\left(\underline{l_{F_t}}\right)^{\chi}, V_2 = \overset{2}{\underset{t=1}{aot}}\left(\overline{l_{F_t}}\right)^{\chi},\ W_2 = \overset{2}{\underset{t=1}{aot}}\left(c_{F_t}\right)^{\chi},$$

$$X_2 = \overset{2}{\underset{t=1}{aot}}\left(\underline{r_{F_t}}\right)^{\chi}, Y_2 = \overset{2}{\underset{t=1}{aot}}\left(\overline{r_{F_t}}\right)^{\chi}.$$

$$\overline{F_1} \oplus \overline{F_2} = \left\{\left[\min\left(U_2^{\frac{1}{\chi}},1\right), \min\left(V_2^{\frac{1}{\chi}},1\right)\right], \min\left(W_2^{\frac{1}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_2^{\frac{1}{\chi}},1\right),\ \min\left(Y_2^{\frac{1}{\chi}},1\right)\right]\right\} \quad (9)$$

C. Multiplication by an Ordinary Number Consider

$$U = \underline{l_F}, V = \overline{l_F}, W = c_F, X = \underline{r_F}, Y = \overline{r_F}$$

$$g \bullet \overline{F} = \left\{\left[\min\left\langle U^{\frac{g}{\chi}},1\right\rangle, \min\left\langle V^{\frac{g}{\chi}},1\right\rangle\right], \min\left\langle W^{\frac{g}{\chi}},1\right\rangle,\right.$$

$$\left.\left[\min\left\langle X^{\frac{g}{\chi}},1\right\rangle, \min\left\langle Y^{\frac{g}{\chi}},1\right\rangle\right]\right\}. \quad (10)$$

D. Power Consider

$$U_3 = \underline{l_F}, V_3 = \overline{l_F}, W_3 = c_F,\ X_3 = \underline{r_F}, Y_3 = \overline{r_F}$$

$$\overline{F}^{g} = \left\{\left[\max\left(1-\left[a_3^{\chi}\right]^{g/\chi},0\right), \max\left(1-\left[b_3^{\chi}\right]^{g/\chi},0\right)\right],\right.$$

$$\max\left(1-\left[c_3^{\chi}\right]^{g/\chi},0\right),$$

$$\left.\left[\max\left(1-\left[d_3^{\chi}\right]^{g/\chi},0\right), \max\left(1-\left[e_3^{\chi}\right]^{g/\chi},0\right)\right]\right\} \quad (11)$$

V. Proposed Theorems

Here the mathematical properties of aggregation properties are proved for TIT2FN using TIT2WA operator which represents the essential qualities of the image processing such as idempotent ability, stability and image contrast.

Let TIT2FNs $\overline{F} = \left([\underline{l_{F_t}}, \overline{l_{F_t}}], c_{F_t}, [\underline{r_{F_t}}, \overline{r_{F_t}}]\right), t = 1, 2, ..., n$, where $0 \le \underline{l_F} \le \overline{l_F} \le c_F \le \underline{r_F} \le \overline{r_F} \le 1$ be a collection of TIT2FNs.

A. Theorem

The aggregation value of these fuzzy numbers using TIT2FYWG operator is again a TIT2FN and

$$TIT2WA_{\rho}\left\langle \overline{F_1}, \overline{F_2}, ..., \overline{F_n}\right\rangle = \left\{\left[\min\left(\left[U_n^{\chi}\right]^{\rho_t/\chi},1\right), \min\left(\left[V_n^{\eta}\right]^{\rho_t/\chi},1\right)\right],\right.$$

$$\left.\min\left(\left[W_n^{\eta}\right]^{\rho_t/\chi},1\right), \left[\min\left(\left[X_n^{\eta}\right]^{\rho_t/\chi},1\right),\ \min\left(\left[Y_n^{\eta}\right]^{\rho_t/\chi},1\right)\right]\right\} \quad (12)$$

Where the weight vector is $\rho = \left(\rho_1, \rho_2, ..., \rho_n\right)^T, \rho_n \ge 0$, the sum of the weight vectors is equal to 1.

Proof:

The aggregation value of these fuzzy numbers using TIT2FYWG operator is again a TIT2FN and

$$TIT2WA_{\rho}\left\langle \overline{F_1}, \overline{F_2}, ..., \overline{F_n}\right\rangle = \left\{\left[\min\left(\left[U_n^{\chi}\right]^{\rho_t/\chi},1\right), \min\left(\left[V_n^{\eta}\right]^{\rho_t/\chi},1\right)\right],\right.$$

$$\left.\min\left(\left[W_n^{\eta}\right]^{\rho_t/\chi},1\right), \left[\min\left(\left[X_n^{\eta}\right]^{\rho_t/\chi},1\right),\ \min\left(\left[Y_n^{\eta}\right]^{\rho_t/\chi},1\right)\right]\right\},$$

Where the weight vector is $\rho = \left(\rho_1, \rho_2, ..., \rho_n\right)^T, \rho_n \ge 0$, the sum of the weight vectors is equal to 1.

Here use mathematical induction method.

Case (i): For $n = 2$

Consider,

$$U_1 = \left(\underline{l_{F_1}}\right)^{\chi}, V_1 = \left(\overline{l_{F_1}}\right)^{\chi}, W_1 = \left(c_{F_1}\right)^{\chi} \quad X_1 = \left(\underline{r_{F_1}}\right)^{\chi}, Y_1 = \left(\overline{r_{F_1}}\right)^{\chi}$$

Using multiplication operation

$$g \bullet \overline{F_1} = \left\{\left[\min\left\langle U_1^{\frac{g}{\chi}},1\right\rangle, \min\left\langle V_1^{\frac{g}{\chi}},1\right\rangle\right], \min\left\langle W_1^{\frac{g}{\chi}},1\right\rangle, .\right.$$

$$\left.\left[\min\left\langle X_1^{\frac{g}{\chi}},1\right\rangle, \min\left\langle Y_1^{\frac{g}{\chi}},1\right\rangle\right]\right\}$$

Consider,

$$U_2 = \left(\underline{l_{F_2}}\right)^{\chi}, V_2 = \left(\overline{l_{F_2}}\right)^{\chi}, W_2 = \left(c_{F_2}\right)^{\chi}, X_2 = \left(\underline{r_{F_2}}\right)^{\chi}, Y_2 = \left(\overline{r_{F_2}}\right)^{\chi}$$

$$g\bullet\overline{F_2}=\left\{\left[\min\left\langle U_2^{\frac{g}{\chi}},1\right\rangle,\min\left\langle V_2^{\frac{g}{\chi}},1\right\rangle\right],\min\left\langle W_2^{\frac{g}{\chi}},1\right\rangle,\right.$$

$$\left.\left[\min\left\langle X_2^{\frac{g}{\chi}},1\right\rangle,\min\left\langle Y_2^{\frac{g}{\chi}},1\right\rangle\right]\right\}.$$

$$r\underset{Y}{\oplus}s=\min\left(\left(r^{\eta}+s^{\eta}\right)^{\frac{1}{\eta}},1\right),\eta>0,\ for\,all\ r,s\in[0,1]^2$$

$$TIT2FWA_{\rho}\left(\overline{F_1},\overline{F_2}\right)=\rho_1\bullet\overline{F_1}\oplus\rho_1\bullet\overline{F_2}$$

$$=\left\{\left[\min\left(\left[\underset{t=1}{\overset{2}{aot}}\left(\min\left(\left(U_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\ \min\left(\left[\underset{t=1}{\overset{2}{aot}}\left(\min\left(\left(V_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right],\right.$$

$$\min\left(\left[\underset{p=1}{\overset{2}{aot}}\left(\min\left(\left(W_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),$$

$$\left.\left[\min\left(\left[\underset{t=1}{\overset{2}{aot}}\left(\min\left(\left(X_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\min\left(\left[\underset{1=1}{\overset{2}{aot}}\left(\min\left(\left(Y_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right]\right\}.$$

Consider,

$$U_4=\left(1-\underline{l_{M_p}}\right)^{\chi},V_4=\left(1-\overline{l_{M_t}}\right)^{\chi},W_4=\left(1-c_{M_t}\right)^{\chi},$$

$$X_4=\left(1-\underline{r_{M_t}}\right)^{\chi},Y_4=\left(1-\overline{r_{M_t}}\right)^{\chi}.$$

$$=\left\{\left[\min\left(\left(\underset{t=1}{\overset{2}{aot}}\left(\min\left(U_4^{\frac{\rho_t}{\chi}},1\right)\right)\right),1\right),\min\left(\left(\underset{t=1}{\overset{2}{aot}}\left(\min\left(1-V_4^{\frac{\rho_t}{\chi}},1\right)\right)\right),1\right)\right],\right.$$

$$\min\left(\left(\underset{t=1}{\overset{2}{aot}}\left(\min\left(W_4^{\frac{\rho_t}{\chi}},1\right)\right)\right),1\right),$$

$$\left.\left[\min\left(\left(\underset{t=1}{\overset{2}{aot}}\left(\min\left(X_4^{\frac{\rho_t}{\chi}},1\right)\right)\right),1\right),\min\left(\left(\underset{p=1}{\overset{2}{aot}}\left(\min\left(Y_4^{\frac{\rho_t}{\chi}},1\right)\right)\right),1\right)\right]\right\}.$$

$$=\left\{\left[\min\left[U_2^{\frac{\rho_t}{\chi}},1\right],\min\left[V_2^{\frac{\rho_t}{\chi}},1\right]\right],\min\left[W_2^{\frac{\rho_t}{\chi}},1\right],\right.$$

$$\left.\left[\min\left[X_2^{\frac{\rho_t}{\chi}},1\right],\min\left[Y_2^{\frac{\rho_t}{\chi}},1\right]\right]\right\}$$

For $n=k,$

$$U_k=\underset{t=1}{\overset{k}{aot}}\left(1-\underline{l_{F_t}}\right)^{\chi},V_k=\underset{t=1}{\overset{k}{aot}}\left(1-\overline{l_{F_t}}\right)^{\chi},W_k=\underset{t=1}{\overset{k}{aot}}\left(1-c_{F_t}\right)^{\chi},$$

$$X_k=\underset{t=1}{\overset{k}{aot}}\left(1-\underline{r_{F_t}}\right)^{\chi},Y_k=\underset{t=1}{\overset{k}{aot}}\left(1-\overline{r_{F_t}}\right)^{\chi}$$

$$TIT2WA_{\rho}\left\langle\overline{F_1},\overline{F_2},...,\overline{F_k}\right\rangle$$

$$=\left\{\left[\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(\left[U_k^{\frac{\rho_t}{\chi}},1\right]\right)\right)\right],1\right),\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(V_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right],\right.$$

$$\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(W_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),$$

$$\left.\left[\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(X_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\ \min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(Y_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right]\right\}.$$

$$=\left\{\left[\min\left[U_k^{\frac{\rho_t}{\chi}},1\right],\min\left[V_k^{\frac{\rho_t}{\chi}},1\right]\right],\ \min\left[W_k^{\frac{\rho_t}{\chi}},1\right],\right.$$

$$\left.\left[\max\left[X_k^{\frac{\rho_t}{\chi}},1\right],\max\left[Y_k^{\frac{\rho_t}{\chi}},1\right]\right]\right\}.$$

For $n=k+1,$

$$TIT2WA_{\rho}\left\langle\overline{F_1},\overline{F_2},...,\overline{F_k}\right\rangle\oplus\left\langle\overline{F_{k+1}}\right\rangle$$

$$=\left\{\left[\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(U_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(V_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right],\right.$$

$$\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(W_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),$$

$$\left.\left[\min\left(\left[\underset{t=1}{\overset{k}{aot}}\left(\min\left(X_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\min\left(\left[\underset{p=1}{\overset{k}{sum}}\left(\min\left(Y_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right]\right\}.$$

$$\oplus\left\{\left[\min\left(U_{k+1}^{\frac{\rho_{k+1}}{\chi}},1\right),\min\left(V_{k+1}^{\frac{\rho_{k+1}}{\chi}},1\right)\right],\min\left(W_{k+1}^{\frac{\rho_{k+1}}{\chi}},1\right),\right.$$

$$\left[\max\left(X_{k+1}^{\frac{\rho_{k+1}}{\chi}},1\right),\max\left(Y_{k+1}^{\frac{\rho_{k+1}}{\chi}},1\right)\right]\Bigg\}.$$

$$=\left\{\left[\min\left(\left[\operatorname*{aot}_{t=1}^{k+1}\left(\min\left(U_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\min\left(\left[\operatorname*{aot}_{t=1}^{k+1}\left(\min\left(V_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right],\right.$$

$$\min\left(\left[\operatorname*{aot}_{t=1}^{k+1}\left(\min\left(W_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),$$

$$\left.\left[\min\left(\left[\operatorname*{aot}_{t=1}^{k+1}\left(\min\left(X_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\min\left(\left[\operatorname*{aot}_{t=1}^{k+1}\left(\min\left(Y_k^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right]\right\}.$$

$$=\left\{\left[\min\left[U_{k+1}^{\frac{\rho_t}{\chi}},1\right],\min\left[V_{k+1}^{\frac{\rho_t}{\chi}},1\right]\right],\ \min\left[W_{k+1}^{\frac{\rho_t}{\chi}},1\right],\right.$$

$$\left.\left[\min\left[X_{k+1}^{\frac{\rho_t}{\chi}},1\right],\min\left[Y_{k+1}^{\frac{\rho_t}{\chi}},1\right]\right]\right\}.$$

Hence the result holds for all the values of n .

B. Theorem (Idempotency)

If $\overline{F_t}=\overline{F}$ for all the values of t then

$$TIT2WA_\rho\left\langle\overline{F_1},\overline{F_2},...,\overline{F_n}\right\rangle=\overline{F}. \tag{13}$$

Proof:

By theorem A,

$$TIT2WA_\rho\left\langle\overline{F_1},\overline{F_2},...,\overline{F_n}\right\rangle=\left\{\left[\min\left(U_n^{\frac{\rho_t}{\chi}},1\right),\min\left(V_n^{\frac{\rho_t}{\chi}},1\right)\right],\ \min\left(W_n^{\frac{\rho_t}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_n^{\frac{\rho_t}{\chi}},1\right),\min\left(Y_n^{\frac{\rho_t}{\chi}},1\right)\right]\right\}$$

$$TIT2WA_\rho\left\langle\overline{F_1},\overline{F_2},...,\overline{F_n}\right\rangle$$

$$=\left\{\left[\min\left(U_4^{\operatorname*{aot}_{t=1}^{k+1}\left(\frac{\rho_t}{\chi}\right)},1\right),\min\left(V_4^{\operatorname*{aot}_{t=1}^{k+1}\left(\frac{\rho_t}{\chi}\right)},1\right)\right],\ \min\left(W_4^{\operatorname*{aot}_{t=1}^{k+1}\left(\frac{\rho_t}{\chi}\right)},1\right),\right.$$

$$\left.\left[\min\left(X_4^{\operatorname*{aot}_{t=1}^{k+1}\left(\frac{\rho_t}{\chi}\right)},1\right),\min\left(Y_4^{\operatorname*{aot}_{t=1}^{k+1}\left(\frac{\rho_t}{\chi}\right)},1\right)\right]\right\}$$

$$=\left\{\left[\min\left(U_5^{\frac{1}{\chi}},1\right),\min\left(V_5^{\frac{1}{\chi}},1\right)\right],\min\left(W_5^{\frac{1}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_5^{\frac{1}{\chi}},1\right),\min\left(Y_5^{\frac{1}{\chi}},1\right)\right]\right\}$$

$$=\left\{\left[U_3,V_3,W_3,X_3,Y_3\right]\right\}=\left\{\left[U,V\right],W,\left[X,Y\right]\right\}=\overline{F}$$

C. Theorem

If $f>0$ for all the values of p then

$$TIT2WA_\rho\left(f\bullet\overline{F_1},f\bullet\overline{F_2},...,f\bullet\overline{F_n}\right)=f\bullet TIT2WA_\rho\left(\overline{F_1},\overline{F_2},...,\overline{F_n}\right) \tag{14}$$

Proof: Using,

$$f\bullet\overline{F}=\left\{\left[\min\left(U_5^{\frac{f}{\chi}},1\right),\min\left(V_5^{\frac{f}{\chi}},1\right)\right],\min\left(W_5^{\frac{f}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_5^{\frac{f}{\chi}},1\right),\min\left(Y_5^{\frac{f}{\chi}},1\right)\right]\right\}$$

$$TIT2WA_\rho\left(f\bullet\overline{F_1},f\bullet\overline{F_2},...,f\bullet\overline{F_n}\right)$$

$$=\left\{\left[\min\left(\left[\operatorname*{aot}_{t=1}^{n}\left(\min\left(\left(U_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\ \min\left(\left[\operatorname*{aot}_{t=1}^{n}\left(\min\left(\left(V_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right)\right],\right.$$

$$\min\left(\left[\operatorname*{aot}_{t=1}^{n}\left(\min\left(\left(W_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),$$

$$\left.\left[\min\left(\left[\operatorname*{aot}_{t=1}^{n}\left(\min\left(\left(X_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\ \min\left(\left[\operatorname*{aot}_{t=1}^{n}\left(\min\left(\left(Y_2\right)^{\frac{\rho_t}{\chi}},1\right)\right)\right],1\right),\right]\right\}.$$

$$=\left\{\left[\min\left(U_n^{\frac{f\rho_t}{\chi}},1\right),\ \min\left(V_n^{\frac{f\rho_t}{\chi}},1\right)\right],\ \min\left(W_n^{\frac{f\rho_t}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_n^{\frac{f\rho_t}{\chi}},1\right),\min\left(Y_n^{\frac{f\rho_t}{\chi}},1\right)\right]\right\} \tag{15}$$

Now consider,

$$f \bullet TIT2WA_{\rho}\left(\overline{F_1}, \overline{F_2}, ..., \overline{F_n}\right)$$

$$= f \bullet \left\{ \left[\min\left(U_n^{\frac{\rho_t}{\chi}}, 1\right), \min\left(V_n^{\frac{\rho_t}{\chi}}, 1\right) \right], \min\left(W_n^{\frac{\rho_t}{\chi}}, 1\right), \right.$$

$$\left. \left[\min\left(X_n^{\frac{\rho_t}{\chi}}, 1\right), \min\left(Y_n^{\frac{\rho_t}{\chi}}, 1\right) \right] \right\}$$

$$= \left\{ \left[\min\left(\left(U_n^{\frac{f}{\chi}} \right)^{\rho_t}, 1\right), \min\left(\left(V_n^{\frac{f}{\chi}} \right)^{\rho_t}, 1\right) \right], \min\left(\left(W_n^{\frac{f}{\chi}} \right)^{\rho_t}, 1\right), \right.$$

$$\left. \left[\min\left(\left(X_n^{\frac{f}{\chi}} \right)^{\rho_t}, 1\right), \min\left(\left(Y_n^{\frac{f}{\chi}} \right)^{\rho_t}, 1\right) \right] \right\}$$

$$= \left\{ \left[\min\left(\left(U_n^{\frac{f\rho_t}{\chi}} \right), 1\right), \min\left(\left(V_n^{\frac{f\rho_t}{\chi}} \right), 1\right) \right], \min\left(\left(W_n^{\frac{f\rho_t}{\chi}} \right), 1\right), \right.$$

$$\left. \left[\min\left(\left(X_n^{\frac{f\rho_t}{\chi}} \right), 1\right), \min\left(\left(Y_n^{\frac{f\rho_t}{\chi}} \right), 1\right) \right] \right\} \qquad (16)$$

Since (15) = (16), hence the result.

D. Theorem (Stability)

If $t > 0$ $\overline{F_{n+1}} = \left(\left[\underline{l_{F_{n+1}}}, \overline{l_{F_{n+1}}} \right], c_{F_{n+1}}, \left[\underline{r_{F_{n+1}}}, \overline{r_{F_{n+1}}} \right] \right)$ then

$$TIT2WA_{\rho}\left(f \bullet \overline{F_1} \oplus \overline{F_{n+1}}, f \bullet \overline{F_2} \oplus \overline{F_{n+1}}, ..., f \bullet \overline{F_n} \oplus \overline{F_{n+1}} \right) \qquad (17)$$

Proof:

$$TIT2WA_{\rho}\left(f \bullet \overline{F_1} \oplus \overline{F_{n+1}}, f \bullet \overline{F_2} \oplus \overline{F_{n+1}}, ..., f \bullet \overline{F_n} \oplus \overline{F_{n+1}} \right)$$

$$= \left\{ \left[\min\left(U_n^{\frac{f\rho_t}{\chi}}, 1\right) \oplus \overline{F_{n+1}}, \min\left(V_n^{\frac{f\rho_t}{\chi}}, 1\right) \oplus \overline{F_{n+1}} \right], \right.$$

$$\min\left(W_n^{\frac{f\rho_t}{\chi}}, 1\right) \oplus \overline{F_{n+1}},$$

$$\left. \left[\min\left(X_n^{\frac{f\rho_t}{\chi}}, 1\right) \oplus \overline{F_{n+1}}, \min\left(Y_n^{\frac{f\rho_t}{\chi}}, 1\right) \oplus \overline{F_{n+1}} \right] \right\}$$

$$= \left\{ \left[\min\left(\left[\underset{t=1}{\overset{n}{aot}} \left(\min\left(\left[\underset{r=\{t,n+1\}}{aot} \left(\underline{l_{F_t}} \right)^{\chi} \right]^{\frac{f}{\chi}}, 1 \right) \right)^{\rho_t} \right], 1 \right), \right. \right.$$

$$\left. \min\left(\left[\underset{t=1}{\overset{n}{aot}} \left(\min\left(\left[\underset{r=\{t,n+1\}}{aot} \left(\overline{l_{F_t}} \right)^{\chi} \right]^{\frac{f}{\chi}}, 1 \right) \right)^{\rho_t} \right], 1 \right) \right]$$

$$\min\left(\left[\underset{t=1}{\overset{n}{aot}} \left(\min\left(\left[\underset{r=\{t,n+1\}}{aot} \left(c_{F_t} \right)^{\chi} \right]^{\frac{f}{\chi}}, 1 \right) \right)^{\rho_t} \right], 1 \right),$$

$$\left[\min\left(\left[\underset{t=1}{\overset{n}{aot}} \left(\min\left(\left[\underset{r=\{t,n+1\}}{aot} \left(\underline{r_{F_t}} \right)^{\chi} \right]^{\frac{f}{\chi}}, 1 \right) \right)^{\rho_t} \right], 1 \right), \right.$$

$$\left. \left. \min\left(\left[\underset{t=1}{\overset{n}{aot}} \left(\min\left(\left[\underset{r=\{t,n+1\}}{aot} \left(\overline{r_{F_t}} \right)^{\chi} \right]^{\frac{f}{\chi}}, 1 \right) \right)^{\rho_t} \right], 1 \right) \right] \right\}$$

$$= \left\{ \left[\min\left(\left(U_n^{\frac{\rho_t}{\chi}} \right) + \left(\left[\left(\underline{l_{F+1}} \right)^{\chi} \right]^{\frac{1}{\chi}} \right)^{\underset{t=1}{\overset{n}{aot}} \rho_t}, 1 \right), \right. \right.$$

$$\left. \min\left(\left(V_n^{\frac{\rho_t}{\chi}} \right) + \left(\left[\left(\overline{l_{F+1}} \right)^{\chi} \right]^{\frac{1}{\chi}} \right)^{\underset{t=1}{\overset{n}{aot}} \rho_t}, 1 \right) \right],$$

$$\min\left(\left(W_n^{\frac{\varepsilon_p}{\eta}} \right) + \left(\left[\left(c_{F+1} \right)^{\chi} \right]^{\frac{1}{\chi}} \right)^{\underset{t=1}{\overset{n}{aot}} \rho_t}, 1 \right),$$

$$\left[\min\left(\left(X_n^{\frac{\rho_t}{\chi}} \right) + \left(\left[\left(\underline{r_{F+1}} \right)^{\chi} \right]^{\frac{1}{\chi}} \right)^{\underset{t=1}{\overset{n}{aot}} \rho_t}, 1 \right), \right.$$

$$\min\left(\left(Y_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(\overline{r_{F+1}}\right)^{\chi}\right]^{\frac{1}{\chi}}\right)^{\overset{n}{\underset{t=1}{aot}}\rho_t},1\right)\Bigg]\Bigg]\Bigg\} \quad (18)$$

$$f\bullet TIT2WA_{\rho}\left(\overline{F_1},\overline{F_2},...,\overline{F_n}\right)\oplus\overline{F_{n+1}}$$

$$=\left\{\left[\min\left(U_n^{\frac{f\rho_t}{\chi}},1\right),\min\left(V_n^{\frac{f\rho_t}{\chi}},1\right)\right],\min\left(W_n^{\frac{f\rho_t}{\chi}},1\right),\right.$$

$$\left.\left[\min\left(X_n^{\frac{f\rho_t}{\chi}},1\right),\ \min\left(Y_n^{\frac{f\rho_t}{\chi}},1\right)\right]\right\}.$$

$$\oplus\left\langle\left[\underline{l_{F_{n+1}}},\overline{l_{F_{n+1}}}\right],c_{F_{n+1}},\left[\underline{r_{F_{n+1}}},\overline{r_{F_{n+1}}}\right]\right\rangle$$

$$=\left\{\left[\min\left(\left(U_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(\underline{l_{F+1}}\right)^{\chi}\right]^{\frac{1}{\chi}}\right),1\right),\min\left(\left(U_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(\overline{l_{F+1}}\right)^{\chi}\right]^{\frac{1}{\chi}}\right),1\right)\right],\right.$$

$$\min\left(\left(W_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(1-c_{F+1}\right)^{\chi}\right]^{\frac{1}{\chi}}\right),1\right),$$

$$\left.\left[\min\left(\left(X_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(\underline{r_{F+1}}\right)^{\chi}\right]^{\frac{1}{\chi}}\right),1\right),\ \min\left(\left(Y_n^{\frac{\rho_t}{\chi}}\right)+\left(\left[\left(\overline{r_{F+1}}\right)^{\chi}\right]^{\frac{1}{\chi}}\right),1\right)\right]\right\}. \quad (19)$$

Here, (18) = (19) Hence the result.

E. Theorem (Image Contrast)

For given arguments $\overline{F_t}, t=1,2,...,n$ and the parameter $\chi\in\left(1,+\infty\right)$ then TIT2WA operator is monotonically non-decreasing (MND) with respect to the parameter.

Proof:

To prove the operator is MND with respect to the parameter, we have to prove the same for every reference point function is MND w.r.t the parameter.

Since $0\le\underline{l_F}\le\overline{l_F}\le c_F\le\underline{r_F}\le\overline{r_F}\le1$, $\min\left(U_5^{\frac{k}{\chi}},1\right)>0.$

And it is true for all the reference points. Hence the result.

$$\max\left(\left(1-C_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-c_{M+1}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\overset{n}{\underset{p=1}{sum}}\varepsilon_p},0\right), \quad (20)$$

Based on the theorem A and the operational law,

$$\text{TIT2FYWG}_{\varepsilon}\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle_{Y}^{\bullet t}\underset{Y}{\otimes}\overline{M_{n+1}}$$

IV. Application of Type-2 Fuzzy Logic for Feature Extraction

Fig. 2 is the proposed algorithm for Brain extraction from a DICOM image using triangular norms.

MRI of the patient DICOM image has been considered for this application, Fig. 3. Using MATLAB 2015a, brain has been extracted from MRI. The image is taken from our empirical data and its description is as follows

Size of the image	: 512 x 512
Mean of the image	: 242 x 4
Standard deviation	: 50.31
Mean absolute deviation	: 22.43

The below figures are the output of the image processing application in edge detection through triangular norms by MATLAB 2015 a.

Fig. 4 is the gradient through x axis and Fig. 5 is the gradient through the y axis.

The figures reveals that the image gradient to identify the region uniformly.

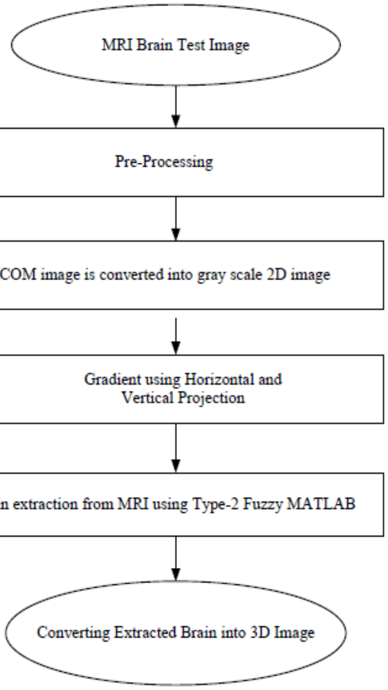

MRI Brain Test Image
Pre-Processing
DICOM image is converted into gray scale 2D image
Gradient using Horizontal and Vertical Projection
Brain extraction from MRI using Type-2 Fuzzy MATLAB
Converting Extracted Brain into 3D Image

Fig. 2. Architecture of Brain Extraction.

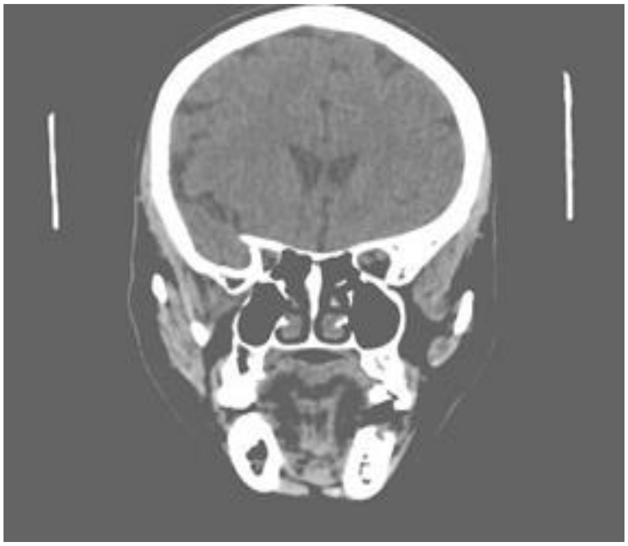


Fig. 3. Original Gray Scale DICOM Image.

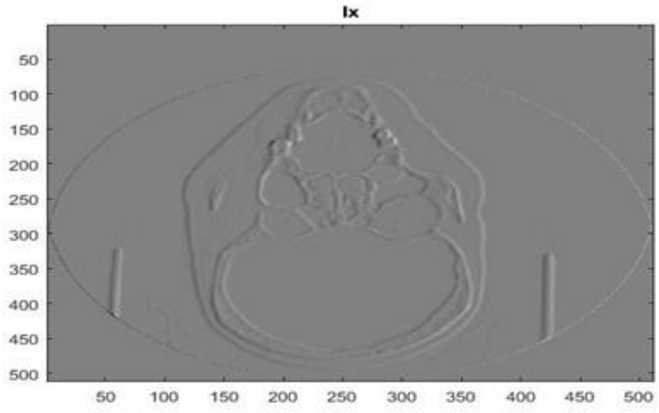


Fig. 4. Gradient through X-Axis.

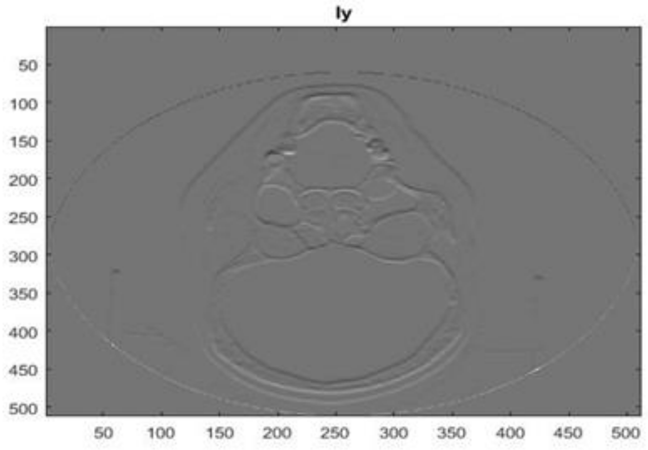


Fig. 5. Gradient through Y-axis.

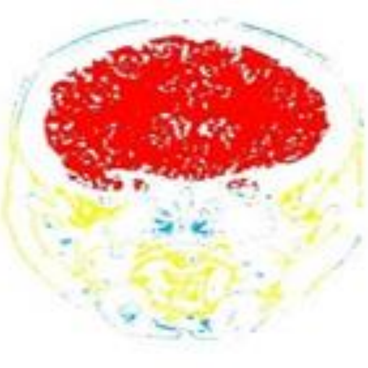


Fig. 6. Extracted Color Image.

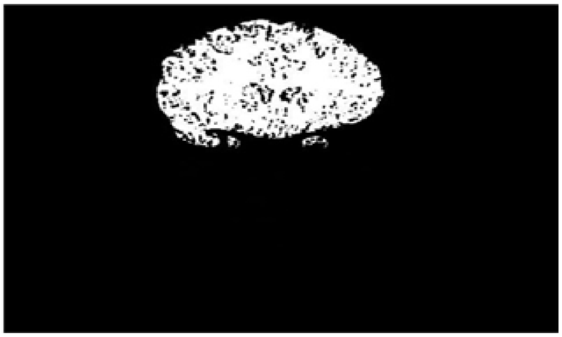


Fig. 7. Extracted Feature.

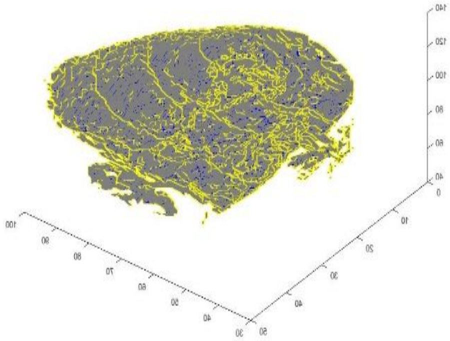


Fig. 8. 3D image of the Extracted Brain Image.

Fig. 6 is the color image output of the Brain extraction using Type-2 Fuzzy based MATLAB 2015a and Fig. 7 is the extracted brain from the DICOM image.

Fig. 8 is the 3D version of the extracted Brain image.

Feature extraction is an essential part of the image processing. In this work, brain has been extracted from patient MRI using MATLAB 2015a. It is examined that, fuzzy logic feature extractor helps to reduce the dimensionality of the image.

Matlab coding based on Interval type-2 fuzzy logic unable to handle non-membership and indeterminacy of the feature which is to be extracted and it is the limitation of the present study.

VI. Discussion

The coding of MATLAB 2015 a under interval type-2 fuzzy has not been used to extract a feature from the image. In Literature review, previous studies have been reviewed on feature extraction and there is no contribution of work for brain extraction using the applied MATLAB coding and 3D version of the extracted color image. This shows the novelty of the proposed work.

VII. Conclusion

Extracting a desired feature plays a key role of image processing. This transforms the pieces from high dimensional space to low dimensional and to decrease the degree of the dimensional. Hence it helps to reduce the dimensionality of the image. In this paper, the proved mathematical properties are related to image processing especially feature extraction such as stability and image contrast and brain has been extracted from patient MRI in a better way and produced 3D image of the extracted brain using interval Type-2 fuzzy MATLAB and it is very helpful for dimensionality reduction while saving the data of the image. Using this technology, extra growth of the cells can be detected and diagnosed if any. In future this work would be extended to intuitionistic and neutrosophic environments.

Data Availability statement

The DICOM data used to support the findings of this study are available from the corresponding author upon request.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Materials

The data set in Fig. 9 is the montage of the images in a single file and is from a patient MRI. This MRI which is in the 3D form is converted to 2D form (DICOM) using MATLAB2015a. The 3D format consists of 25 DICOM file formats; the montage of the images is obtained as a single frame. Out of these 25 DICOM images a clear full image is chosen as in Fig. 10. Using Dilation and erosion methods, the gradient is identified. The edge detection is performed through triangular norms using MATLAB 2015a.

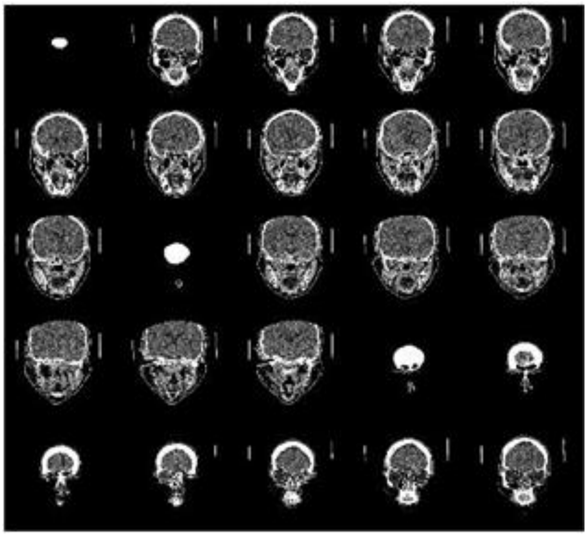


Fig. 9. Montage of the Images.

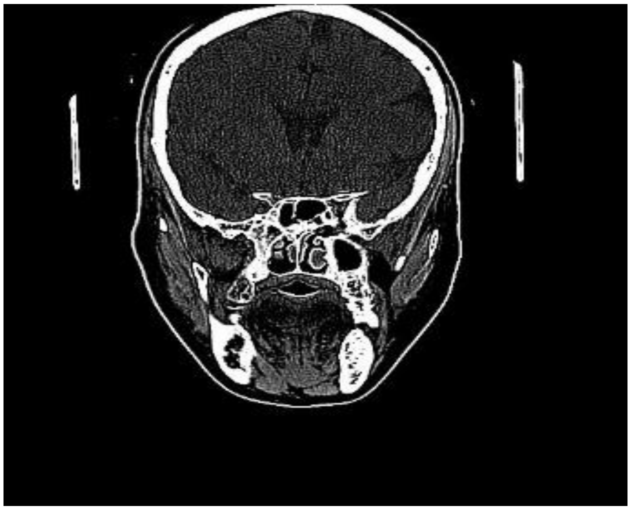


Fig. 10. Clear Image from Montage.

References

[1] L. A. Zadeh, “The concept of a linguistic variable and its application to Approximate Reasoning-I,” InformationSciences, Vol.8, No.3, 1975, pp.199-249.

[2] J. Dombi, “A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators,” Fuzzy Sets and Systems, Vol. 8, No.2, 1982, pp.149-163.

[3] E. Czogala and H. Zimmermann, “The Aggregation Operators for Decision Making in Probabilistic Fuzzy Environment,” Fuzzy Sets and Systems, Vol.9, No.1, 1984, pp.183-196.

[4] D. Dubois and H. Prade, “A review of fuzzy set aggregation Connectives, Information Sciences,” Vol. 36, No.1-2, 1985, pp. 85-121.

[5] M. M. Gupta and J. Qi, “Theory of t-norms and Fuzzy Inference Methods,” Fuzzy Sets and Systems, Vol. 40, No. 3, 1991, pp.431-450.

[6] R. R. Yager, “Aggregation operators and fuzzy systems modeling,” Fuzzy Sets and Systems, Vol. 67, No.2, 1994, pp.129-145.

[7] Y. Yao and J. Wang, "Interval based Uncertain Reasoning using Fuzzy and Rough Sets," Advances in Machine Intelligence & Soft-Computing, 1996, pp.1-20.

[8] M. Detynieki, "Mathematical Aggregation Operators and their Application to Video Querying," Ph.D. Thesis in Artificial Intelligence Specialty of Paris, 2000, pp.1-104.

[9] K Franke, M. Koppen and B.Nickolay, "Fuzzy Image Processing by using Dubois and Prade Fuzzy Norms," Pattern Recognition, 2000, pp.1-4.

[10] K. Franke and M. Koppen, "A computer-based system to support forensic studies on handwritten documents," International Journal on Document Analysis and Recognition, 2000, pp.1-13.

[11] T. Calvo and R. Mesiar, "Weighted Means Based on Triangular Conorms," International Journal of Uncertain Fuzziness Knowledge Based Systems, Vol.29, No.12, 2001, pp.1173-1180.

[12] C. H. Chiu and W. J. Wang, "A simple computation of MIN and MAX operations for fuzzy numbers," Fuzzy Sets and Systems, Vol.126, No. 2, 2002, pp.273-276.

[13] D. Dubois and H. Prade, "On the Use of Aggregation operations in Information Fusion Processes," Fuzzy Sets and Systems, Vol. 142, No.3, 2004, pp.143-161.

[14] K. Hirota, H. Nobuhara, K. Kawamoto and S. I. Yoshida, "On a Lossy Image Compression/ Reconstruction Method Based on Fuzzy Relational Equations," Iranian Journal of Fuzzy Systems, Vol.1, No.1, 2004, pp.33-42.

[15] S. Zhou, F. Chiclana, R. John and J. Garibaldi, "Type-1 OWA operators for aggregating uncertain information with uncertain weights induced by type-2 linguistic Quantities," Fuzzy Sets and Systems, Vol.159, No.24, 2008, pp.3281-3296.

[16] R. Mesiar, A. Kolesarova, T. Calvo and M. Komornikova, "A review of aggregation functions," In: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, H.Bustince et al. (Eds.), Springer, Berlin, 2008, pp.121-144.

[17] Z. Gera and J. Dombi, "Exact calculations of extended logical operations on fuzzy truth values," Fuzzy Sets and Systems, Vol.159, No.11, 2008, pp.1309-1326.

[18] R. Maini and H. Aggarwal, "Study and Comparison of Various Image Edge Detection Techniques," International Journal of Image Processing, Vol.3, No.1, 2009, pp.1-12.

[19] M. G. Hidalgo, A. M. Torres, D. R. Aguilera, J. T. Sastre, "Image Analysis Applications of Morphological Operators on Uninorms," Soft computing techniques for uncertainty management in image processing, 2009, pp. 630-635. ISBN: 978-989- 95079-6-8.

[20] A. A. Kiaei, S. B. Shouraki, S.H. Khasteh, M. Khademi and A. R. G. Samani, "New S-norm and T-norm Operators for Active Learning Method," Fuzzy Optimization and Decision Making, 2010, pp.1-11.

[21] T. Calvo, G. Mayor and R. Messiar, "Aggregation Operators: New Trends and Application," 2012, Springer Publications.

[22] J. Qin and X. Liu, "Frank Aggregation operators for Triangular Interval Type-2 Fuzzy Set and its Application in Multiple Attribute Group Decision Making," Journal of Applied Mathematics, Vol. 1, 2014, pp.1-24.

[23] F. Castro-Company and P. Tirado, "On Yager and Hamacher t-norms and Fuzzy Metric spaces," International Journal of Intelligent Systems, Vol.29, No.12, 2014, pp.1173-1180.

[24] T. Chaira, "Medical Image Processing: Advanced Fuzzy Set Theoretic Techniques," CRC Publications, 2015, ISBN: 9781498700450-CAT#K24522.

[25] D. Vivona and M. Divari, "On Information Reception and Some of its Aggregation Operators," 8th International Summer School on Aggregation Operators, Kotowice, Poland, University of Silesia, 2015, pp.233-237.

[26] O. Castillo, M. A. Sanchez, C. I. Gonzalez and G. E. Martinez, "Review of Recent Type-2 Fuzzy Image Processing Applications," Information, Vol. 8, No. 97, 2017, pp. 1-18.

[27] I. Sheikh and K. A. Khan, "Simulation of Image Edge Detection using Fuzzy Logic in MATLAB," International Journal of Computer & Mathematical Sciences, Vol. 6, No. 5, 2017, pp. 19-22.

[28] M. Lathamaheswari, D. Nagarajan, A. Udayakumar, and J. Kavikumar, "Review on Type-2 Fuzzy in Biomedicine," Indian Journal of Public Health Research & Development.

[29] Kenjharayoobchandio and Yasarayaz, "Fuzzy Logic Based Digital Image Edge Detection," International Journal of Electrical, Electronics and Data Communication, Vol. 6, No. 2, 2018, pp.18-22.

[30] D. Nagarajan, M. Lathamaheswari, R. Sujatha and J. Kavikumar, "Edge Detection on DIOM Image using Triangular Norms in Type-2 Fuzzy," International Journal of Advanced Computer Science and Applications, Vol. 9, No. 11, pp.462-475.

[31] A. K. Dutta, "Intuitionistic Fuzzy Logic Implementation in Image Fusion Technique," Asian Journal of Research in Computer Science, Vol. 1. No. 1, 2018, pp. 1-7.

[32] M. Lathamaheswari, D. Nagarajan, J. Kavikumar and C. Phang, "A Review on Type-2 Fuzzy Controller on Control System," Journal of Advanced Research in Dynamical and Control Systems, Vol.10, No.11, 2018, pp.430-435.

[33] D. G. Privezentsev, A. L. Zhiznyakov, A. V. Astafiev and E. V. Pugin, "Using fuzzy fractal features of digital images for the material surface analysis," Journal of Physics, Vol. 944, 2018, pp. 1-5.

[34] D.Nagarajan, "Three dimensional visualization of brain using machine learning". International Journal of pure and applied mathematics, Vol:117 no.7, 2017,p.459-466.

[35] D.Nagarajan, Nagarajan.V, Abitha Gladis N.K,, "Block Processing And Edge Detection For A Dicom Image". International Journal of Pure and Applied Mathematical Sciences Vol 9,No 1 , 2016,pp 9-16.

[36] D.Nagarajan, "Image denoiseing using LU decomposition and Features extraction using GLCM", International Journal of Advanced Research in Computer Science, Vol:8 no.7,2017,pp.675-677.

A Personalized Hybrid Recommendation Procedure for Internet Shopping Support

R. Shanthi*
*Research Scholar,
Sathyabama Institute of Science and Technology,
Jeppiaar Nagar, Rajiv Gandhi Salai,
Chennai- 119

Dr. S.P. Rajagopalan
Professor of Computer Science and Engineering,
GKM College of Engineering and Technology,
Chennai-63

***Abstract*—Lately, recommender systems (RS) have offered a remarkable breakthrough to users. It lessens the user time cost thereby delivering faster and better results. After purchasing a product there are recommendations according to the different comments provided by users. Within a short span of product utilization and quality, the users receive a product recommendation. But this doesn't work out good so as to make it much better;feedbacks, commands and reviews are fetched on the basis of in-depth commands, globally like and normal keys. Recommendation systems are crucially important for the delivery of personalized product to users. With personalized recommendation to product, users can enjoy a variety of targeted recommendations such as online product; the current paper suggests hybrid recommendation system (HRS) that makes use of rating and review to recommend any product to user. The main objective of this paper is to personalize recommendation of product that have become extremely effective revenue drivers for online shopping business. Despite the great benefits, deploying personalized recommendation services typically requires the collection of users' personal data for processing and analytics, which undesirably makes users. To implement product recommendations following are incorporated that is retrieving personal data, Logical Language based Rule Generation *(*LLRG), ranking and Hybrid recommendation system. The stages in the suggested recommendation system include, Data Gathering, preprocessing, filtering and Ranking. The Ranking algorithm ranks the products in relation to the sales count. The top list displays the product having greatest count number. In the LLRG strategy, the logic rule generation methodology retrieves useful and mandatory data from reviews, commands, products original state and thereafter comes the recommendation. The HRS enforces two techniques, namely, location based and the other being heterogeneous domain based. Also the recommendations presented to the user are in context to the user's activities, choices and conduct that are in accord with user's personal likings and aids in decision making. When comparing the outcome, it is clear that the suggested method is superior than the traditional with regard to clarity, effective recommendation and coverage rate. It's evaluated that Hybrid Recommendation System yields in greater results compared with rest of the existing recommendation techniques**. **We, also identity to some future research directions.**

Keywords—Web mining; web search; products; ranking; recommendation system; hybrid approach; e-commerce; online shopping market

I. Introduction

There is an enormous increase and demand in Online shopping as it offers cheap rates, multiple options and fast logistic systems[1] [2]. People have become an avid user of Online Shopping for purchasing any product. To examine the quality of a product, user's comments and reviews becomes very necessary information [3] [4]. The manufactures can enhance the quality of a product by referring to the user comments related to those products. The question is how to retrieve such necessary information and produce neutral products. The research field talks about the emerging quality test system which can handle this massive textual information. The technology of Web Mining is founded on logical language processing and web mining. It projects a way to deal with this issue [5]. Document-level recommendation system analysis focuses on Emotional and rating orientation of every feedback. It judges the viewpoint of the information expressed by the authors, discussing the sentence-level web mining and handles the statements of the product ‟considering each view as an analysis object, thereafter trace the authors" opinion inclinations. This technique aids in tracing clear cut details of the comments keeping it highly confidential, but performing such task can be extremely complicated. Consider a laptop for example its various attributes can be classified as brand type, endurance time, cost, performance, look and so on. Every feature expressed by the author for every comment is analyzed accordingly, thereafter a complete and thorough evaluation is performed to omit over generalization. A products name is spoiled if it receives a bad user review or rating. Also there are user comments after a product is purchased. Any negative comments may recommend the product as of low quality compared to good product based on previous user comments. This can be bad; hence feedbacks, commands and reviews are fetched on basis of in-depth commands, globally like and normal keys.

The suggested paper hybrid recommendation system (HRS) helps in recommending the products to user keeping in account of the user's ratings and reviews. To implement product recommendations following are incorporated that is retrieving personal data, ranking Logical Language based Rule Generation *(*LLRG) and Hybrid recommendation system. The stages in the suggested recommendation system include, Data Gathering, Preprocessing, Filtering, Ranking, Prediction, Recommendations. Initially, Data gathering process aims to gather personal information via Internet. The role of preprocessing involves eliminating void columns, null values and noisy data, thereafter comes the filtering process that remove unnecessary data like junk characters and commands. The Ranking algorithm ranks the products in relation to the sales count. The top list displays the product having greatest count number. Then Prediction uses the logic rule generation methodology (from the LLRG approach) which retrieves useful and mandatory data from reviews, commands and products original state and lastly comes the recommendation. The HRS enforces two techniques, namely, location based and the other being heterogeneous domain based. Also the recommendation presented to the user is in context to the user's activities, choices and conduct that are in accord with user's personal likings and aids in decision making. For granting personalized services on Internet Recommender Systems are useful and important. Keeping in account the user's preference and recommending products based on that have been under long period of research followed by many approaches. The benefit of this approach lies in presenting the data in a visual classification format relying on given structure and remarkable size reduction in the search space per result. Also with this approach any product can be searched anytime and anywhere. By analyzing the reviews, ratings and emoticons they can be organized under good/positive and bad/negative feedbacks. Though Product Classification still remains as issue in recommender systems, this lies in building a quick and dynamic method for product classification which can aid in online shopping.

Following is the journal classification. Section 2 briefly elaborates former author's work. Section 3 proposes retrieving personal information, prediction, Hybrid recommendation system and the outlook of different stages. Section 4 displays test results. The paper concludes with, Section proposing future research study.

II. Literature Survey

Concerning news recommendations news that is available from only one website is more appropriate rather than coming from various websites. The authors suggest a hot news recommendation model that rely Bayesian model, taking in account large number of news websites. The model judges if the news is hot or not by computing the joint probability of the news. The suggested recommendation model is computed and then compared with the results of human experts on real time data sets. In year 2013, this model was also implemented in hot news recommendation system of Hangzhou city government, and the results were pretty good [6] [7].

The intelligent recommender system utilizes knowledge, comprehends, explores new data, draw out criticisms and preferences, knowledge representation paradigm, learning methodologies and reasoning mechanisms. Present work implements one intelligent recommender system based on the Fuzzy Cognitive Maps (FCMs). Next, to evaluate the performance and versatility execution of the intelligent recommender system is tested based on specialized criteria making use of the knowledge [8]. In [9] the Team Recommender Systems (TRS) is proposed. It's a knowledge based RS helping the organizations to build up a team in order to perform work that requires several skills. It helps in solving two main issues.

The reputation of Recommender System has raised, helping the researchers to trace out papers related to them amidst a huge collection. All the more, recommendation methods like collaborative-filtering or content-based restrict the user from explicitly providing any personal data. The work presents a customized Efficient Incremental High-Utility Item set Mining algorithm (EIHI), introduced recently in the survey which is designed to work along with dynamic datasets [10].

Yun Wan et al. (2015) [11] suggested an Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis by customer reviews. In the suggested system ensemble sentiment classification strategy was implemented on the basis of Majority Vote principle of multiple classification methods, including Naive Bayes, SVM, Bayesian Network, C4.5, Decision Tree and Random Forest algorithms. Hazem M. El-Bakry et.al (2016) [12] presented an efficient hybrid system for breast cancer document classification. It combines K-means clustering algorithm, fuzzy rough feature selection (FRFS), and Discernibility nearest neighbor classifier. It is claimed that the proposed model stands out with accuracy up to 98.9%. Anita kumar (2015) [13] suggested data mining classification methods implemented for cancer data twittering Perpetuation making use of cancer data set. Following Classification techniques such as CART, Random Forest, LMT, and Naive Bayesian are made use of. The best accuracy is provided by Random forest in comparison with other algorithms. Hakizimana Leopordet.al (2016) [14] suggested survey and analysis for current techniques on both classification and regression models that are implemented for product outbreak prediction in datasets.

There is a proposed movie recommendation system. People usually seek movie recommendation which can help out them with features like list of movies suggestion based on user's interest, or movie popularity wise. The system mines movie databases to fetch the required information that is, popularity and attractiveness, which further helps for recommendation [15]. MOVREC [16] is a movie recommendation system presented by D.K. Yadav et al. That relies on collaborative filtering approach. This approach utilizes the information provided by the user. Also there is a provision for the user to check mark the features of the movie to be recommended. Luis M Capos et al. [17] has proposed two traditional recommender systems they are content based filtering and collaborative filtering. Both these systems have few shortcomings as a result he suggested another system that's a combination of Bayesian network and collaborative filtering.

Harpreet Kaur et al. [18] have proposed a Hybrid System. The system uses a blend of content and also the collaborative filtering algorithm. The relationships that are user - user and user - item hold significance in the recommendation. The information's that is user related or item related is put together to build a cluster by Utkarsh Gupta et al. [19] using chameleon. This effective methodology relies on the

Hierarchical clustering for recommender system. Voting mechanism is used for the prediction of rating any product. Such system has reduced error rate and performs good clustering of same items. Urszula Kużelewska et al. [20] stated clustering as a means to work with recommender systems. Two techniques of computing cluster representatives were suggested and computed. Centroid-based solution and memory-based collaborative filtering methods were implemented for comparing effectiveness of the mentioned 2 methods. When compared to centroid based method, the outcome resulted in noticeable hike in the accuracy of the generated recommendation. Costin-Gabriel Chiru et al. [21] suggested Movie Recommender that makes use of the user information to provide movie recommendations. It's a hybrid model which implements both content based filtering and collaborative filtering.

The product added by the seller is predicted by the machine for cost prediction with Naïve Bayer method. The user thus can do online trading without even predicting the cost for C2C student trading [22]. Search and recommendation confronts the issue of Matching that is to compute the relatedness of a document to a query or users interest in a product. Earlier, to address this problem machine learning techniques were made use that learns a matching function from labeled data, known as "learning to match" [23]. Lately, deep learning is being implemented in Matching and reportedly there has been remarkable success. Search uses deep semantic matching models [24] and Recommendation makes use of neural collaborative filtering models [25] which has become the state-of-the-art technologies. A successful deep learning approach relies on its capability in learning of representations and generalization of matching from raw data (e.g., users, items, queries and documents especially in raw format).

The most often implemented algorithms are association rule mining that focus on the following: Apriori and FP-tree. Apriori generates quite a huge number of rules; whereas the FP-tree algorithm produces just a main tree and additional trees. With such tree generation the experimenter can get puzzled. Hence a concept is suggested in the present work in which the optimal rules are selected for the applications from both set of rules [26]. Current work is analyzed that takes information from user item interaction logs that are sequentially-ordered in the recommendation process. It is thereby suggested to categorize the recommendation works and targets, outline present algorithmic solutions, analyze methodological techniques when gauzing sequence-aware recommender systems, and portray open challenges in the concerned area [27]. In coordination with mobile e-commerce, this approach utilize an enhanced radial basis function (RBF) network to identify the weights of recommendations, and an enhanced Dempster–Shafer study merging the multi-source information. Thereafter Power-spectrum estimation is applied to deal with the fusion results and helps in decision-making. The test results depicts that the traditional approach is lower compared to suggested method related to simplicity, recommendation recall rate accuracy and coverage rate [28].

III. Proposed Work

A. Overview

The HRS techniques squeeze the options of available items that are in large number and propose the best ones, depending on the internal/external user ratings. A Recommender system can be considered as a system comprising of: user interface, a user, a dataset and few recommendation techniques. It can be stated that "Recommender systems are for the user, from the user and by the user" as the recommendations are derived implicitly or explicitly from the user ratings and thereafter presented to the users itself. Hence users can be considered as the lifeline of Recommender systems. The dataset composes of feedbacks or user ratings. At last, after receiving the required information a recommendation algorithm is implemented. Fig. 1 outlines the working structure of recommender systems.

B. Web Mining

In customer relationship management (CRM), Web mining combines the algorithms and methodologies to fetch information from traditional data mining and from the World Wide Web. (Mining actually refers to extraction of something which is valuable or useful, like mining gold or coal from the earth.) Web Usage Mining basically applies various data mining techniques to withdraw useful patterns from Web data to comprehend and fulfill the requirements of Web-based applications. Usage data detects the Web user's identity also how their browsing pattern and conduct on a website. The concept of Web mining uses various data mining methods and algorithms to withdraw information straight away from the Web, fetching it from Web content, Web documents and services, server logs and hyperlinks. Web mining aims to draw out patterns in Web data by gathering and examining information to have an overview of current trends, organizations and users in common.

C. Data Gathering

The surveyed data is manually fed in the system and further acts as the dataset. Admin gathers all the reviews and have a wide collection of them. They can be in form of text, rating and similar reviews. For any future evaluation all the reviews like ratings and emoticons are stored in the database. Reviews, Ratings and Emoticons form the evaluated data concerning quantity, quality (like rating a novel) or sometimes combination of both. Recommendation of a product using product based datasets is gathered from the UCI datasets via online [29].

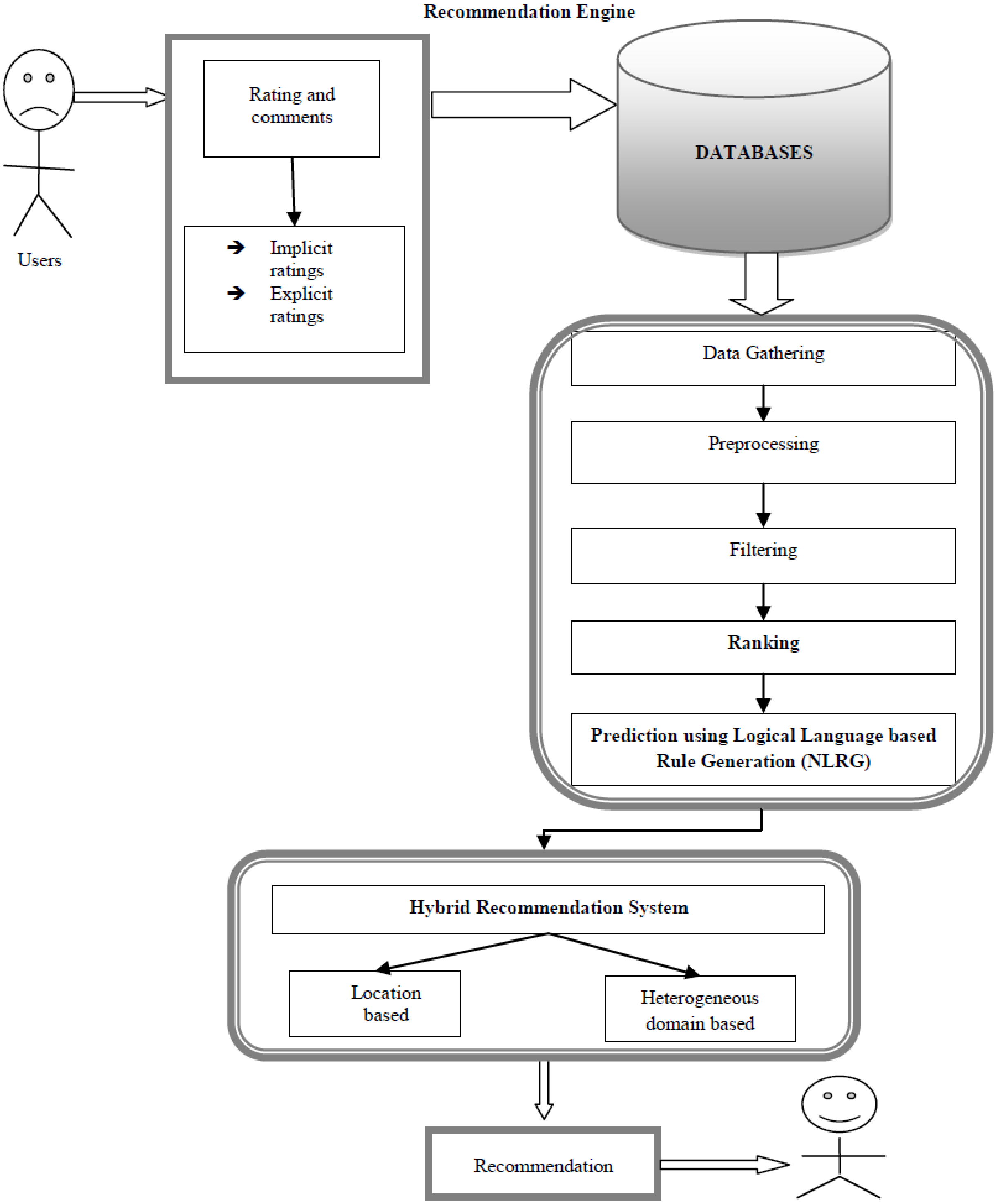


Fig 1. Overall Proposed Recommendation Engine.

How to fetching the personalized information:

Algorithm:

```
extractInfo ← User personalization(User Id, Uname, age,
gender and emotions ,city and others );
Clear_Sentiment_Term ← String and characters;
Context_Term ← [PositiveContextTerm1..
PositiveContextTerm2] &&
                              [NegativeContextTerm1,..
                         NegativeContextTerm2,, ]
                       ←[GoodContextTerm1..
GoodContextTerm2,,] &&
                         [BedContextTerm1,..BedContextTerm2,..]
                          ←[MixRatingContextTerm1..
             MixRatingContextTerm2,,] &&
                         [MinRatingContextTerm1,..
MinRatingContextTerm2,..]
                          ←[LogoSymbolsTerm1..
             LogoSymbolsContextTerm2,,] &&
                         [LogoSymbolsTerm3,..
LogoSymbolsTerm4,..]
Valid language ==”en”;
ContextVector ← CreatModels(ContextTerm);
Declaration ←GetConceptNetDeclration(Clear_Sentiment_Te
rm);
ValidDeclaration ← Apply extractionRules(Declation,
stages=”identifysenses” ,lang=      validLaunuages,
Context_Term=”String”)
Senses ←extractsenses(Valid_declaration)
senseDirectly=0;
for all CurrentSenses in sensed Do
Declaration ←GetConceptNetDeclration(Clear_Sentiment_Te
rm && Current senses);
ValidDeclaration ← Apply extractionRules(Declation,
stages=”identifysenses” ,lang=      validLaunuages,
Context_Term=”String”, Logo=”manually”)
senseTaskTerm ←extract Textual _Information( Valid_
declaration ,Text && Manully);
senseVectorModel ←CreateVectorModel(senseTaskTerm)
Existing_Emotions ←contextVector.SensesVector;
Current_Emotions
←ContextVector.SensesVector.Exsiting_emotions(Valid_
declaration);
Sense[Task(Sense)] ←crrent_emotions;
End For
Return  sense_dirctly,GetMax_Emotion(SensesDictly);
```

D. Preprocessing

The real or actual data is often noisy, inconsistent and fragmented. The preprocessing technique aids in filling up the unspecified data, smooth out noise by detecting outliers, and make the data consistent. The missing values can be populated manually or by making use of global constant. Then the undesirable data is eliminated from the original content. Data inconsistencies found in any transactions can be rectified manually with the help of external references. Also data integration may lead to some inconsistencies, say an attribute having various names in different databases or similar data value is referred by various names. Hence in preprocessing complete raw data is fetched from personal reviews and commands which aim to eliminate any null values, void columns and noisy data.

E. Filtering

Filtering is responsible to display the data as you desire. It’s helpful to review and command attributes to filter unwanted data. In Information Technology Data filtering provides a variety of solutions or techniques for filtering data sets. In simple terms data sets are refined according to user’s requirement or need, excluding any irrelevant or duplicate data. Hence filtering aids in eliminating unnecessary data thereby reducing size of dataset for further processing (Table 1).

F. Ranking

Ranking of a product is performed on the sale count value at which the promised items are rated. Next the promising items are found. In the Ranking algorithm products are ranked according to the sales count. The product ID acts as the primary or main key for the identification of the products. When user ID visits a particular product ID, the count linked associated with product ID is incremented by one. With re ranking algorithm products are ranked according to the rank count. Product having the greatest count number is included in the top most lists for the users. The theory behind this is that the similarity or resemblance among various entities in the dataset is computed with help of similarity measures. The similarities are picked depending on the ratings, reviews, users’ preferences, likes, dislikes, actions and comments. The Ranking model needs a vast dataset. Rating is on the basis of users’ visits on a product. Hence the product ID having the greatest visit count by the user is included at the top of the product list.

TABLE I. DATASET PARAMETERS (ATTRIBUTES)

S. NO	PARAMETERS	DESCRIPTIONS
1	IP address	System address
2	Time	Specific time
3	Name	Users name
4	Age	18+
5	Gender	Men/women
6	Location	Villagers/ city/ others
7	Product info	Product type
8	Activity	Active/ active less
9	Commands/reviews	Text, stars, logos

G. Prediction using Logical Language based Rule Generation (NLRG)

Product prediction analysis indicates to the use of logical language rule generation process, text analysis, and computational linguistics to identify, extract, quantify, and study affective states and subjective information in a systematic way. The review analysis is implemented to commands of the customer materials such as reviews and ratings for applications that vary from marketing to customer service to purchasing the products effectively. The products predict module is enforced to add the product with the 'Product ID that's considered as a word token. The PRODUCT ID keeps a track of user visits related to the product. This method records the user visits. The product id and visits done related to it are stored in the database. The product id key is incremented by one, which makes the product appear as most visited one. As the product ID count keeps on increasing, product with greatest count appears at the top position. The increase in product count is one of the ways of prediction.

Algorithm

```
Input: 1. User-id for the end-user for whom the suggestion
&& reviews will be given.
        2. Record Dataset Models
        3. Threshold values (LOW-LIMIT, HIGH-LIMIT) for
determining the
Prediction Results
        4. Number of ranked output, suggestion, commands;
Output: The list of predicted the product for recommendations
by their weights (or counts).
Initialize P = find -track-data-count (),
Initialize Q = find-max-track- sales-count (training-data-set);
Initialize R = P / Q (R← sales count ratio);
Initialize Product_RANKED_RESULTS =
find_PPrediction_Results (Dataset, user_id,
number_of_recommendations, commands);
Initialize Manully_RANKED_RESULTS =
find_Manully_Results(CaseBase, user_id,
number_of_recommendations, commands and logos);
If (R == 0) then {
                                                        //
EXTREME IMPORTANCE
P = 0; Q=100;
} else if (0 < R < LOW-LIMIT) then {
                                                        //
STRONG IMPORTANCE
P = 10; Q=90;
} else if (LOW-LIMIT ≤ R < HIGH-LIMIT) then {
                                                        //
EQUAL IMPORTANCE
P = 50; Q = 50;
} else if (HIGH-LIMIT ≤ R < 1) {
                                                        //
STRONG IMPORTANCE
P = 90; Q = 10;
} else if (R == 1) then {
                                                        //
EXTREME IMPORTANCE
P = 0; Q = 100;
}
Perdiction_RESUTS = MERGE (TOP P percent of
Prediction_RESULTS, TOP Q percent of Final
Prediction_RESULTS);
Return [highest count of product is displayed in the top
position]
End If
```

H. Hybrid Recommendation System

Each Recommender System has its own set of drawbacks; it's not possible to resolve all sorts of problems just with one method. As a result the Hybrid recommendation technique is taken into consideration. This technique merges two recommendation methods, one is the location based method and the other is heterogeneous domain method. Its purpose is to overcome the shortcomings like the cold-start problem, thereby enhancing the RS output and performance. The techniques of Hybrid recommendation is implemented and evaluated. Using this system it can be claimed that it produces results which exceeds single component systems by merging multiple set of techniques. Though the approach of Hybrid recommendation is a better option for various RS methods, there occurs for further efforts and in depth information to make this technique implement. The hybrid recommendation approaches combined for best product recommending as follows of Fig, 2.

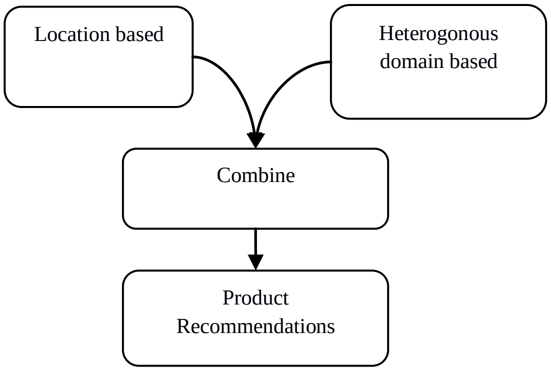


Fig 2. Recommendation System.

I. Location based

A social network can be elaborated as a system wherein people from various regions or locations of the world having different cultures, professions, age groups and social circle interact with one another to build up various relations like friends, global awareness and mutual concern. The locations are based on various parameters for example whether it's a small town or a big town or city. Depending on how big or small the location is, it can be organized in a hierarchical order in which locations that are small in geographical area are placed at bottom. Reviews and commands based on Location are fetched for further product recommendation. The product quality can be judged on basis of location based rating and commands.

J. Heterogeneous Domain based

Building up user product relationship domain wise can often be tedious because of the heterogeneous parameters. That is a product may be required to be recommended to any personal users but the products in these domains are referred overall in some other name or tag/feature. Generally Web tags classifies a product attributes viz. size, quality, advantages, disadvantages , availability etc. and few users may post their commands and review depending on the multiple domain for example in case of mobiles (its features, version, specialties, software, drawbacks, advantages, simplicity etc.) recommended for the users. Whereas, web tags are basically specifies a product price, production Nation, year and category.

Algorithm of hybrid recommendation system:

```
List of Recommendation (User t, List I, List U) {
Val featureVectorsUDF(vector, string, string, string, string){
Product _Id, sales _Rank, average _rating, Prediction _rating, categories, }
Prodcut_count(P_$C_{t,i}$) ← GetCartCount,GEtBuyCount,GetClickCount,GetSearchCount($_{t,i}$)
totalItem=TI;
TI=TI.allItem(List of Datas);
Temp = salerating=0.0, averagerating=0.0;
Try { sr=salerank.todouble } catch {
Case_Exception: →}
Try { ar=averagerank.todouble} catch {
Case_exception : →}
RecommendationList ←emptyset();
$L_t$ ← recommendationOntology.isLocation(t);
$H_t$ ←recommendationOntology.isHdomains(H);
$P_t$ ←Recommendationontology.hasPerformance(P);
useSimilarityMAP ←userSmil(t,i,U);
for (Item i=I) {
$Pr_i$ ←recommenderontology.hasProperty(i);
$comPr_i$ ← commanPerformanceProperty(pt, $pr_i$);
if ($cpprn_{t,I}$ > 0) {
PPVti ←prefProValues($p_t$, $pr_i$,$cpprnt_{,i}$,i)
IVi ←item Values(i,i);
$USV_{t,i,u}$ ←userRatings(userSimilerMap,$_{r,u,i}$, GetRatingN(u,i)); }
$UV_{t, i}$ ← UserValues($USV_{t, i,u}$);
$HDV_{t, i}$ ←TrakingDifferentDomainValues(getTimePlace(t,i),getDifferntfeature(t,i))
getMostFeaturevalues(t), getDifferentDate(t,i))
$FV_{t,i}$ ←Final V(P_C $_{t,I,}$ $PPV_{t,i}$ , IV $_{t,i,}$ UV $_{t,i}$, USV $_{t,I}$, HDV $_{t,i}$);
recommendationItemList.Adding ($_i$,FV $_{t,i}$)
}}
SortList(MostRecommendationList)
Return RecommentdationList;
}
```

K. Recommendation

Recommender systems (RS) are a subset of information filtering system that aids in predicting the "preference" or "rating" given by the user to a particular item. The product is searched using search bar which results in a complete set of products related to cost and reviews. On the basis of users ratings and reviews the positive or right products are put forth in recommendation panel. The product recommendation system, that is location and heterogeneous based displays the top list product. The issue of recommendation lies for those users who have not purchased any product. For such users product having greatest count is listed at the top. The top most products are highlighted based on the product id and its count. Every time the product is visited, the product id count is incremented by one. As a result product having greatest visit count is highlighted at the top list.

L. Evaluation Metrics

This paper considers large data based evaluation and large volume of data to be tested providing efficient results. Recommender Systems are also considered to examine large Data, keeping in accord some main tests related with Big Data analytics. System stability evaluation and performance can be computed using parameters that are evaluated and examined. Few of them are as given below:

The suggested Personalized recommended system performance is computed with Root Mean Square Error (RMSE), Recall, precision, F-score, probability of the misclassification error (PME) and accuracy of the training set, testing set and complete performance was examined by using the Equation (1-6) respectively where Y_i is actual and R_i is the result of the i^{th} recommended to the obtained,

- True Positive (TP): If the case in point is positive (recommended results) and it is recommended as positive.
- False Negative (FN): If the case in point is positive (recommended results) but it is recommended as negative.
- True Negative (TN): If the case in point is negative (recommended results) and it is recommended as negative.
- False Positive (FP): If the case in point is negative (recommended results) but it is Recommended as positive

A general way to compute a Recommendation is to evaluate the deviation of the recommended from the true or real value. This forms the basis for the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - R_i)^2} \tag{1}$$

Where N relates to set of all user-item pairings (i) for which rating is predicted as ˆR_i and a known rating R_i which was unused to learn the recommendation model. The other wel-known measure is the probability of the misclassification error (PME).

$$PME = \sqrt{\frac{\sum_{i \in N}(R_i - \widehat{R_i})^2}{|N|}} \tag{2}$$

Performance measures used for computation of these algorithms are deeply rooted in machine learning. Most popularly used measure is accuracy, the fraction of correct recommendations to total possible recommendations.

TABLE II. CONFUSION MATRIX

Actual/Predicted	**Negative**	**Positive**
Negative	True Positive	True Negative
Positive	False Negative	False Positive

Table 2, shows the confusion matrix for to detecting results. From the confusion matrix various performance measures can be extracted. In the data mining method of a recommender system the algorithms performance relies on its capability to imbibe significant patterns in the data set. Most popularly used measure is accuracy, the fraction of correct recommendations to total possible recommendations.

$$Accuracy\ = \frac{Correct\ Recommendations}{Totla\ Possible\ recommendations}$$

$$= \frac{TP + TN}{TP + FP + TN + FN} \qquad (3)$$

$$Recall\ \ = \frac{Correctly\ recommended\ Items}{Total\ useful\ recommendation}$$

$$= \frac{TP}{TP + FN} \qquad (4)$$

$$Precision\ = \frac{Correctly\ recommended\ Items}{Total\ recommended\ items}$$

$$= \frac{TP}{TP + FP} \qquad (5)$$

Popular single-valued measure is the F-measure which is defined as the harmonic mean of precision and recall.

$$F - score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

$$= \frac{2}{1/Precision + 1/Recall} \qquad (6)$$

M. Advantages

- Recommended is more precise, fats and accurate.
- Nature of outcome.
- Reduced time complexity.
- Detecting quality and best product easily.

IV. RESULTS AND DISCUSSION

The product recommendation making use of product based datasets were gathered from the UCI datasets via online [29]. This work have been framed and implemented in Big data Domain with large volumes of datasets to be computed successfully and effectively. Tests were carried out in concern with the following specifications: Windows 7, Intel Pentium (R), CPU G2020 and processor speed 2.90 GHz respectively. The required software specifications are given below, Operating System→Windows 7, Front End→JAVA, Back End→MYSQL.

The research paper implementation is performed on big data domain and JAVA based environment and entire dataset is managed using MYSQL databases and the resultant output is stored in MYSQL databases. Since there is a massive quantity of information loaded on INTERNET, it becomes difficult for the user to search requires and necessary information. Luckily, based on behavior of any user, their likings and priorities can be judged. The efficiency of the suggested method can be computed by conducting a set of tests. Considering the following experiment, the Logical Language based Rule Generation (NLRG) is being implemented to detect the ratings of the non-rated products. In the recommendation systems the effectiveness of the suggested method is tested using the hybrid recommendation system by judging it with existing recommendation methods. The comparison parameters are: recommendation accuracy, simplicity, processing time and error rate.

TABLE III. COMPARISON OF OVERALL PERFORMANCE

S No	**Algorithm**	**Accuracy (%)**	**Simplicity**	**Time (MS)**
1	Traditional algorithm	80.01	2.01	0.934
2	Radial Basis Function (RBF) Neural Network	91.23	1.47	0.803
3	Hybrid Recommendation system	94.21	1.03	0.521

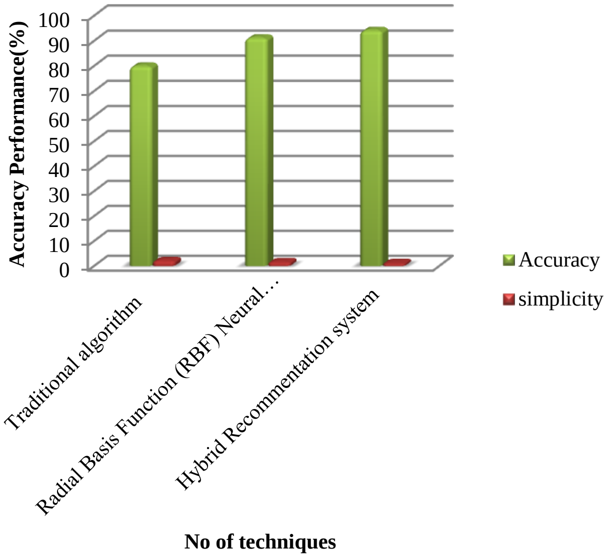


Fig 3. Comparison of Accuracy Performance.

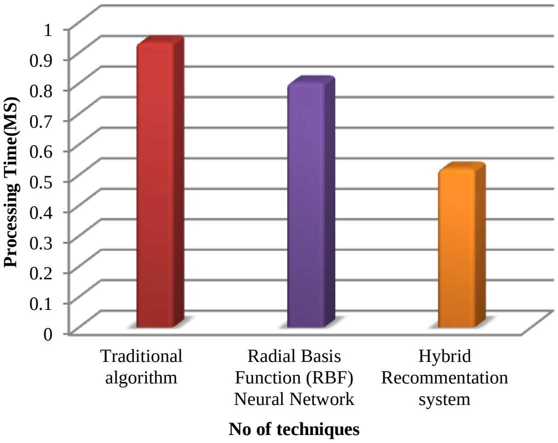


Fig 4. Comparison of Processing Time Calculation.

Fig. 3 and 4 depicts the comparison of product recommendation Techniques output with the overall performance of product recommendation and presenting comparison with various other existing techniques namely, Traditional algorithm, Radial Basis Function (RBF) based Neural Network and Hybrid Recommendation system. Then the research aims to evaluate the complete performance factors like accuracy, simplicity and time. The suggested product recommendation using Location based and heterogeneous based Hybrid Recommendation system performs better with good output compared to other existing techniques.

Fig. 5 depicts the Prediction Techniques performance over product recommendation and displaying comparison with various other existing techniques C5.0, Numerical Prediction and Deep Learning Prediction and Logical language based rule generation (LLRG) method. The suggested product recommendation of prediction using Logical language based rule generation (LLRG) Prediction method reveals better performance and output compared to other existing techniques.

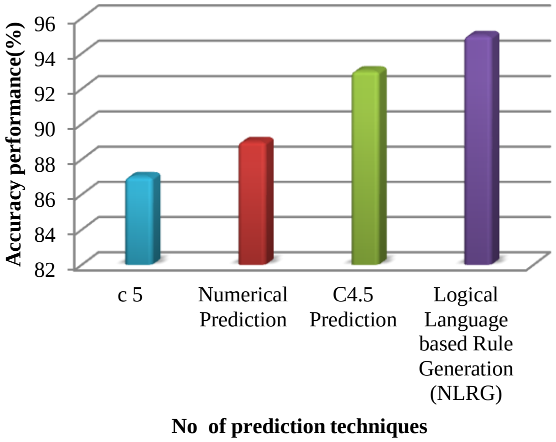


Fig 5. Comparison of Prediction Techniques.

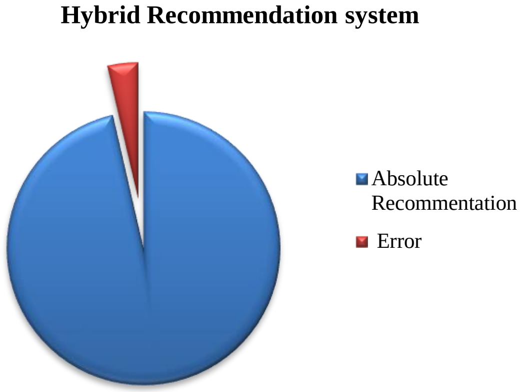


Fig 6. Error Performance.

Fig. 6 is depicting the Error Performance for recommendation Techniques output over the product recommendation revealing 96.4% accuracy and error as 3.6%.

V. CONCLUSION

The prime agenda of the study is to offer a new approach to enhance hybrid recommender systems for users. Based on the available types of items, prioritization of the suggested items is enhanced and thereafter made available to the user. The research paper aims at the enhancement of the hybrid recommendation system making use of web mining techniques to rectify the shortcomings of the cost and time prediction using recommendation algorithm, to judge the products cost by the machine related to the previously stored data. The method of Location based and heterogeneous based Hybrid Recommendation system yields 94.21%. Traditional algorithm has 80.1%, Radial Basis Function (RBF) based Neural Network has 91.23% derived from analyzing and evaluation. The outcome produced presents that apart from higher speed of processing and accuracy it provides a remarkable improvement. The work evaluated that the personalized hybrid recommendation system is delivering better performance and output compared to other existing recommendation techniques.

FUTURE WORK

Choose the best cloud web services for specific applications are very challenging. This is because there are many services with similar functionalities but varying non functional properties. This includes the use of selected quality and best of cloud service properties coupled with user feedback data to determine the most suitable service. In future work a recommended mechanism for selecting the best cloud web service at the levels of cloud computing environment for personalized users.

REFERENCES

[1] Ricci, Francesco, Lior Rokach, and Bracha Shapira, “Introduction to recommender systems handbook”, Springer US, 2011.

[2] Jannach, Dietmar, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich, “Recommender systems: an introduction”, Cambridge University Press, 2010.

[3] Ekstrand, Michael D, John T. Riedl, and Joseph A. Konstan, “Collaborative filtering recommender systems”, Foundations and Trends in Human-Computer Interaction, Vol. 4, no. 2 (2011): 81-173.

[4] Jannach, Dietmar, and Gerhard Friedrich, “Tutorial: recommender systems”, In Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona. 2011.

[5] Yazdanfar, Nazpar, and Alex Thomo, “Link recommender: Collaborative-Filtering for recommending URLS to Twitter users”, Procedia Computer Science 19 (2013): 412-419.

[6] Zhengyou Xia, Shengwu Xu, Ningzhong Liu, and Zhengkang Zhao, “Hot News Recommendation System from Heterogeneous Websites Based on Bayesian Model”, Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 734351, 8 pages.

[7] Cai Y. fung Leung H. Li Q. Min H. Tang J. Li J, “Typicality-based collaborative filtering recommendation, IEEE Trans. Knowl. Data Eng., 26 (3), 766–779, 2014.

[8] Jose Aguilar, Priscila Valdiviezo-Dı´az , Guido Riofrio,” A general framework for intelligent recommender Systems”, Available online 21 September 2016, Applied Computing and Informatics (2017) 13, 147–160.

[9] M. Ayub, A. Cian, M. Caliusco, E. Reynares, “Developing an ontology-based team recommender system using EDON method: an experience report, SADIO: EJIOR, Vol. 13 (2014), pp. 1–13.

[10] Mahak Dhanda and Vijay Verma,” Recommender System for Academic Literature with Incremental Dataset”, Twelfth International Multi-Conference on Information Processing 2016(IMCIP-2016), © 2016 The Authors. Published by Elsevier B.V.

[11] Yun Wan, Dr. QigangGao, “An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis”, 2015 IEEE 15th International Conference on Data Mining Workshops.

[12] Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh, “Classification of Breast Cancer Using Softcomputing Techniques”, International Journal of Electronics and Information Engineering, Vol.4, No.1, PP.45-54, Mar. 2016.

[13] Anita kumar, “A Study on Cancer Perpetuation Using the Classification Algorithms”, International Journal of Recent Research in Mathematics Computer Science and Information Technology Vol. 2, Issue 1, pp: (96-99), Month: April 2015 – September 2015, Available at: www.paperpublications.org

[14] Hakizimana Leopord, Dr. Wilson KiprutoCheruiyot, Dr. Stephen Kimani, “A Survey and Analysis on Classification and Regression Data Mining Techniques for Diseases Outbreak Prediction in Datasets”, The International Journal Of Engineering And Science (IJES), Volume. 5, Issue. 9, pp. 01-11, 2016.

[15] Geetha G,Safa M , Fancy C , Saranya D, “A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System”, National Conference on Mathematical Techniques and its Applications (NCMTA 18).

[16] Manoj Kumar, D.KYadav, Ankur Singh, Vijay Kr. Gupta, “A Movie Recommender System: MOVREC”, International Journal of Computer Applications, Volume 124 – No.3, August 2015.

[17] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Miguel A. Rueda-Morales, “Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks”, International Journal of Approximate Reasoning, revised 2010.

[18] Harpreet Kaur Virk, Er. Maninder Singh, “Analysis and Design of Hybrid Online Movie Recommender System”, International Journal of Innovations in Engineering and Technology (IJIET) Volume 5 Issue 2,April 2015.

[19] Utkarsh Gupta1 and Dr Nagamma Patil, “Recommender System Based on Hierarchical Clustering Algorithm Chameleon”, 2015 IEEE International Advance Computing Conference (IACC).

[20] Urszula Kuzelewska, “Clustering Algorithms in Hybrid Recommender System on MovieLens Data”, Studies in Logic, Grammar and Rhetoric, 2014.

[21] Costin-Gabriel Chiru, Vladimir-Nicolae Dinu, Ctlina Preda, Matei Macri, “Movie Recommender System Using the User's Psychological Profile”, IEEE International Conference on ICCP, 2015.

[22] Megha K, Ms. P Devaki, “Web Application for student trading using Data Mining Techniques”, International Journal of Scientific Development and Research (IJSDR), 2018, IJSDR, Volume 3, Issue 5.

[23] Hang Li and Jun Xu. 2014, “Semantic Matching in Search. Foundations and Trends” Information Retrieval, Vol. 7, 5 (2014), pp. 343–469.

[24] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng, 2017, “DeepRank: A NewDeep Architecture for Relevance Ranking in Information Retrieval”, International Conference on Information and Knowledge Management (CIKM’17).

[25] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017, “Neural Collaborative Filtering”, International Conference on World Wide Web (WWW ’17), pp. 173–182.

[26] Devyani Ojha, Pragya Pandey”, “Optimizing Association Rule using Genetic Algorithm and Data Sampling Approach”, International Journal of Computer Applications, Volume 179, No.11, January 2018.

[27] Massimo Quadrana, Paolo Cremonesi, Dietmar Jannach, “Sequence-Aware Recommender Systems”, ACM Computing. Survey, 2018, 35 pages.

[28] Yan Guo, Chengxin Yin, Mingfu Li, Xiaoting Ren, and Ping Liu, “Mobile e-Commerce Recommendation System Based on Multi-Source Information Fusion for Sustainable e-Business”, 2018, Vol. 10.

[29] https://archive.ics.uci.edu/ml/datasets.html

Integration of REST-Based Web Service and Browser Extension for Instagram Spam Detection

Antonius Rachmat Chrismanto[1], Willy Sudiarto Raharjo[2], Yuan Lukito[3]
Program Studi Informatika, Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Yogyakarta, Indonesia

***Abstract*—In this paper, a REST-based Web Service developed in previous work was integrated with a newly developed browser extension that works in modern browser (Firefox and Google Chrome) using Greasemonkey. It uses previous collected datasets which comprised of 17.000 postings and comments from 10 Indonesian actresses whom followers are more than 10 million on Instagram. The performance of the developed web services has been evaluated and the average response time is 1678.133ms using AWS platform located in Ohio (US East 2). The proposed work is working as expected and in accuracy test, it has reached 63.125% in overall, 72% for non-stemmed data and 70% for stemmed data using 1000 test data with a processing time needed for classification is under 2s. The new extension works in Firefox and Chrome and it can utilize the web services to classify spam comments in Instagram.**

Keywords—Instagram; spam comments; REST service; web service testing; browser extension

I. Introduction

Social media is no longer just a mean for sharing information along relatives and colleagues, but it has transformed into a bigger scope and touching every aspect of human life. Social media is already used in many situations, like emergency situation [1], traveling [2], and health [3]. However, it comes with a price. According to [4], [5], and [6] there are a lot of spam comments in media social, such as YouTube, Facebook, Twitter, and Instagram. These spammers may cause some information misleading, mixed information, wasting valuable network resources, and decreasing the quality of online social networking sites [7], [8], and [9].

Nowadays, most people are using Instagram because of its characteristic of being an image-based social media. A picture speaks for thousand words by nature. According to [10], Instagram has reached 1 billion monthly users in June 2018, a significant raise from 800 million in September 2017. It shows that Instagram is gaining a huge popularity among many people, including Indonesian actress who proactively engaged with their fans to help them gain more popularity and brings more business opportunities for them.

Instagram is gradually introducing new features as posted in their press web sites (https://instagram-press.com/), but rarely seen a posting about spam detection. One of the reasons is that because spam may come in many ways and sometimes it’s context-based, so it’s hard to find a good balance for creating an algorithm that can detect spam comments nowadays, especially in Indonesian language. There is no implemented solution for automated Indonesian language spam detection in Instagram yet. Many previous work [11], [6], [12] used Instagram data for spam detection, but so far, there are no real implemented solution for spam detection. The research done so far was more focused on testing the accuracy of each model. Especially on Indonesian-based language, which according to [13] is still considered as one of the resource-poor languages.

In this paper, an implemented solution for automated Indonesian language spam detection is proposed by building an integration between a REST-based web service and a browser extension that can be used to detect Instagram spam comments in Indonesian language. This research contributes in enriching Indonesian language related researches and creates a ready to use Instagram spam detector. Browser extension is the option we chose since it allows us to interact with the content on Instagram without breaking same-origin policy [14].

II. Related Work

Hardinata and Tirtawangsa [11] developed spam detector in Indonesian Twitter trending topics. The spam detector works by detecting spam that utilized trending topics hashtags. The spam detection process involved human input that collected using monster game interface. Zhang and Sun [15] has published their work on a model to decrease number of spam posts in Instagram, but only applicable for English language. Ali and Okiriza [12] published their work on detecting spam comments on Indonesia’s Instagram post using three different algorithms: Naïve Bayes, SVM, and XGBoost. They concluded that SVM and XGBoost got the best scores of 0.9601 and 0.9512. In all the researches, not a single of them proposed a real implemented and practical solution since they all are focusing on the accuracy of the models being tested.

This work was started in 2017 by building Indonesian spam comments detector using Naïve Bayes [16] and collected more than 25.000 postings and comments from Indonesian actress with more than 10 million followers. After data cleansing process, the final data used are 17.000 postings. From this datasets, some experiments were conducted using different algorithms and it was concluded that K-Nearest Neighbors (k-NN) gave the best results with 88.4% of accuracy [17], followed by Support Vector Machine with 78.5% [18], and Naïve Bayes with 75.5% [16].

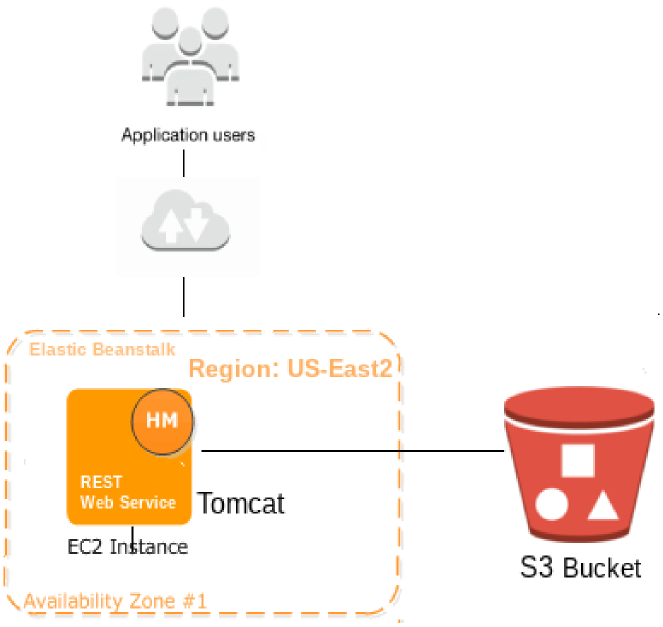


Fig. 1. Web Service Architecture.

Next, a REST-based web service to detect Indonesian language Instagram spam comments using k-NN algorithm design was designed and deployed on top of AWS platform and evaluate the performance based on response time. [19].

III. Methodology

A. Architecture

This work is using the same AWS architecture that was developed in previous work [19] for the web service architecture, which was deployed on US-East 2 region (Ohio). The system is using Tomcat as the main web server and all datasets are stored in the S3 bucket for durability and performance reason. The architecture is illustrated in Fig. 1.

The web service does not deploy SSL certificate for this machine as there are no confidential data that are communicated, and the system never stored any data transmitted to the server during spam detection process. The dataset is stored in the S3 bucket which is only accessible via the web server and not directly accessible for public.

All the communication between client (browser) and the server will be done using REST [20] which has some advantages over SOAP such as better throughput and response time, as demonstrated on [21] and [22].

B. Algorithm

In this work, k-NN algorithm is used based on previous work [17] that gives best results compared to other algorithms (Support Vector Machine [18] and Naïve Bayes [16]). K-NN is learning directly while performing classification process by finding some adjacent data object or patterns based on the input and choose a class with the highest number of patterns [19]. K-NN can be implemented as follows (Fig. 2):

1) Load the data

2) Initialize the value of k

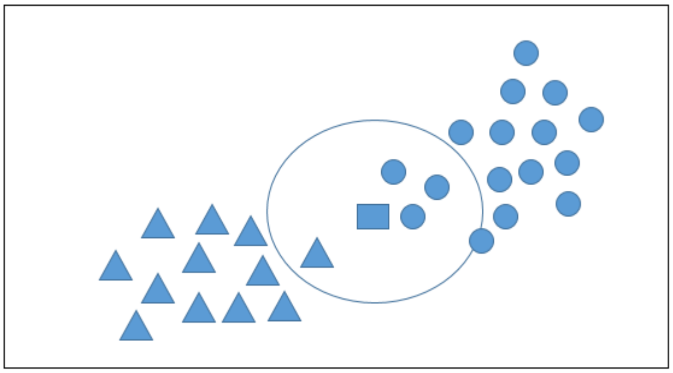


Fig. 2. k-NN Algorithm

3) For getting the predicted class, iterate from 1 to total number of training data points:

a) Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.

b) Sort the calculated distances in ascending order based on distance values

c) Get top k rows from the sorted array

d) Get the most frequent class of these rows

e) Return the predicted class

C. Browser Extension

The browser extension is developed using Greasemonkey and works as follows:

- Script will check visited page. If it is coming from Instagram, it will add a new entry in the browser's context menu (accessed via right click)
- When user highlighted some text in Instagram posting, the extension will read the highlighted text and send it to the web service in AWS
- Web service will process the request and reply the results back to the browser
- Browser extension will display the results to user in form of a dialog box.

D. Evaluation

Several tests were conducted to evaluate some metrics. The first test was performed using SOAPUI tool which is used to perform load testing, method testing, simple load testing, burst load testing, thread load testing, variance load testing, and data-driven testing. It used 160 data for data-driven test.

The second test was testing the web service accuracy by using PHP scripts to automate the test. The test used 1000 random data taken from dataset using shuffled sampling. The dataset was generated using 10 smaller dataset which consisted of 100 data to reduce the slow processing time. Afterwards, it's merged with the rest. Next, the dataset is tested against 8 test datasets which have been stored in the web service already. The parameters used for the k-NN validation shown in Table 1.

TABLE I. TESTING PARAMETERS

Parameters	**Values**
Number of data	1000
Output criteria	Accuracy
Sampling type	Shuffled Sampling
Dataset type	8 types
Dataset Criteria	Unbalanced non-stemmed, unbalanced-stemmed, balanced non-stemmed, and balanced stemmed

The 8 datasets that were used are as follow:

- Generated using PHP
  - Unbalanced non-stemmed data
  - Unbalanced stemmed data
  - Balanced non-stemmed data
  - Balanced stemmed data
- Generated using R
  - Unbalanced non-stemmed data
  - Unbalanced stemmed data
  - Balanced non-stemmed data
  - Balanced stemmed data

IV. RESULTS AND DISCUSSIONS

A. Web Service Accuracy

The results of the web service accuracy after tested against 8 datasets can be seen on the Table II through Table IX.

TABLE II. DATASET 1: UNBALANCED-NON STEMED DATA GENERATED USING PHP

TEST DATA	**STEM**		**%**	**NON-STEM**		**%**
	T	***F***		***T***	***F***	
1-100	71	29	71%	73	27	73%
101-200	68	32	68%	68	32	68%
201-300	83	17	83%	83	17	83%
301-400	85	15	85%	85	15	85%
401-500	74	26	74%	73	27	73%
501-600	72	28	72%	72	28	72%
601-700	61	39	61%	61	39	61%
701-800	87	13	87%	87	13	87%
801-900	75	25	75%	75	25	75%
901-1000	58	42	58%	57	43	57%
			73%			73%

TABLE III. DATASET 2: UNBALANCED STEMMED DATA GENERATED USING PHP

TEST DATA	**STEM**		**%**	**NON-STEM**		**%**
	T	***F***		***T***	***F***	
1-100	66	34	66%	67	33	67%
101-200	53	47	53%	54	46	54%
201-300	49	51	49%	48	52	48%
301-400	44	56	44%	46	54	46%
401-500	43	57	43%	43	57	43%
501-600	58	42	58%	54	46	54%
601-700	44	56	44%	38	62	38%
701-800	55	45	55%	55	45	55%
801-900	45	55	45%	45	55	45%
901-1000	53	47	53%	55	45	55%
			51%			51%

TABLE IV. DATASET 3: BALANCED NON-STEMMED DATA GENERATED USING PHP

TEST DATA	**STEM**		**%**	**NON-STEM**		**%**
	T	***F***		***T***	***F***	
1-100	71	29	71%	73	27	73%
101-200	68	32	68%	66	34	66%
201-300	76	24	76%	75	25	75%
301-400	82	18	82%	83	17	83%
401-500	69	31	69%	69	31	69%
501-600	71	29	71%	72	28	72%
601-700	60	40	60%	60	40	60%
701-800	88	12	88%	88	12	88%
801-900	88	12	88%	88	12	88%
901-1000	65	35	65%	62	38	62%
			74%			74%

TABLE V. DATASET 4: BALANCED STEMMED DATA GENERATED USING PHP

TEST DATA	**STEM**		**%**	**NON-STEM**		**%**
	T	***F***		***T***	***F***	
1-100	69	31	69%	69	31	69%
101-200	55	45	55%	53	47	53%
201-300	51	49	51%	47	53	43%
301-400	48	52	48%	43	57	43%
401-500	50	50	50%	40	60	40%
501-600	48	52	48%	45	55	45%
601-700	62	38	62%	58	42	58%
701-800	53	47	53%	47	53	47%
801-900	48	52	48%	45	55	45%
901-1000	55	45	55%	51	49	51%
			54%			49%

TABLE VI. DATASET 5: UNBALANCED NON-STEMMED DATA GENERATED USING R

TEST DATA	**STEM**		**%**	**NON-STEM**		**%**
	T	***F***		***T***	***F***	
1-100	86	14	86%	86	14	86%
101-200	78	22	78%	77	23	78%
201-300	81	19	81%	84	16	84%
301-400	89	11	89%	86	14	86%
401-500	83	17	83%	83	17	83%
501-600	81	19	81%	77	23	77%
601-700	79	21	79%	79	21	79%
701-800	83	17	83%	85	15	85%
801-900	84	16	84%	84	16	84%
901-1000	75	25	75%	74	26	74%
			82%			82%

TABLE VII. DATASET 6: UNBALANCED STEMMED DATA GENERATED USING R

TEST DATA	STEM		%	NON-STEM		%
	T	*F*		*T*	*F*	
1-100	86	14	86%	86	14	86%
101-200	78	22	78%	77	23	77%
201-300	81	19	81%	84	16	84%
301-400	89	11	89%	86	14	86%
401-500	83	17	83%	83	17	83%
501-600	81	19	81%	77	23	77%
601-700	79	21	79%	79	21	79%
701-800	83	17	83%	85	15	85%
801-900	84	16	84%	84	16	84%
901-1000	75	25	75%	74	26	74%
			82%			82%

TABLE VIII. DATASET 7: BALANCED NON STEMMED DATA GENERATED USING R

TEST DATA	STEM		%	NON-STEM		%
	T	*F*		*T*	*F*	
1-100	83	17	83%	86	14	86%
101-200	77	23	77%	77	23	77%
201-300	82	18	82%	79	21	79%
301-400	87	13	87%	82	18	82%
401-500	82	18	82%	77	23	77%
501-600	77	23	77%	71	29	71%
601-700	77	23	77%	76	24	76%
701-800	83	17	83%	80	20	80%
801-900	83	17	83%	78	22	78%
901-1000	69	31	69%	66	34	66%
			80%			77%

TABLE IX. DATASET 8: BALANCED STEMMED DATA GENERATED USING R

TEST DATA	STEM		%	NON-STEM		%
	T	*F*		*T*	*F*	
1-100	84	16	84%	82	18	82%
101-200	80	20	80%	73	27	73%
201-300	84	16	84%	79	21	79%
301-400	84	16	84%	81	19	81%
401-500	80	20	80%	75	25	75%
501-600	83	17	83%	71	29	71%
601-700	86	14	86%	80	20	80%
701-800	83	17	83%	73	27	73%
801-900	83	17	83%	74	26	74%
901-1000	75	25	75%	65	35	65%
			82%			75%

B. *Web Service Comprehensive Testing*

a) Method Testing

This test is used to ensure the output of all the web service are according to what we expected in terms of formatting and the content itself. This is the simplest test but also crucial to be performed so that the system gives the same output as what it is expected to do. The result of the method testing can be seen on Table X. All the methods we developed have produced expected results.

TABLE X. METHOD TESTING RESULTS

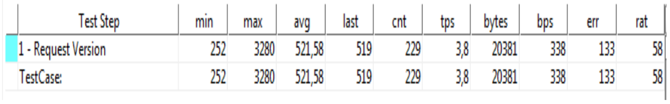
No	Method	Expected Results	Actual Results	Remarks
1	Version (GET)	JSON 200 OK	JSON 200 OK	Match
2	No (GET)	JSON 200 OK	JSON 200 OK	Match
3	Dataset (GET)	Text Plain 200 OK	Text Plain 200 OK	Match
4	File (GET)	JSON 200 OK	JSON 200 OK	Match
5	Classify (POST)	JSON 200 OK	JSON 200 OK	Match

b) Load Testing

This test is used to see how the system behaves under high load. The instance used in this work is t2 micro which only have 1 vCPU and 1 GB of RAM. In the first test, we used the Version method to represents GET method, with the following parameters:

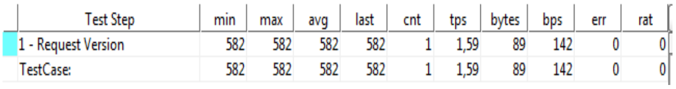
- Number of threads: 10
- Intervals: 10 s
- Variance: 0.5
- Time limit: 1 s
- Burst delay: 60 s
- Burst duration: 10 s

In load testing, there are 4 sub tests: simple, burst, thread, and variance. The result of the load testing are shown in Fig. 3, Fig. 4, Fig. 5, and Fig. 6.



Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
1 - Request Version	252	3280	521,58	519	229	3,8	20381	338	133	58
TestCase:	252	3280	521,58	519	229	3,8	20381	338	133	58

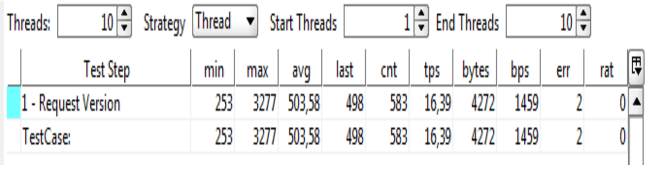
Fig. 3. Simple Load Testing Result.



Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
1 - Request Version	582	582	582	582	1	1,59	89	142	0	0
TestCase:	582	582	582	582	1	1,59	89	142	0	0

Fig. 4. Burst Load Testing Result.

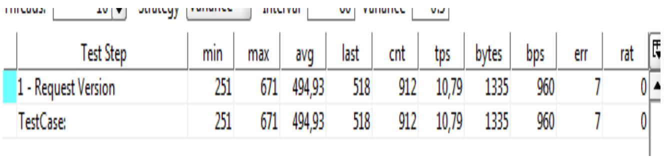
In simple load test, it has minimal request of 252 ms, maximum request is 3280 ms, and average request is 521,58 ms.



Threads: 10 Strategy Thread Start Threads 1 End Threads 10

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
1 - Request Version	253	3277	503,58	498	583	16,39	4272	1459	2	0
TestCase:	253	3277	503,58	498	583	16,39	4272	1459	2	0

Fig. 5. Thread Load Testing Result.



Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
1 - Request Version	251	671	494,93	518	912	10,79	1335	960	7	0
TestCase:	251	671	494,93	518	912	10,79	1335	960	7	0

Fig. 6. Variance Load Testing Result.

In burst load test, it has minimal request of 582 ms, maximum request is 582 ms, and average 582 ms.

In thread load test, it has minimal request 253 ms, maximum request is 3277 ms, and average is 503,58 ms. There are 11 requests that has more than 1000 ms (more than time limit of the system). The average request time is around 3 seconds.

In variance load test, it has minimal request of 251 ms, maximum request is 671 ms, and average is 494,93 ms. There are 9 requests that has more than 1000 ms (more than time limit of the system). The average request time is around 3 seconds.

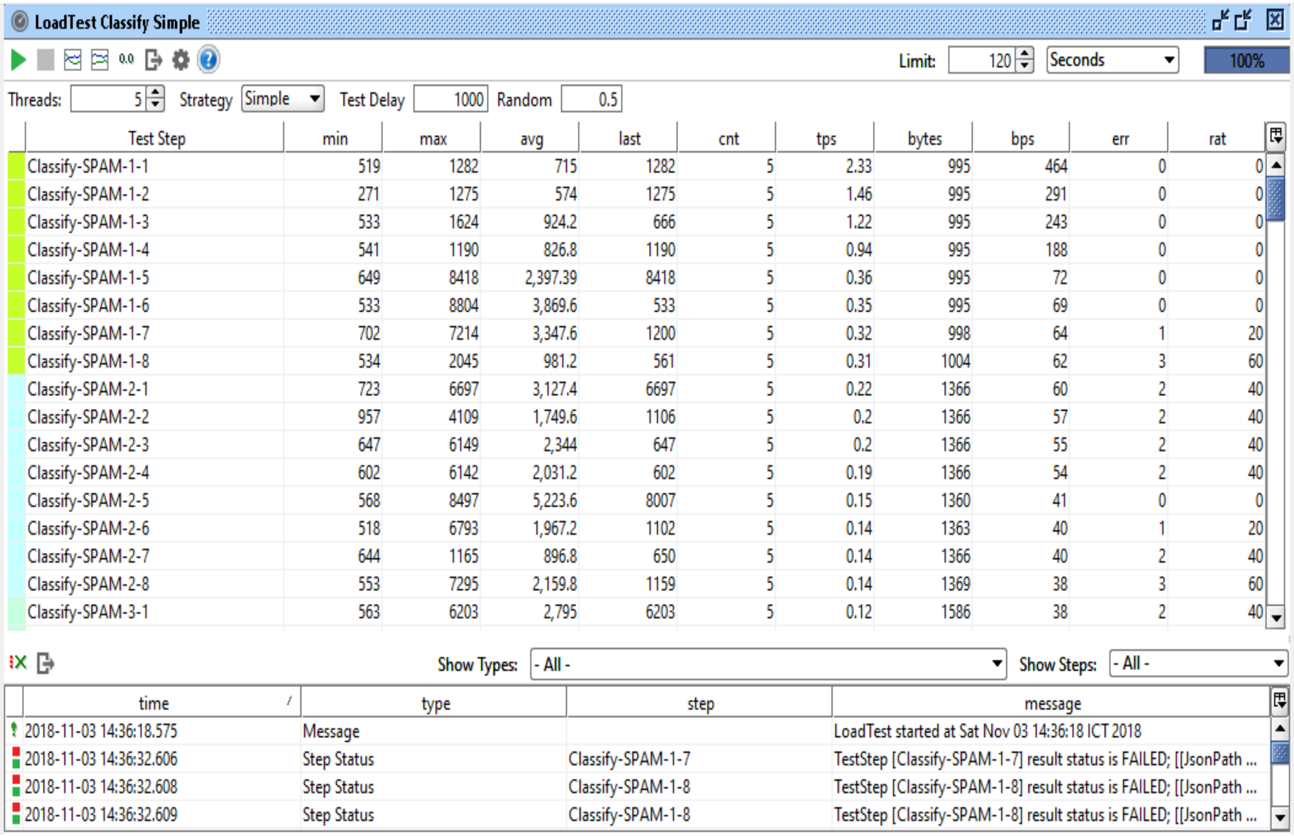
The second test is the Classify method to represents POST method, with the following parameters:

- Number of threads: 5
- Intervals: 60ms
- Variance: 0.5
- Limit: 120-180s
- Burst delay: 10s
- Burst duration: 10s

The result of the load testing is shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10.

In simple load test, it has minimal request about 0.5-6s, maximum request is 0.5-8ms, and average 0.5-6ms. In thread load test are, it has minimal request about 0.2-4s, maximum request is 0.2-3 ms, and average 0-0.7 ms. The system cannot continue all of the test data as it only finish 19 of 160 test data in 120s. In thread load test, it has minimal request about 0.4-11s, maximum request is 0.4-12ms, and average 0.4-12ms. The system cannot continue to load all the test data as it only finishes 87 of 160 test data in 120s. In variance load test, it has minimal request about 0.4-11s, maximum request is 0.4-12ms, and average 0.4-12ms. The system cannot continue all the test data as it only finished 87 of 160 test data in 120s.

The load testing results are summarized in Table XI and Table XII. The description for the Table are: I is minimum load time, X is maximum load time, and A is Average load time.



LoadTest Classify Simple

Limit: 120 Seconds 100%

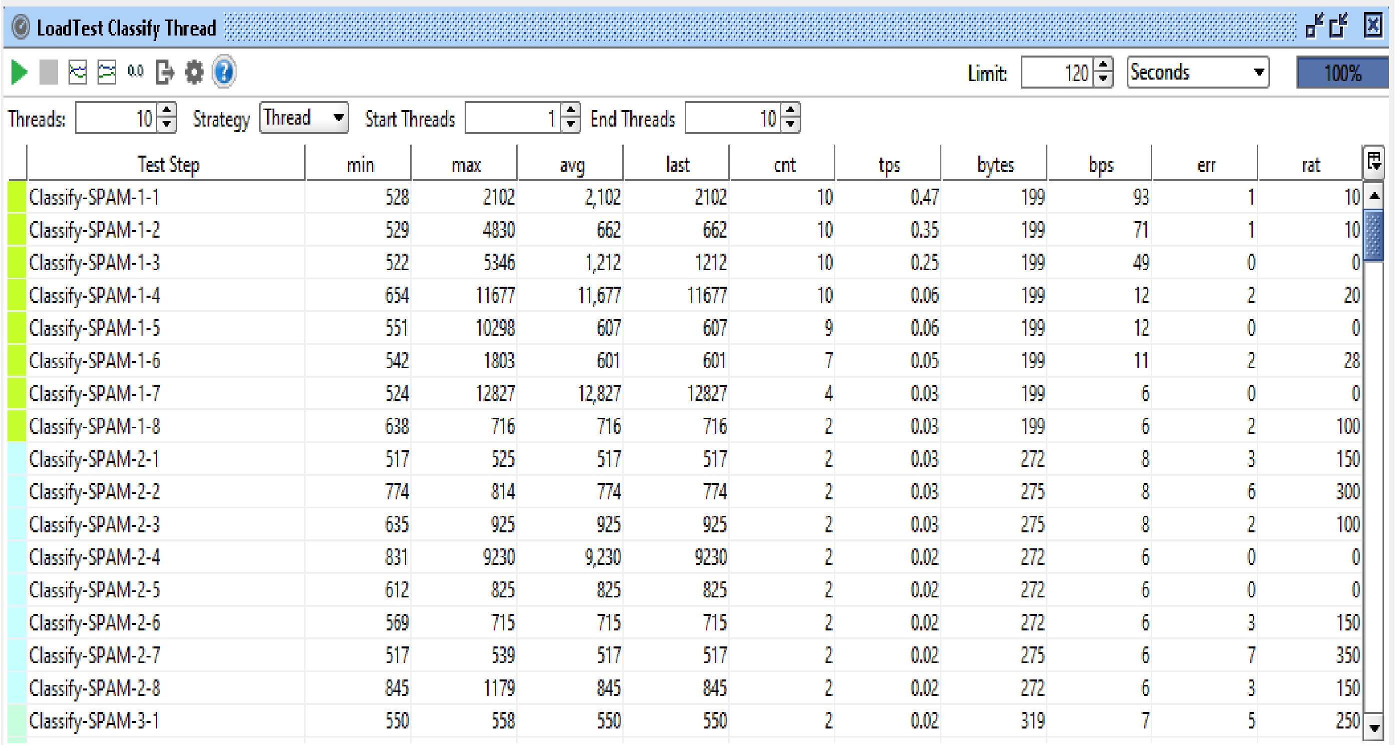
Threads: 5 Strategy Simple Test Delay 1000 Random 0.5

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
Classify-SPAM-1-1	519	1282	715	1282	5	2.33	995	464	0	0
Classify-SPAM-1-2	271	1275	574	1275	5	1.46	995	291	0	0
Classify-SPAM-1-3	533	1624	924.2	666	5	1.22	995	243	0	0
Classify-SPAM-1-4	541	1190	826.8	1190	5	0.94	995	188	0	0
Classify-SPAM-1-5	649	8418	2,397.39	8418	5	0.36	995	72	0	0
Classify-SPAM-1-6	533	8804	3,869.6	533	5	0.35	995	69	0	0
Classify-SPAM-1-7	702	7214	3,347.6	1200	5	0.32	998	64	1	20
Classify-SPAM-1-8	534	2045	981.2	561	5	0.31	1004	62	3	60
Classify-SPAM-2-1	723	6697	3,127.4	6697	5	0.22	1366	60	2	40
Classify-SPAM-2-2	957	4109	1,749.6	1106	5	0.2	1366	57	2	40
Classify-SPAM-2-3	647	6149	2,344	647	5	0.2	1366	55	2	40
Classify-SPAM-2-4	602	6142	2,031.2	602	5	0.19	1366	54	2	40
Classify-SPAM-2-5	568	8497	5,223.6	8007	5	0.15	1360	41	0	0
Classify-SPAM-2-6	518	6793	1,967.2	1102	5	0.14	1363	40	1	20
Classify-SPAM-2-7	644	1165	896.8	650	5	0.14	1366	40	2	40
Classify-SPAM-2-8	553	7295	2,159.8	1159	5	0.14	1369	38	3	60
Classify-SPAM-3-1	563	6203	2,795	6203	5	0.12	1586	38	2	40

Show Types: - All - Show Steps: - All -

time	type	step	message
2018-11-03 14:36:18.575	Message		LoadTest started at Sat Nov 03 14:36:18 ICT 2018
2018-11-03 14:36:32.606	Step Status	Classify-SPAM-1-7	TestStep [Classify-SPAM-1-7] result status is FAILED; [[JsonPath ...
2018-11-03 14:36:32.608	Step Status	Classify-SPAM-1-8	TestStep [Classify-SPAM-1-8] result status is FAILED; [[JsonPath ...
2018-11-03 14:36:32.609	Step Status	Classify-SPAM-1-8	TestStep [Classify-SPAM-1-8] result status is FAILED; [[JsonPath ...

Fig. 7. Simple Load Testing on Classify Method.



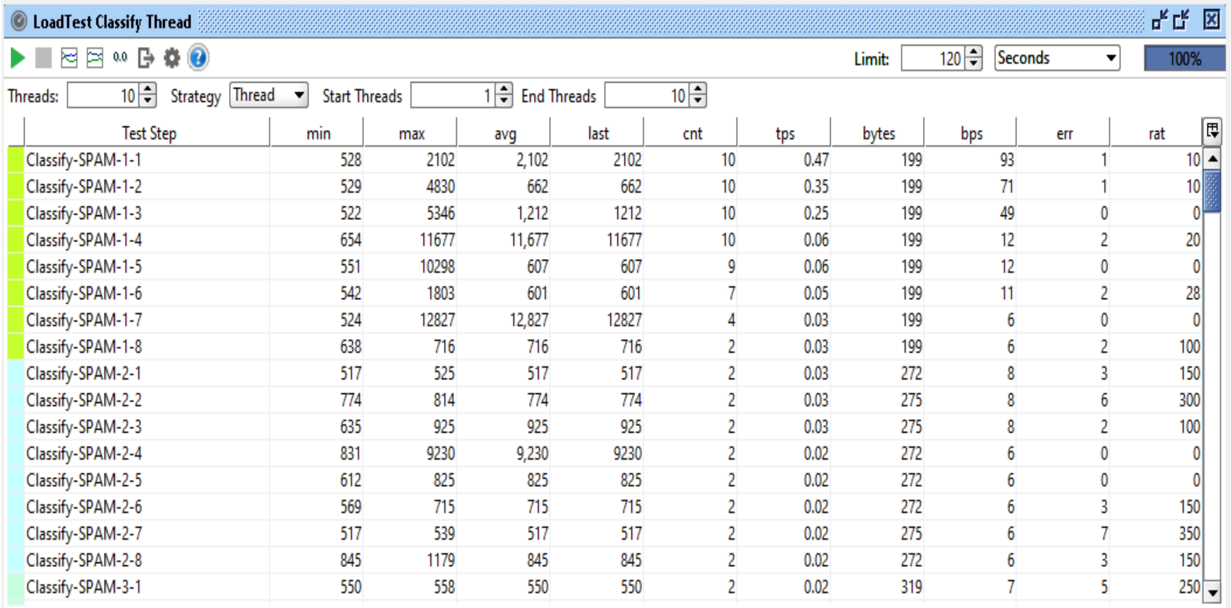
LoadTest Classify Thread

Limit: 120 Seconds 100%

Threads: 10 Strategy Thread Start Threads 1 End Threads 10

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
Classify-SPAM-1-1	528	2102	2,102	2102	10	0.47	199	93	1	10
Classify-SPAM-1-2	529	4830	662	662	10	0.35	199	71	1	10
Classify-SPAM-1-3	522	5346	1,212	1212	10	0.25	199	49	0	0
Classify-SPAM-1-4	654	11677	11,677	11677	10	0.06	199	12	2	20
Classify-SPAM-1-5	551	10298	607	607	9	0.06	199	12	0	0
Classify-SPAM-1-6	542	1803	601	601	7	0.05	199	11	2	28
Classify-SPAM-1-7	524	12827	12,827	12827	4	0.03	199	6	0	0
Classify-SPAM-1-8	638	716	716	716	2	0.03	199	6	2	100
Classify-SPAM-2-1	517	525	517	517	2	0.03	272	8	3	150
Classify-SPAM-2-2	774	814	774	774	2	0.03	275	8	6	300
Classify-SPAM-2-3	635	925	925	925	2	0.03	275	8	2	100
Classify-SPAM-2-4	831	9230	9,230	9230	2	0.02	272	6	0	0
Classify-SPAM-2-5	612	825	825	825	2	0.02	272	6	0	0
Classify-SPAM-2-6	569	715	715	715	2	0.02	272	6	3	150
Classify-SPAM-2-7	517	539	517	517	2	0.02	275	6	7	350
Classify-SPAM-2-8	845	1179	845	845	2	0.02	272	6	3	150
Classify-SPAM-3-1	550	558	550	550	2	0.02	319	7	5	250

Fig. 8. Burst Load Testing of Classify Method.

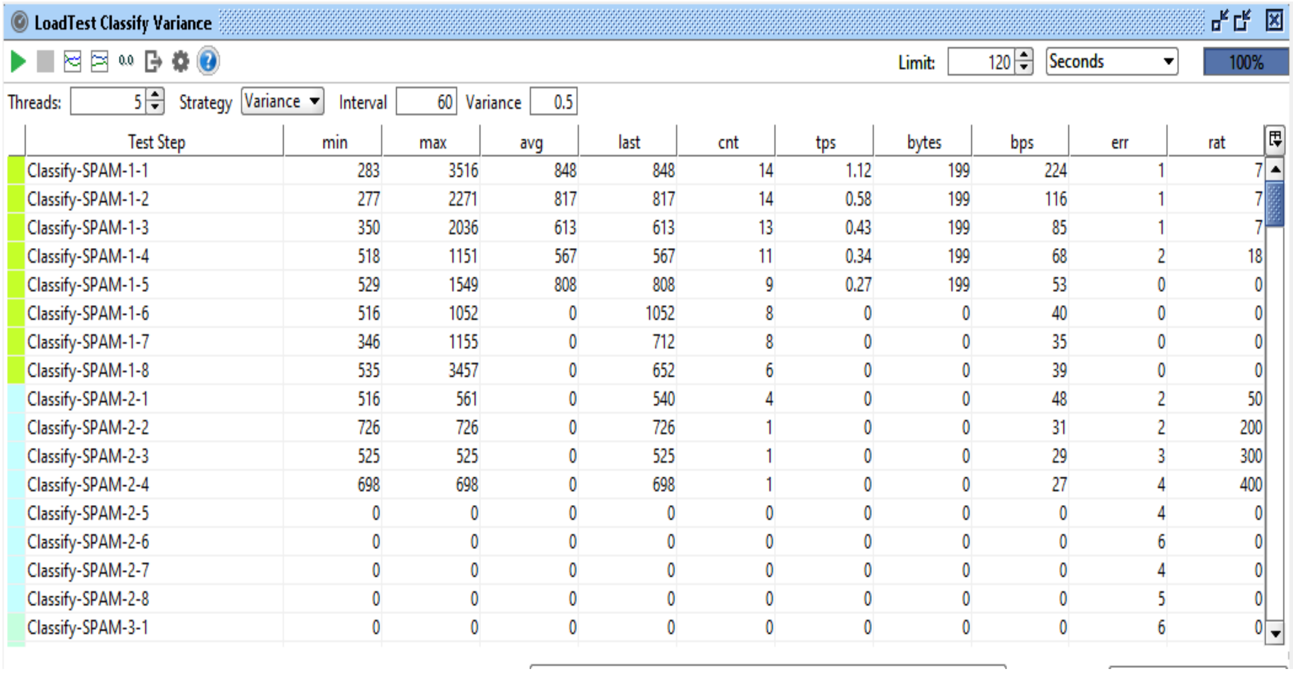


Limit: 120 Seconds 100%

Threads: 10 Strategy Thread Start Threads 1 End Threads 10

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
Classify-SPAM-1-1	528	2102	2,102	2102	10	0.47	199	93	1	10
Classify-SPAM-1-2	529	4830	662	662	10	0.35	199	71	1	10
Classify-SPAM-1-3	522	5346	1,212	1212	10	0.25	199	49	0	0
Classify-SPAM-1-4	654	11677	11,677	11677	10	0.06	199	12	2	20
Classify-SPAM-1-5	551	10298	607	607	9	0.06	199	12	0	0
Classify-SPAM-1-6	542	1803	601	601	7	0.05	199	11	2	28
Classify-SPAM-1-7	524	12827	12,827	12827	4	0.03	199	6	0	0
Classify-SPAM-1-8	638	716	716	716	2	0.03	199	6	2	100
Classify-SPAM-2-1	517	525	517	517	2	0.03	272	8	3	150
Classify-SPAM-2-2	774	814	774	774	2	0.03	275	8	6	300
Classify-SPAM-2-3	635	925	925	925	2	0.03	275	8	2	100
Classify-SPAM-2-4	831	9230	9,230	9230	2	0.02	272	6	0	0
Classify-SPAM-2-5	612	825	825	825	2	0.02	272	6	0	0
Classify-SPAM-2-6	569	715	715	715	2	0.02	272	6	3	150
Classify-SPAM-2-7	517	539	517	517	2	0.02	275	6	7	350
Classify-SPAM-2-8	845	1179	845	845	2	0.02	272	6	3	150
Classify-SPAM-3-1	550	558	550	550	2	0.02	319	7	5	250

Fig. 9. Thread Load Testing of Classify Method.



Limit: 120 Seconds 100%

Threads: 5 Strategy Variance Interval 60 Variance 0.5

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
Classify-SPAM-1-1	283	3516	848	848	14	1.12	199	224	1	7
Classify-SPAM-1-2	277	2271	817	817	14	0.58	199	116	1	7
Classify-SPAM-1-3	350	2036	613	613	13	0.43	199	85	1	7
Classify-SPAM-1-4	518	1151	567	567	11	0.34	199	68	2	18
Classify-SPAM-1-5	529	1549	808	808	9	0.27	199	53	0	0
Classify-SPAM-1-6	516	1052	0	1052	8	0	0	40	0	0
Classify-SPAM-1-7	346	1155	0	712	8	0	0	35	0	0
Classify-SPAM-1-8	535	3457	0	652	6	0	0	39	0	0
Classify-SPAM-2-1	516	561	0	540	4	0	0	48	2	50
Classify-SPAM-2-2	726	726	0	726	1	0	0	31	2	200
Classify-SPAM-2-3	525	525	0	525	1	0	0	29	3	300
Classify-SPAM-2-4	698	698	0	698	1	0	0	27	4	400
Classify-SPAM-2-5	0	0	0	0	0	0	0	0	4	0
Classify-SPAM-2-6	0	0	0	0	0	0	0	0	6	0
Classify-SPAM-2-7	0	0	0	0	0	0	0	0	4	0
Classify-SPAM-2-8	0	0	0	0	0	0	0	0	5	0
Classify-SPAM-3-1	0	0	0	0	0	0	0	0	6	0

Fig. 10. Variance Load Testing of Classify Method.

TABLE XI. SUMMARY OF SIMPLE AND BURST TESTING

Methods	**Average (in milliseconds)**					
	Simple			***Burst***		
	I	***X***	***A***	***I***	***X***	***A***
GET Version	252	3280	521	582	582	582
POST Classify	1255	7403	3355	404	1425	386

TABLE XII. SUMMARY OF THREAD AND VARIANCE TESTING

Methods	**Average (in milliseconds)**					
	Thread			***Variance***		
	I	***X***	***A***	***I***	***X***	***A***
GET Version	253	3277	503	251	671	494
POST Classify	913	1547	1319	484	1558	304

Version method is considerably faster than Classify method because it only returns static text, while Classify method is more slower because it does spam detection process. This characteristics is also shown in the simple and burst testing and thread and variance testing results. The Classify method performance is also affected by the length of the input and the size of the datasets used for spam detection process.

c) Data Driven Testing

Data driven test is using test data that has been stored in some external storage and use it iteratively. 8 datasets were used in which each dataset consists of 20 test data and divided into 2 more categories: 10 data categorized as SPAM and 10 data categorized as NON-SPAM so in total, it has 160 tests. The metrics measured were response time and accuracy. The results can be seen in Table XIII.

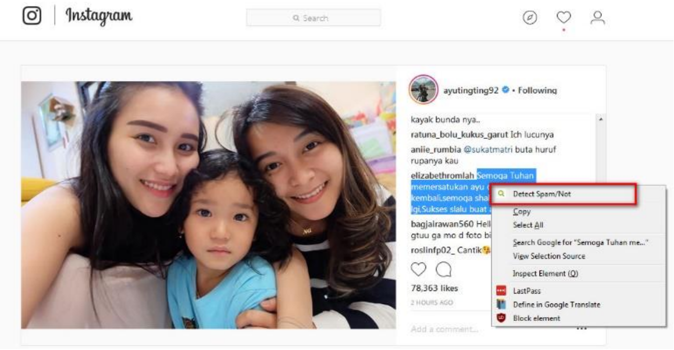
The result of this accuracy on data driven test with SOAPUI are: the accuracy is 63.125 % and average response time is about 2 seconds.

C. Browser Extension Development

The browser extension was developed extensively for Mozilla Firefox since it was using Greasemonkey plugin although it is also working in Google Chrome.

The extension is dynamically detecting the URL loaded in the address bar. If it is coming from Instagram's URL, it will add a new entry in the context menu (right click menu) as the user highlight some comment as shown in Fig. 11. When user clicked the entry, it will send the text to the Classify method in our web services and it will return the results ('spam' or 'not spam') in clear text and show it to user Fig. 12. In Google Chrome, the results are displayed as inFig. 13.

The browser extension developed is working as expected and able to do the spam detection process utilizing REST-based web service that were deployed in earlier work. The extension's user interface still need some improvements to make it easier to use for common user.



ayutingting92 • Following

kayak bunda nya..

ratuna_bolu_kukus_garut Ich lucunya

aniie_rumbia @sukatmatri buta huruf rupanya kau

elizabethromlah Semoga Tuhan memersatukan ayu kembali.semoga sha lgi.Sukses slalu buat

bagjalrawan560 Hell gtuu ga mo d foto bi

roslinfp02_ Cantik

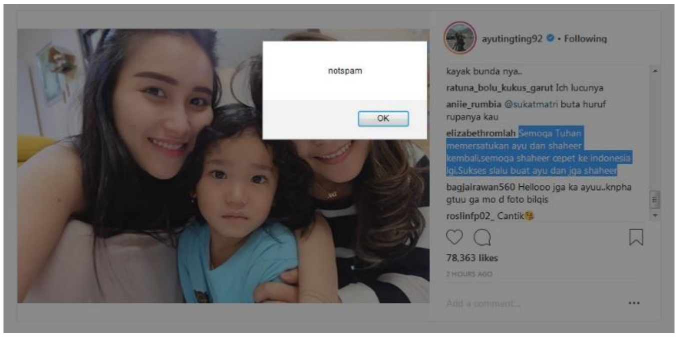
78,363 likes

2 HOURS AGO

Fig. 11. New Entry in Firefox's Context Menu.

TABLE XIII. DATA DRIVEN TESTING RESULT

STEP	TEST ID	RESULT	CATEGORY	TIME
Step 1	[Classify-SPAM-1-1]	OK	SPAM	1162 ms
Step 2	[Classify-SPAM-1-2]	OK	SPAM	521 ms
Step 3	[Classify-SPAM-1-3]	OK	SPAM	1638 ms
Step 4	[Classify-SPAM-1-4]	FAILED	NONSPAM	1543 ms
Step 5	[Classify-SPAM-1-5]	OK	SPAM	2208 ms
Step 6	[Classify-SPAM-1-6]	OK	SPAM	2271 ms
Step 7	[Classify-SPAM-1-7]	OK	SPAM	1554 ms
Step 8	[Classify-SPAM-1-8]	OK	SPAM	1921 ms
Step 9	[Classify-SPAM-2-1]	OK	SPAM	2468 ms
Step 10	[Classify-SPAM-2-2]	FAILED	NONSPAM	1125 ms
...	...	...	...	...
...	...	...	...	...
Step 150	[Classify-NOSPAM-9-6]	FAILED	SPAM	2211 ms
Step 151	[Classify-NOSPAM-9-7]	FAILED	SPAM	1512 ms
Step 152	[Classify-NOSPAM-9-8]	FAILED	SPAM	1883 ms
Step 153	[Classify-NOSPAM-10-1]	FAILED	SPAM	1964 ms
Step 154	[Classify-NOSPAM-10-2]	OK	NONSPAM	1123 ms
Step 155	[Classify-NOSPAM-10-3]	OK	NONSPAM	4039 ms
Step 156	[Classify-NOSPAM-10-4]	OK	NONSPAM	1660 ms
Step 157	[Classify-NOSPAM-10-5]	OK	NONSPAM	1988 ms
Step 158	[Classify-NOSPAM-10-6]	FAILED	SPAM	3511 ms
Step 159	[Classify-NOSPAM-10-7]	FAILED	SPAM	1863 ms
Step 160	[Classify-NOSPAM-10-8]	FAILED	SPAM	1797 ms
AVERAGE ACCURACY / TIME			**63.125%**	**1991,244 ms**



notspam

kayak bunda nya..

ratuna_bolu_kukus_garut Ich lucunya

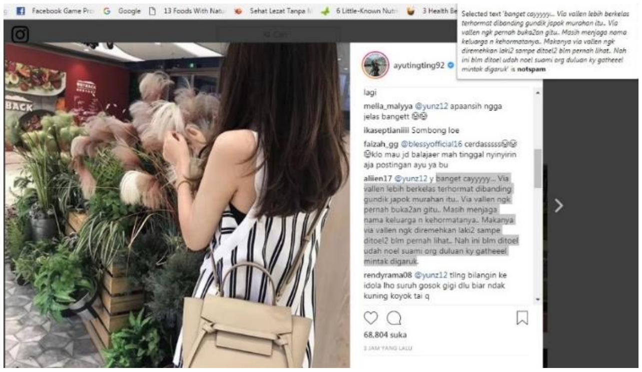
aniie_rumbia @sukatmatri buta huruf rupanya kau

elizabethromlah Semoga Tuhan memersatukan ayu dan shaheer kembali,semoga shaheer cepet ke indonesia lgi,Sukses slalu buat ayu dan jga shaheer

bagjairawan560 Hellooo jga ka ayuu..knpha gtuu ga mo d foto bilqis

roslinfp02_ Cantik😘

Fig. 12. Result of Classify Method in Mozilla Firefox.



Selected text 'banget cayyyyy... Via vallen lebih berkelas terhormat dibanding gundik japok murahan itu.. Via vallen ngk pernah buka2an gitu.. Masih menjaga nama keluarga n kehormatanya.. Makanya via vallen ngk diremehkan laki2 sampe ditoel2 blm pernah lihat.. Nah ini blm ditoel udah noel suami org duluan ky gatheeel mintak digaruk' is **notspam**

lagi

mella_malyya @yunz12 apaansih ngga jelas bangett 😂😂

ikaseptianiiii Sombong loe

faizah_gg @blessyofficial16 cerdasssss😂😂😂klo mau jd balajaer mah tinggal nyinyirin aja postingan ayu ya bu

aliien17 @yunz12 y banget cayyyyy... Via vallen lebih berkelas terhormat dibanding gundik japok murahan itu.. Via vallen ngk pernah buka2an gitu.. Masih menjaga nama keluarga n kehormatanya.. Makanya via vallen ngk diremehkan laki2 sampe ditoel2 blm pernah lihat.. Nah ini blm ditoel udah noel suami org duluan ky gatheeel mintak digaruk.

rendyrama08 @yunz12 tling bilangin ke idola lho suruh gosok gigi dlu biar ndak kuning koyok tai q

Fig. 13. Result of Classify Method in Google Chrome.

V. Conclusions

In this paper, a browser extension for Firefox & Chrome has been successfully developed and integrated into a REST-based web service [19] deployed on top of AWS Platform. Accuracy of the web service were measured using three datasets (whole datasets, 1000 stemmed dataset and 1000 non-stemmed dataset) and achieved accuracy level of 63.125% for whole datasets, 72% for non-stemmed dataset, and 70% for stemmed dataset. The average response time is under 2s, minimum load time test is between 0.2 – 1.2s, and, maximum load time test is between 3 – 7s. Although the browser extension is working as expected, the user interface and data accuracy still have room for improvements.

Acknowledgment

This research is partially funded by UKDW Research and Community Service (LPPM) unit under grant no. 072/D.01/LPPM/2018 and Faculty of Information Technology.

References

[1] L. Vries, S. Gensler and P. S. Leeflang, "Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing," Journal of Interactive Marketing, vol. 26, no. 2, pp. 83-91, 2012.

[2] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," Tourism Management, vol. 31, no. 2, pp. 179-188, 2010.

[3] W.-y. S. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser and B. W. Hesse, "Social Media Use in the United States: Implications for Health Communication," Journal of Medical Internet Research, vol. 11, no. 4, 2009.

[4] M. Chakraborty, S. Pal, R. Pramanik and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey," Information Processing & Management, vol. 52, no. 6, pp. 1053-1073, 2016.

[5] M. Salehi, S. Shehnepoor, R. Farahbakhsh and N. Crespi, "NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media.," in IEEE Transactions on Information Forensics and Security, 2017.

[6] W. Zhang and H. M. Sun, "Instagram spam detection," in 22nd IEEE Pacific Rim International Symposium on Dependable Computing, Christchurch, New Zealand, 2017.

[7] A. R. Chrismanto and Y. Lukito, "Klasifikasi Komentar Spam Pada Instagram Berbahasa Indonesia," in Seminar Nasional Teknologi Informasi Kesehatan (SNATIK), Yogyakarta, 2017.

[8] F. Fathaliani and M. Bouguessa, "A Model-Based Approach for Identifying spammers in social networks," in IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 2015.

[9] N. Agarwal and Y. Yiliyasi, "Information quality challenges in social media," in The 15th International Conference on Information Quality, Little Rock, Arkansas, USA, 2010.

[10] J. Constine, "TechCrunch," 20 June 2018. [Online]. Available: https://techcrunch.com/2018/06/20/instagram-1-billion-users/. [Accessed 1 November 2018].

[11] R. Hardinata and J. Tirtawangsa, "A game with purpose to filter spams from Indonesian Twitter trending topics," in 2016 4th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2016.

[12] A. A. Septiandri and O. Wibisono, "Detecting spam comments on Indonesia's Instagram posts," Journal of Physics: Conference Series, vol. 801, no. 1, 2017.

[13] A. Barth, "The Web Origin Concept," Infosec Institute, December 2001. [Online]. Available: https://tools.ietf.org/html/rfc6454. [Accessed 1 November 2018].

[14] W. S. Raharjo and A. Ashari, "IMPLEMENTASI ANNOTEA CLIENT BERBASIS WEB UNTUK MENGATASI ATURAN SAME ORIGIN POLICY," in KNASTIK, Yogyakarta, Indonesia, 2009.

[15] Z. Wuxain and S. Hung-Min, "Instagram Spam Detection," in IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC), Christchurch, New Zealand, 2017.

[16] A. Rachmat and Y. Lukito, "Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes," Ultimatics, vol. 9, no. 1, 2017.

[17] A. R. Chrismanto and Y. Lukito, "KLASIFIKASI KOMENTAR SPAM PADA INSTAGRAM BERBAHASA INDONESIA MENGGUNAKAN K-NN," in Seminar Nasional Teknologi Informasi Kesehatan (SNATIK) 2017, Yogyakarta, 2017.

[18] A. R. Chrismanto and Y. Lukito, "Identifikasi Komentar Spam Pada Instagram," Lontar Komputer, vol. 8, no. 3, pp. 219-231, 2017.

[19] A. R. Chrismanto, W. S. Raharjo and Y. Lukito, "Design and Development of REST-based Instagram Spam Detector for Indonesian Language," in 3rd International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2018.

[20] R. Fielding, Architectural Styles and the Design of Network-based Software, California: University of California, 2000.

[21] S. Malik and D.-H. Kim, "A comparison of RESTful vs. SOAP web services in actuator networks," in 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), Milan, Italy, 2017.

[22] S. Kumari and S. K. Rath, "Performance comparison of SOAP and REST based Web Services for Enterprise Application Integration," in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 2015.

Spectral Efficiency of Massive MIMO Communication Systems with Zero Forcing and Maximum Ratio Beamforming

Asif Ali[1], Imran Ali Qureshi[2], Abdul Latif Memon[3], Sajjad Ali Memon[4], Erum Saba[5]
[1,2,3,4] Department of Telecommunication Engineering, [5] Information Technology Center
[1,2,3,4] Mehran University of Engineering and Technology, [5] Sindh Agriculture University
[1,2,3,4] Jamshoro, Pakistan, [5] Tandojam, Pakistan

***Abstract*—The massive multiple-input-multiple-output (MIMO) is a key enabling technology for the 5G cellular communication systems. In massive MIMO (M-MIMO) systems few hundred numbers of antennas are deployed at each base station (BS) to serve a relatively small number of single-antenna terminals with multiuser, providing higher data rate and lower latency. In this paper, an M-MIMO communication system with a large number of BS antennas with zero-forcing beamforming is proposed for the improved spectral efficiency performance of the system. The zero forcing beamforming technique is used to overcome the interference that limits the spectral efficiency of M-MIMO communication systems. The simulation results authenticate the improvement in the spectral efficiency of M-MIMO system. The spectral efficiency value using zero-forcing beamforming is near to the spectral efficiency value with the no-interference scenario.**

Keywords—Massive MIMO; Base station; channel capacity; Spectral efficiency; latency; cellular communication; beamforming techniques; throughput; mobile communication

I. Introduction

In a few years, several studies have been focused on massive MIMO systems which play an important role in emerging trends in communication networks. The massive MIMO transceiver contains the various number of smart antenna arrays, which enable to get higher spectral efficiency and vigorously efficient to be achieved [1][2]. In various antennas were attached to an array form to organize the base station and mobile stations of the wireless communication link, in order to manage the signal in appropriate direction to improve the system performance [3]. The multi-beam forming needs to join the multiple signals received or transmitted by an array of antennas. Designing and implementation challenges are associated with its configuration and applications. The massive MIMO antennas are integrated with beamforming array antenna technologies for next-generation cellular communication and deployed in 2020 as recommended to [4]. Various numbers of communication model and antennas are unable to increase the capacity of wireless communication systems to mitigate the multipath fading and channel interference [5]. The beamforming technique is improving signal radiation accordingly to the environment. In communication networks, transmitter and receiver used the beamforming signal transmission from base station to mobile station with multiple antennas [6]. The key theme of enabling the beamforming in the communication networks, to enhance the power capacity and minimize the interference [7]. To transmit the same signal with various values of phase and amplitude, it can pass from the different MIMO channel was constructively added the desired signal and destructively on the other users. In MIMO array gain would contest the increased path losses but necessary to provide the suitable link budget [8]. The millimeter wave communication creates the massive MIMO (multi-input multi-output) system more smarts for the size of the antenna is small but makes a possible a huge number of array antennas in a small region at the BS and massive MIMO can conflict the high path loss and fading of mm-wave channels [9]. The massive MIMO principle is previously working in the Wi-Fi and 4G standards, it will play a major role when the 5G networks are coming. Certainly, the massive MIMO is extensively estimated key enable technology and fundamental component of 5G [10][11].

II. Related Work

In the related work of massive MIMO mostly consider the omnidirectional antennas for the base station. It is eminent to know that base station directional antennas are used along with sectorization antennas [12] to increase SINR of the cellular network. In massive MIMO multiples antennas are used, it poses the major challenges in it [13], a number of antennas are to increase the energy consumption and cost as well. In millimeter wave (mm-Wave) bands have high path loss at their operating frequencies which are essential for the high gain and sufficient to signal-to-noise ratio (SNR) at a very low distance [14]. In the low range wavelengths, of the mm-wave frequencies where the large antennas array is packed with small form factor [8], the higher gain array can be a conflict to increase the path loss and crucial to provide the suitable link budget [15]. Therefore, massive MIMO is a preferred technology to overcome these problems according to their desirable achievements.

III. System Model

We have considered the uplink and downlink of one cell multi-user massive MIMO system consists of K number of single antenna users and BS prepared with M number of antennas. In our consideration, k is the all active users share the resources at the same time with perfect channel interference with base station end user [16]. Considering the multi-carrier frequency of flat channel and with symbol rates of sampling, the Mx1 signals received at the BS from the K users. In general, the propagation channels are modeled on large scale fading and small-scale fading, the zero mean and unit variances are estimated in this work. The postulation of the symbol rate of sampling is the matched at the receiver end; it must be executed in the analog sphere. The better presentation might be achieved by oversampling the ADC especially those of one-bit resolution [17]. The fading block representation through coherence bandwidth Wc and coherence time Tc. In this model both channels are constant for an interval of length $T = TcWc$, those symbols were independently changed from various intervals. Note that T is a fixed for a minimum duration of all the users.

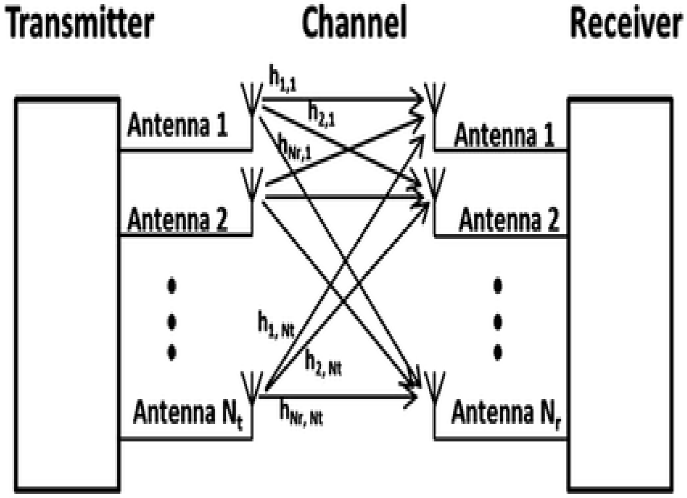


Fig. 1. Transmitter and Receiver System Model.

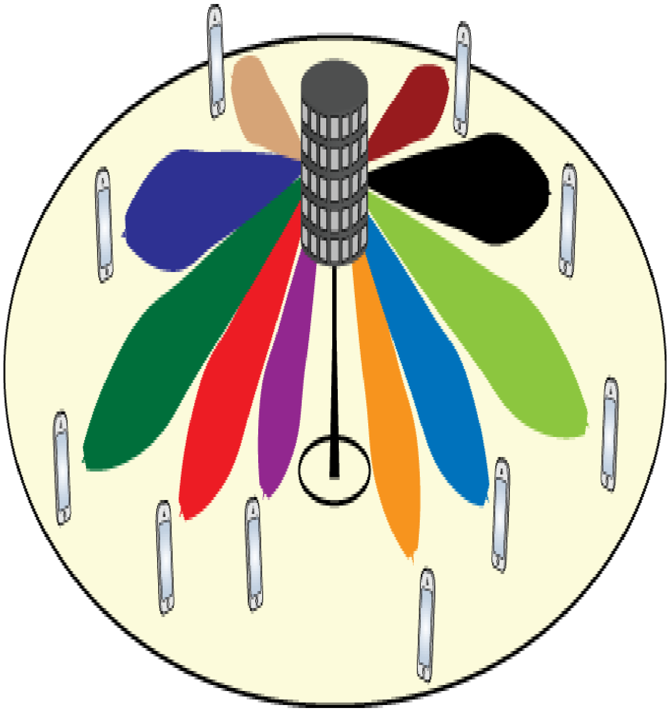


Fig. 2. The Uplink and Downlink Transmission in a Massive MIMO System.

IV. Massive MIMO System

Massive MIMO is the primary component of the fast 5G networks in the future. Because the MIMO wireless networks allow the transmitting and receiving a large number of the data signal at the same time with the same radio channels. To fulfill the need of high data rates and good quality of service constraint, massive multiple-input-multiple-output (MIMO) systems will be preferred as they contain a la number of antennas at the access point, which is a suitable technology for 5G communication [18]. The massive MIMO technology is familiarized with the last two decades and useful to many wireless standards due to drastically improve the capability and reliability of wireless systems. The number of antennas is increased to massive MIMO to enhance the performance likewise: spectrum efficiency, energy efficiency and network coverage [19]. The massive MIMO systems are widely adopted in cellular communication as a vital component in the future. The array elements of massive MIMO systems are formulated to the shape of RF energy and reusing the spectrum between separated users. In this phenomenon, the energy falls into places outside the intended users, causing the unwanted interference with cell boundary and wasting the spectral power at the transmitter end. The MIMO technology utilizes entire bandwidth in multi-antenna BS spatially multiplex a large number of user workstation is compatible to communicate. The underlying principle of the Massive MIMO theory with its communication protocol is enlightened from a chronological point of view on [20] presents a fundamental theoretic performance analysis. The multi-cell simulate the ones were presented to showcase the massive MIMO ability may offer 100-fold improvements in spectral efficiency over existing tools, lacking for highly developed signal processing.

The massive MIMO system consists of base stations with M antennas that provides K single antenna terminal as shown in Fig.1 for graphic footprints of massive MIMO. The base station multiplexes the received the data stream per user for uplink and downlink. The base station uses the antennas to direct the signal in the direction of the preferred receiver in the downlink and to divide the multiple signals for the uplink transmission as given in Fig. 2. To mitigate interference and enhance the SNR just enabling extra antennas rather than sending multiple data streams [21]. The main advantage of the MIMO system over the existing network is that it can multiply the capability of wireless connectivity without need the additional spectrums. It is a considerable point that competence enhancement and it could potentially give up as a 100-fold in the future.

V. Problems and Issues

In massive MIMO systems, the major problem is interference while the antennas originate the same signal. To overcome this problem, 3D beamforming massive MIMO is a preferable solution over 2x2 Massive MIMO systems. In a massive MIMO system, pilot contamination is independent of the pilot sequence so it limits the reuse for uplink. The effect of reuse produce the conflict within different cells in the antenna array of the BS has correlated with the desired received pilot signal. The array antenna at the BS obtains a channel approximate that is corrupted by a combination of signals from other terminals using the same pilot sequence and it will create the interference. The hybrid beamforming is well-matched with mm-wave bands for massive MIMO systems. In comparison the lower frequency band MIMO systems which reduce the complexity and performance gap between beamforming techniques. The mm-wave and massive MIMO applications are limited due to LOS/near LOS properly. To mitigate the limitation of interference through superposition and beamforming techniques. The studies have focused on multiuser transmission schemes for massive MIMO systems. It is necessary to find the mm-wave channel propagation characteristics due to high path loss. The beamforming techniques and mm-wave system provide a solution to improve system performance. The source emits a signal and received at BS with estimated beamforming value for localization. The source is founded through triangulation and the dense multipath environment in the urban region. The localization based beamforming techniques may further be developed to get more benefits from it.

VI. Antenna Placement

One of the major problems with a MIMO system is to replace the antennas. For many systems using actually undersized units, the antenna deployment offers some problem. To accurately run the MIMO system, the contact between antennas should be smaller. According to the thumb rule, λ/2 spacing (where λ is the signal wavelength) is measured essential to offer nearly no connection among the antennas. There are different methods to get taken. The optimal methods are commonly used in conservative small scale MIMO systems because the incredible number of antenna subsets is used in massive MIMO system which is necessary for antenna placement [22].

A. *Use High Frequencies*

To accommodate a large number of antennas are necessary for large MIMO systems, using the basic data onto frequency and wavelength as, high frequencies shorter the signal wavelength, therefore permit the dimension related antenna will be adjusted within the substantial gap. Many systems are taking into consideration the use of 10 GHz frequency, to the extent that 60GHz and more [23].

B. *Use Volumetric Instead of Linear Spacing*

It is feasible to use the three dimensions inside a range of three dimension spacing as a linear fashion within two dimensions. Although there are many things that include mobile phones, are often slim and for that reason, this move toward cannot be appropriate, in some cases, a cube may be capable to adjust the maximum antennas with three dimensional spaced. The signal model and regard the effect of cluster size or spacing, the massive MIMO system contains the transmitter and receiver antennas [24].

C. *Use of Spatial Modulation*

The number of RF chains is necessary for a massive MIMO system can be condensed without understanding the spectral efficiency by spatial modulation. Spatial modulation is used to transmit a single chain for multiple antennas. It transmits the data and selects a single antenna from an array antenna for transmission at the same time. It has adopted an easy but effective coding mechanism, in which the bits of the transfer information and the local position of the transmitter antenna in the general antenna array. If the spreader signal reaches your destination at the receiving antenna is very low and the spacing antenna elements are inadequate the Massive MIMO capacity could be ruined. Therefore vigilant deliberation is needed for the antenna design and deployments [25].

VII. Simulation Results

In this section, the simulation results of the massive MIMO system for a cellular communication system with simulation parameters mentioned in Table I. To examine the numerical results where the number of users k=10 simultaneously provide the BS by M antennas. For the ease of all user is unspecified to have typical SNR of 5dB, there is an ideal channel state interference accessible everywhere. Fig. 3 illustrate that typical spectral efficiency like a purpose of M antennas, attain as a result of sum capability of non-linear processing and basic processing schemes known as the zero-forcing (ZF), which crack to repress all obstruction. The uplink and downlink transmissions show in the form of results.

This simulation presents that the non-linear processing largely outperforms linear ZF when $M \approx K$. The working point $M = K$ creates exacting logic from a multiplexing perception for the multiplexing gain min($M;K$) does not develop, let M raise for a fixed K. Fig. 3 demonstrate that there is some other cause to regard as $M > K$; the capability boosts and the performance will be linear ZF processing advance the capacity. Previously at $M = 20$, there is only a little difference between linear ZF and optimum non-linear processing. Actuality, these two methods also approached the upper curve in Fig. 3, which characterizes the upper bounce which indicates where the hindrance among the users is ignoring. It illustrates the fundamentally that only K users can be served as every one of them was single handed in the cell.

TABLE I. Simulation Parameters

Simulation Parameters	Values
Number of antennas(M)	10 -100
Number of users(K)	10
SNR	5dB
Channel variance	Ones(1,K)
Coherence time	400symbol

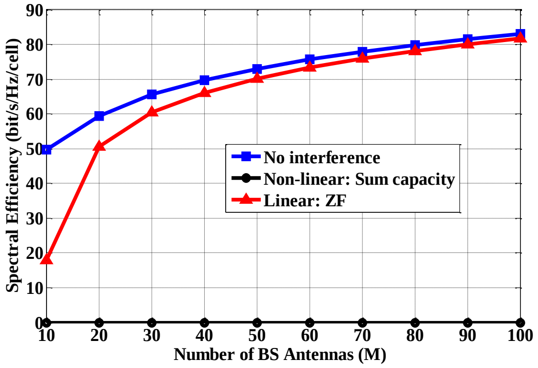


Fig. 3. The Average Spectral Efficiency of Massive MIMO for Both Uplink and Downlink.

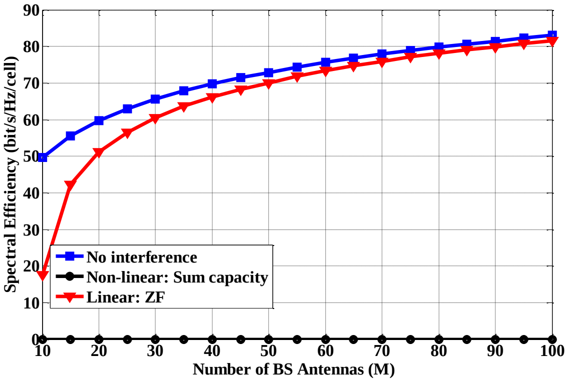
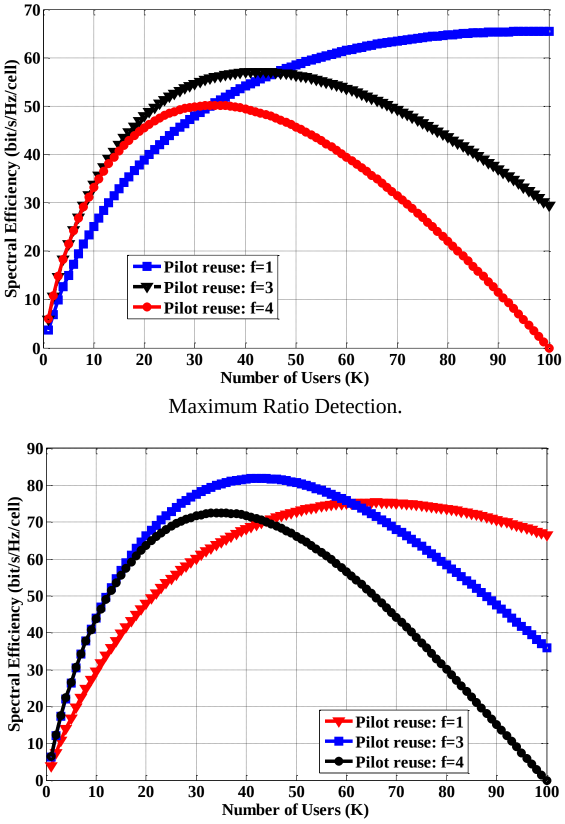


Fig. 4. An Improved Version of Spectral Efficiency for Uplink and Downlink Transmission.



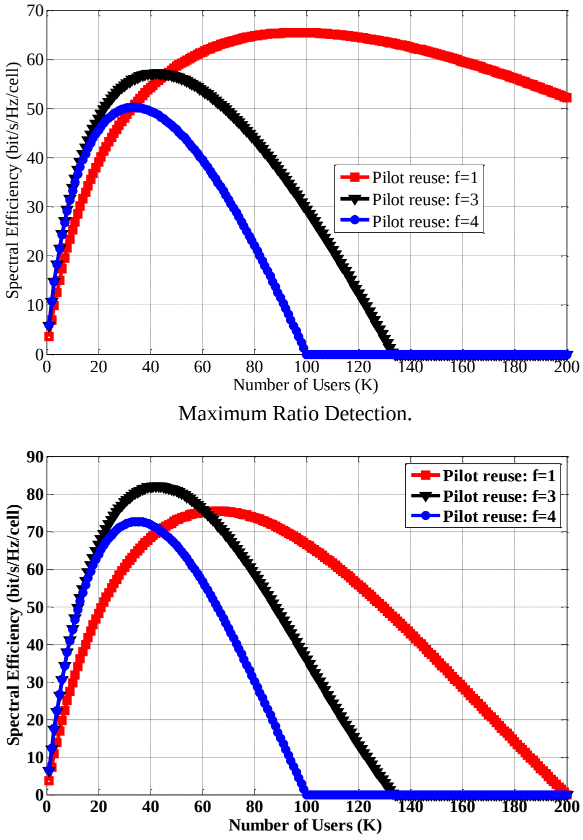
Maximum Ratio Detection.

Zero Forcing Detection.

Fig. 5. Average Spectral Efficiency with different Pilot Reuse Factor at Two Different SNR Levels, 0dB, and 20dB.

In this scenario, the f=1 shows the universal pilot reuse and f is greater than one known as the non-universal pilot reuse. In every tier has six cells of the hexagonal cell topology. The minimum reuse pilot feature improves the symmetric pilot reuse prototype are f=1, f=3, and f=4. These reuse patterns are shown in Fig. 4, where different color shows different reuse pilot sequences. The same color of cells and same division of pilots, reason the pilot contagion to everywhere, while the other hand the different colors have no pilot contamination. We have noted the pilot reuse factor of f=4, can be divided into four cells with different disjoint groups, or divide into subcells. Later on, this is recognized as the fractional pilot reuse and it has less common pilot reuse at the corners of the cell centers [26], due to its users at the cell borders are most susceptible to pilot contamination. In this scenario, massive MIMO contain M=200 base station antennas and coherence time of *tc*= 400symbols. The users are unspecified to be regularly circulated in the cell and the channel is a representation as uncorrelated Rayleigh fading with non-line of sight attenuation with path loss exponent 3.7 [27].

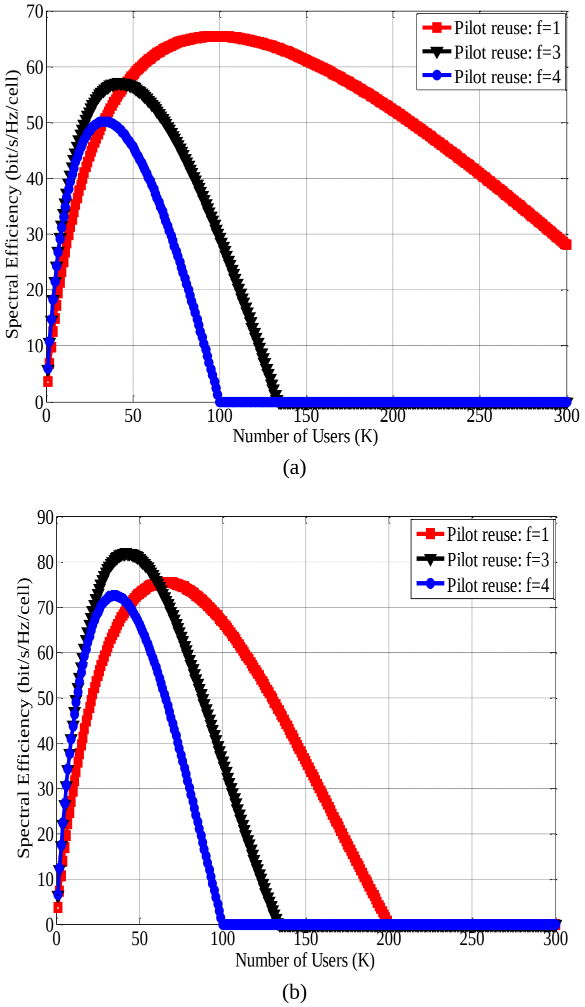
The spectral efficiency of different users of both ZF and MR detection is shown in Fig. 5. In this inspection, the two SNR levels, 0dB, and 20dB give the same performance. It noted that massive MIMO works similarly at both SNR levels whether its high or low, therefore array gain make it the interference limited and rather than noise limited. As the value of f improved properly it will operate top on the curve and massive MIMO can give the high spectral efficiency over a large number of users. When the number of users is raised like K>10 as suggested in Fig. 6, it removes the complications from the massive MIMO networks because the active user serve the simultaneously in every coherence time interval of data, which number is typically more than hundreds and it shared between all the users. If the number of users increased beyond 200 then the spectral efficiency also decreases after a certain point, which is demonstrated in Fig. 7 that is an improved version of spectral efficiency.



Maximum Ratio Detection.

Zero Forcing Detection.

Fig. 6. Improved Spectral Efficiency at two Different Levels of SNR, 0dB and 20dB at K=200 users.



(a)

(b)

Fig. 7. Improved Spectral Efficiency and Pilot Reuse Factor at different Levels of SNR, 0dB and 20dB at K=300 users.

The M-MIMO achieved a high sum of spectral efficiency which distributes among all users. The difference between MR and ZF in terms of gain is relatively small, the ZF provides the gain performance ratio between the 4% and 50% but depends on the user as well. In the single cell simulation where the zero-forcing offered more than twofold the spectral efficiency as compared to the MR. The cause for the unassuming performance space is suffered from ZF from the obstruction in the multicell given that the pilot contamination and countless inter-cell interferes very much possible to cancel all interferences. The pilot reuse factors are a very important parameter to design for the M-MIMO system and its good selection which depends on the user capability and propagation environment in which more number of antennas are used at the base station. The achieved spectral efficiency is 80 (bit/s/Hz/cell) and 82 (bit/s/Hz/cell) with 100 users using non-linear-sum capacity and no interference technique respectively, while maintaining same bandwidth of base stations as in current system. The pilot reuse is a limitation of spectral efficiency when the value of pilot reuse (f=1) the efficiency is sustainable to provide better communication otherwise the efficiency decreases.

VIII. Conclusion

The massive MIMO technique introduces more efficiency in the present scenario of wireless communication systems. The single antenna was converted into arrays of antennas for advanced beamforming techniques; the major advantages are that it can be improving the SNR of the overall system along with data transmission within the cell during the time of interval. In the massive MIMO, it is necessary for linear detection to perform the zero-forcing to allow the reliable detection to accumulate high-quality performance through enabling an increase in the dimension of arrays which mitigate the processing problem and considered as a tolerable area. Massive MIMO sustainably increased the spectral efficiency and enhance capacity and coverage. A base station uses a large number of antennas in a massive MIMO system, therefore it achieves high throughput and more capacity. The small cell system uses the low power mini BS to avoid interference easily and enhance efficiency. Nowadays, it is used in smart cities and densely populated areas and not only suitable for improving the spectral efficiency but also higher throughput can be achieved in 5G technology.

References

[1] H. Q. Ngo, S. Member, E. G. Larsson, S. Member, and T. L. Marzetta, “Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems,” vol. 61, no. 4, pp. 1436–1449, 2013.

[2] Y. Huang, S. Member, S. He, J. Wang, and S. Member, “Spectral and Energy Efficiency Tradeoff for Massive MIMO,” IEEE Trans. Veh. Technol., vol. 67, no. 8, pp. 6991–7002, 2018.

[3] R. Choudhury, “A Network Overview of Massive MIMO for 5G Wireless Cellular: System Model and Potentials,” Int. J. Eng. Res. Gen. Sci., vol. 2, no. 4, pp. 338–347, 2014.

[4] J. Huang, C. Wang, R. Feng, and J. Sun, “Multi-Frequency mmWave Massive MIMO Channel Measurements and Characterization for 5G Wireless Communication Systems,” vol. 35, no. 7, pp. 1591–1605, 2017.

[5] L. I. Wang, J. Li, S. Member, J. Zhang, and S. Member, “Uplink Sum Rate Analysis of Massive Distributed MIMO Systems Over Composite Fading Channels,” IEEE Access, vol. 6, pp. 25970–25978, 2018.

[6] E. Ali, M. Ismail, R. Nordin, and N. F. Abdulah, “Beamforming techniques for massive MIMO systems in 5G : overview, classification, and trends for future research,” Front. Inf. Technol. Electron. Eng., vol. 18, no. 6, pp. 753–772, 2017.

[7] E. Björnson, J. Hoydis, and M. Kountouris, “Massive MIMO Systems With Non-Ideal Hardware : Energy Efficiency,Estimation, and Capacity Limits,” vol. 60, no. 11, pp. 7112–7139, 2014.

[8] J. Jing, C. Xiaoxue, and X. Yongbin, “Energy-efficiency based downlink multi-user hybrid beamforming for millimeter wave massive MIMO system,” J. China Univ. Posts Telecommun., vol. 23, no. 4, pp. 53–62, 2016.

[9] O. El Ayach and R. W. Heath, “Multimode Precoding in Millimeter-Wave MIMO Transmitters with Multiple Antenna Sub-Arrays,” pp. 3476–3480, 2013.

[10] Sajjad Ali, Z. Chen and F. Yin "Eradication of pilot contamination and zero forcing precoding in the multi-cell TDD massive MIMO systems" IET Communications vol. 11 no.13, pp. 2027-2034, 2017.

[11] Sajjad Ali, Zhe Chen, and Fuliang Yin "Pilot decontamination in TDD multi-cell massive MIMO systems with infinite number of BS antennas" Canadian Journal of Electrical and Computer Engineering (IEEE Canada), vol. 40, no. 3, pp. 171-180, Summer 2017.

[12] S. Shahsavari, P. Hassanzadeh, A. Ashikhmin, and E. Erkip, “Sectoring in multi-cell massive MIMO systems,” Conf. Rec. 51st Asilomar Conf. Signals, Syst. Comput. ACSSC 2017, vol. 2017–October, pp. 1050–1055, 2018.

[13] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, “Hybrid Beamforming for Massive MIMO: A Survey,” IEEE Commun. Mag., vol. 55, no. 9, pp. 134–141, 2017.

[14] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Commun. Mag., vol. 52, no. 2, pp. 186–195, 2014.

[15] S. Rajoria, A. Trivedi, and W. W. Godfrey, “A comprehensive survey: Small cell meets massive MIMO,” Phys. Commun., vol. 26, pp. 40–49, 2018.

[16] H. Wang, D. Yang, X. Li, and P. Pan, “How Many Signals Can Be Sent in a Multi-Cell Massive MIMO System,” IEEE Wireless. Commun. Lett., vol. 7, no. 3, pp. 368–371, 2018.

[17] H. Pirzadeh, S. Member, and A. L. Swindlehurst, “Spectral Efficiency of Mixed-ADC Massive MIMO,” no. c, pp. 1–15, 2018.

[18] X. Wei et al., “Uplink Channel Estimation in Massive MIMO Systems Using Factor Analysis,” vol. 7798, no. c, pp. 1–4, 2018.

[19] L. Lu, S. Member, G. Y. Li, and A. L. Swindlehurst, “An Overview of Massive MIMO : Benefits and Challenges,” vol. 8, no. 5, pp. 742–758, 2014.

[20] T. Van Chien, S. Member, E. Björnson, S. Member, and E. G. Larsson, “Joint Pilot Design and Uplink Power Allocation in Multi-Cell Massive MIMO Systems,” vol. 17, no. 3, pp. 2000–2015, 2018.

[21] E. Björnson, M. Kountouris, M. Bengtsson, and S. Member, “Receive Combining vs. Multi-Stream Multiplexing in Downlink Systems With Multi-Antenna Users,” vol. 61, no. 13, pp. 3431–3446, 2013.

[22] C. Bounds, “Massive MIMO Antenna Selection : Switching,” vol. 66, no. 5, pp. 1346–1360, 2018.

[23] L. Zhao, K. Li, K. Zheng, S. Member, and M. O. Ahmad, “An Analysis of the Tradeoff Between the Energy and Spectrum Efficiencies in an Uplink Massive MIMO-OFDM System,” vol. 62, no. 3, pp. 291–295, 2015.

[24] Y. Yang et al., “Fast Optimal Antenna Placement for Distributed MIMO Radar with Surveillance Performance,” vol. 22, no. 11, pp. 1955–1959, 2015.

[25] A. E. Forooshani, A. A. Lotfineyestanak, S. Member, D. G. Michelson, and S. Member, “Optimization of Antenna Placement in Distributed MIMO Systems for Underground Mines,” vol. 13, no. 9, pp. 4685–4692, 2014.

[26] Italo Atzeni, Jesus Arnau, and Merouane Debbah “Fractional Pilot Reuse in Massive MIMO Systems,” IEEE ICC, pp. 1030–1035, 2015.

[27] L. Shen, Y. Yao, H. Wang, H. Wang, and S. Member, “ICA Based Semi-Blind Decoding Method for a Multicell Multiuser Massive MIMO Uplink System in Rician / Rayleigh Fading Channels,” vol. 16, no. 11, pp. 7501–7511, 2017.

Cluster based Routing Protocols for Wireless Sensor Networks: An Overview

Muhammad Nadeem Akhtar[1], Arshad Ali[2], Zulfiqar Ali[3], Muhammad Adnan Hashmi[4] and Muhammad Atif[5]
Department of Computer Science & Information Technology
The University of Lahore, Lahore, 55150, Pakistan

***Abstract*—Energy consumption of nodes in Wireless Sensor Networks (WSNs) is a very critical issue, particularly in scenarios where the energy of nodes cannot be recharged. Optimal routing approaches play a key role in energy utilization, so there is great importance of energy efficient routing protocols in WSNs. Energy efficient routing protocols in WSNs are categorized into four schemes, namely (i) communication model, (ii) topology based model, (iii) reliable routing, and (iv) network structure. Network structure category is further divided into flat and cluster-based approaches. This work focuses on a subtype of "network structure" scheme known as clustered based routing protocols, which are mainly used in WSNs for reduction in energy consumption. This work reviews and provides an overview of prominent cluster based energy efficient routing protocols on the basiss of some primary performance metrics such as (i) energy efficiency, (ii) algorithm complexity, (iii) scalability, (iv) data delivery delay, and (v) clustering approach. Finally, this work discusses some latest research trends with respect to cluster based energy efficient routing protocols in WSNs.**

Keywords—Wireless sensor networks; network structure; clustering protocols; energy efficiency

I. Introduction

Since the last twenty years, there is a rapid growth with respect to technologies in the field of data communication networks. This technological progress facilitates organizations by providing very easy and secure working environment. Wireless networks enable organizations to get rid of expensive procedure of using cables for the purpose of connecting equipment located at different locations. This motivates organizations to use wireless networks for communication purpose.

From topological perspective, wireless networks are commonly classified into two modes i.e., (i) infrastructure mode, and (ii) ad hoc mode. The former supports communication between nodes through a Base Station (BS), while in ad hoc mode, all nodes can communicate with each other directly without requiring any infrastructure and no node is superior to any other node in the absence of any central entity.

A Wireless Sensor Networks (WSN) is a set of low-cost and small-sized sensor nodes having limited communication range, energy, processing, and storage capacity. From network design perspective, WSNs are classifies into structured and unstructured networks. In the former, the deployment of nodes is made with proper planning while in latter the same is done in an ad hoc manner [1].

A wireless sensor network is a combination of various sensor nodes connected with each other. Physical location where these nodes are deployed is known as a sensor field. Data from any node is transferred to other linked nodes and aggregated at sink node in order to be accessible to end users as shown in Fig.1 [2].

Each sensor node has four major hardware components i.e., (i) sensing unit, (ii) a processing unit, (iii) transmitting / receiving unit, and (iv) power unit. Each sensor comprises of application dependent two additional components, namely (i) location finder system, and (ii) mobilizer. Sensor and analog to digital converter (ADC) are two sub-parts of sensing unit. Initially, the data is observed by a sensor which is forwarded to ADC for conversion into digital form. Then, digital data is sent to the processing unit, which is usually linked with a storage unit consisting of a small storage capacity. In order to perform assigned activities, sensor units cooperate with each other by using procedures organized by processing unit. The transceiver helps a node in connecting with the network. Power unit, considered as the most important part, provides power to all the remaining units. The power may be provided through solar cells or by using power generator system (refer to Fig.2) [2].

Sensors used in WSNs have various kinds like acoustics, seismic visual, low sampling magnetic, thermal, radar and infrared. These sensors are capable of monitoring several conditions such as noise level, soil makeup, lightening, vehicular movement, temperature, humidity and pressure etc.

WSNs have various application areas i.e., performance monitoring of industrial machines, environmental monitoring, monitoring of health and military battlefield [3].

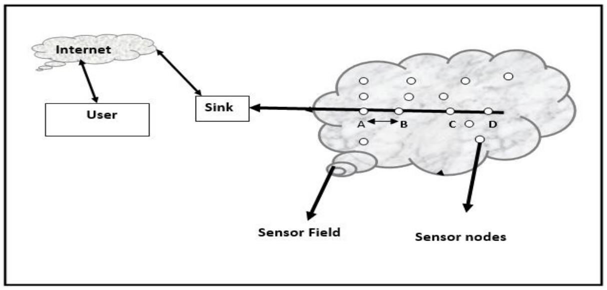


Fig. 1. Sensor Nodes in a Sensor Field.

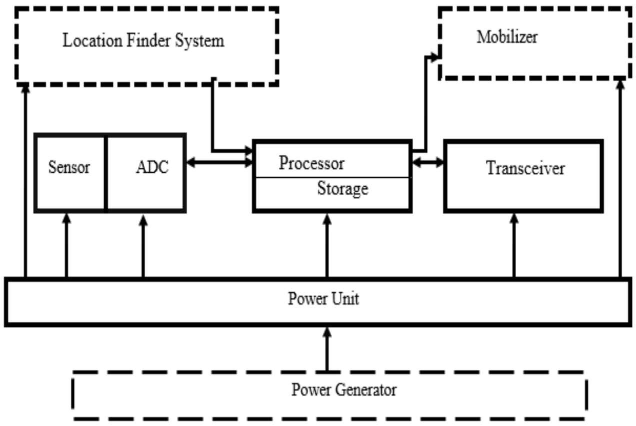


Fig. 2. Sensor Nodes' Hardware Components.

Routing protocols are of great attention because of versatilities lying in network architecture as well as in applications using wireless sensor networks. Traditional routing protocols are not applicable in WSNs; therefore WSNs require routing protocols different than traditional ones. Consequently, many energy-efficient routing protocols were developed for WSNs for providing efficient delivery of to the destination. Keeping in view the nature of application and network architecture, each energy-efficient routing protocol may possess specific features.

This research work provides an overview of existing energy-efficient cluster based routing protocols in the context of WSNs. Moreover, it lists a brief comparison of the studied protocols. The rest of the paper is organized as follows: Section II describes the related work. Section III provides the review of the cluster based hierarchical routing protocols of WSNs. Section IV gives comparative analysis of prominent hierarchical clustering protocols in terms of performance metrics and final Section V concludes the study.

II. Related Work

Many surveys have already been conducted in the area of WSNs energy efficient routing protocols, from different perspectives. However, this paper reviews hierarchal / cluster-based routing protocols in WSNs.

In 2004, Al karaka and A E Kamal [4] surveyed routing techniques in WSNs. This work provided a taxonomy about WSN routing protocols by dividing them into two major categories (i) network structure and (ii) protocol operation. The network structure is further divided into flat, hierarchal and location-based routing. The location based is further divided into (i) negation based, (ii) multipath based, (iii) query based, (iv) QoS based, and (v) coherent based routing. This work also exposed some design tradeoff between communication overheads and energy saving in some routing protocols paradigm.

In a survey in 2005, Kemal Akkaya et al. [5] classified WSNs routing protocols as data-centric, hierarchal and location-based. Each protocol was placed in one basic category, while a few protocols belonged to more than one class conserving various metrics such as QoS, network flow, and data aggregation. All of the clustering protocols were not discussed in this work.

In 2007, Abbasi et al. [6] presented a taxonomy of different clustering schemes and provided an overview of clustering protocols and algorithms from the perspective of variable convergence time and constant convergence time. Moreover, their study provided a comparison of some popular clustering methods.

In 2008, Deosarkar et al. [7] discussed cluster head (CH) selection techniques on the basis of classification as (i) deterministic, (ii) adaptive, and (iii) combined metric schemes. The authors compared the cost of CH selection from various angles like cluster information, creation, and distribution of clusters.

In 2010, Shio Kumar Singh et al. [8] described cluster based energy efficient routing protocols in WSN. The authors highlighted some pitfall and disadvantages of individual protocols along with some future trends and constraints lying in this area.

In 2012, Xuxun Liu [9] comparatively expressed a better survey on cluster-based energy efficient WSN routing protocols. The author developed a novel taxonomy about clustering methods on WSN rather than detailed clustering attributes. This work analyzed some prominent clustering routing protocols in WSNs and compared them through different approaches as discussed in the taxonomy about the cluster (refer to Fig.3). The author described three clustering approaches i.e., (i) centralize, (ii) distributed, and (iii) hybrid. Centralize clustering approach is responsible for making clusters and CH selection. Distributed approach allows all cluster nodes to work as CH for the current round. Hybrid approach combines the properties of both centralize and distributed approaches.

In 2013, Nikolaos A. P et al. [10] presented a detailed survey on overall energy efficient WSN routing protocols by dividing them into four main categories on the basis of energy efficiency, nemly (i) network structure, (ii) communication model based, (iii) topology based, and (iv) reliable routing based. The first scheme is further divided flat and cluster based approaches. The second scheme is classified in three subtypes i.e., (i) query based, (ii) non query based/negation based, and (iii) coherent based. The third scheme is further divided into location-based and mobile agent based ones. The fourth scheme is divided into QoS based and multipath based schemes. Fig. 4 presents the complete picture of their division. The present work focuses cluster based routing protocols in detail.

In 2014, Agam Gupta and Anand Nayyar [11] discussed many routing protocols. Traditional routing protocols being used in WSNs lack in load balancing and efficiency of energy. The use of clustering not only improves network life time but also supports load balancing. There are many clusters and each cluster consists of many inter-connected sensor nodes, while one of them works as a cluster head (CH). Each cluster head gathers data from the nodes belonging to the cluster and transfer that data to the BS (refer to Fig.5). There is intra cluster as well as inter cluster data communication between cluster head and member nodes of the cluster [11].

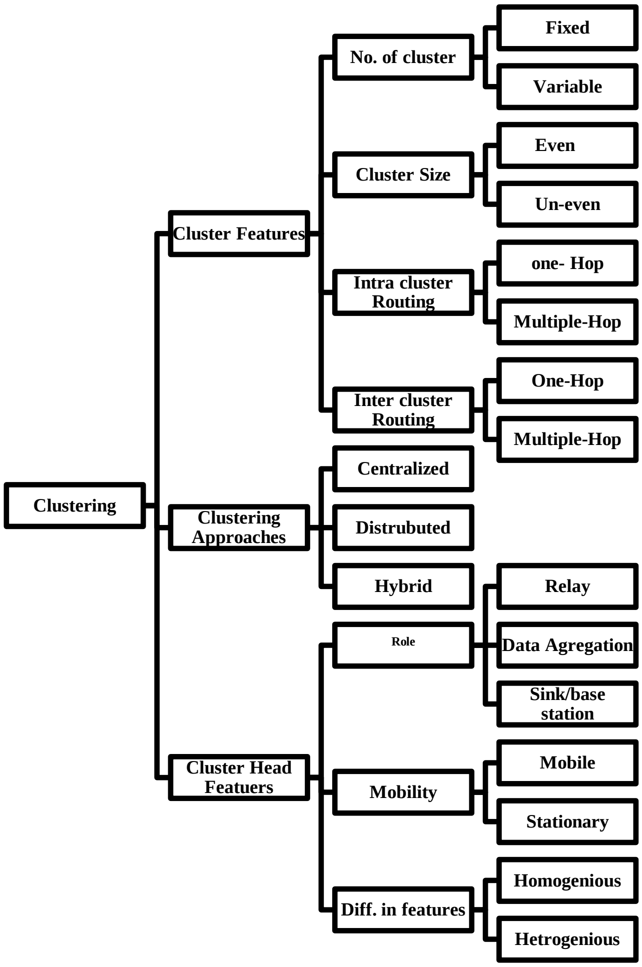


Fig. 3. Different Aspects of Cluster in WSNs [9].

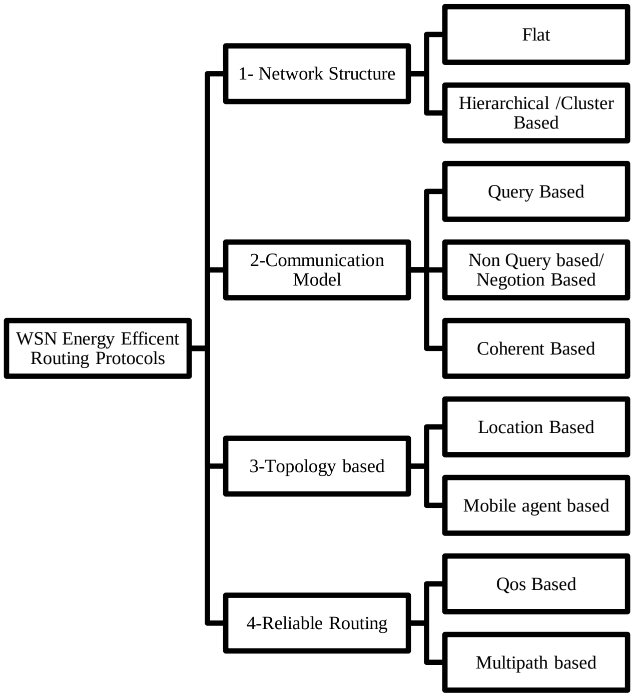


Fig. 4. Energy Efficient Routing Protocols in WSNs [10].

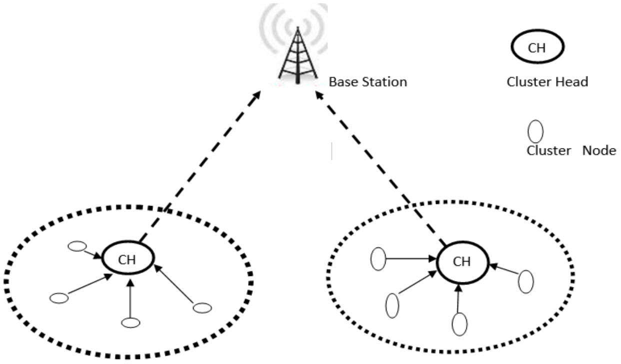


Fig. 5. Cluster in WSNs [11].

In 2015, Santar Paul Singh and SC Sharma [12] conducted a survey on cluster based energy efficient WSN routing protocols. The authors described taxonomy of WSN routing protocols into five categories, namely (i) initiator of communication, (ii) Path establishment, (iii) Network structure, (iv) protocol operation, and (v) next hope selection. Network structure scheme is Specifically further classified into (i) flat, (ii) cluster based, and (iii) location based. The authors in this survey did not review clustering protocols individually. They classified cluster-based routing protocols into three sub-categories, i.e., (i) block cluster-based, (ii)grid cluster based, and (c) chain cluster based. According to this classification, different clustering protocols lie under these three clustering schemes. In the end, the authors discussed some merits and limitations of some prominent cluster-based routing protocols.

In 2015, Priyanka Sharma and Inderjeet kaur [13] discussed WSNs routing protocols by classifying them into three main categories, i.e., (i) path establishment, (ii) network structure, and (c) protocol operation. First scheme path establishment is further divided into proactive, reactive and hybrid. Second scheme network structure is further divided into flat, hierarchal and location based. The third scheme is further classified into eight sub-types, namely (i) query, (ii) bio-Inspired, (iii) multipath, (iv) negation based, (v) QoS, (vi) non-coherent, (vii) coherent, and (viii) mobility. The authors discussed some metrics, pros, cons, and applications of some clustering protocols lying in above-mentioned categories.

In 2015, Ibrihich Ouafaa et al. [14] discussed and compared some prominent cluster-based routing protocols by classifying them into WSN and ad-hoc categories. The authors also compared these prominent protocols considering some important metrics.

In 2016, Yan et al. [15] classified WSNs routing protocols into data-centric, location-based and hierarchal depending on network structure. In data-centric approach, metadata approach is used by the protocols to sense and transmit information to base station. Hierarchal approach adopts clustering technique which can be made by grouping sensor nodes. The cluster reduces the energy utilization of sensor nodes. Clustering technique is more scalable and is used in a number of various applications. The location-based approach uses position/ location of nodes to route the data intelligently.

In 2017, Syed Bilal Hussain Shah et al. [16] conducted a survey on hierarchal routing protocols in WSNs to increase network lifetime and conserve energy. In this survey, they reviewed some of the hierarchal routing protocols like LEACH, LEACH- TLCH, APTEEN, TEEN and proposed new scheme adaptive threshold. They also discussed some limitations of LEACH and some merits of newly proposed adaptive threshold. Adoptive Threshold attained good results as compared to some of the previously discussed schemes. But the authors did not studied or compared all hierarchal/ clustering protocols with the newly proposed scheme.

In 2018, G. Beni and C. Selden Christopher [17] discussed a few cluster-based protocols like LEACH, PEGASIS, HEED, TEEN and APTEEN from the perspective of comparing performance metrics like energy efficiency, cluster stability, delivery delay, and scalability. But the authors did not present a detailed survey on all hierarchal / cluster-based protocols of WSN.

Our work, to the best of our knowledge, is a more comprehensive study covering maximum number of famous hierarchal /cluster based protocols of WSNs with different clustering approaches.

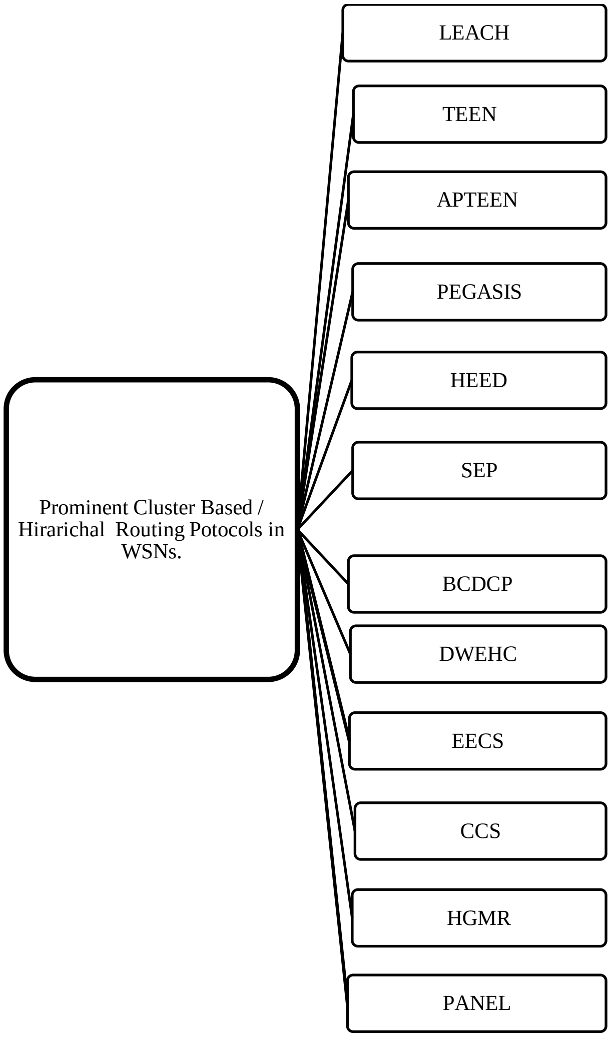

Prominent Cluster Based / Hirarichal Routing Potocols in WSNs.
LEACH
TEEN
APTEEN
PEGASIS
HEED
SEP
BCDCP
DWEHC
EECS
CCS
HGMR
PANEL

Fig. 6. Cluster-Based Routing Protocols in WSNs.

III. Cluster-based Routing Protocols in WSNs

This section provides an overview of prominent energy efficient cluster-based hierarchal routing protocols of WSNs. In cluster-based routing protocols, multiple nodes connected with each other in a sensor field make a group having one node among them as a cluster head. Data transformation from that particular cluster to sink node occurs through the cluster head (CH). In this way, the energy of other nodes is saved. So in this classification of protocols, the major aspect is clustering. Fig. 6 depicts prominent cluster based / hierarchal routing protocols in WSNs while details of these protocols are provided in subsequent sub-sections.

A. LEACH

In 2000, Heinzelman et al. [18] proposed one of the famous cluster based routing protocol for WSNs namely "Low Energy Adaptive Clustering Hierarchy" (LEACH). LEACH evenly distributed the load of energy between all sensors of the network using random based rotation of cluster head. In order to make dynamic networks more scalable and robust, LEACH used localize coordination. In LEACH, sensor nodes scattered in field organize themselves to make local clusters, among these sensor nodes one node becomes local base station or cluster head (CH). This CH works as router, transfers the signals from all sensor nodes to the sink. LEACH saves energy due to transfer of data by CH rather than transferring of the data individually by all sensor nodes. Optimal number of nodes considered as cluster head are about five percent of the total nodes. In LEACH, all data processing including "data aggregation" and "data fusion" is held locally in the cluster which results in reduction of energy dissipation as well as increasing life time of the system. Cluster head is changed randomly, so energy dissipation between all nodes becomes balanced. CH changing decision is made by randomly selecting a number between 0 and 1.

If selected number (shown in Eq.1) is below the threshold, node may become the CH for some specific round.

$$\mathrm{T(n)} = \begin{cases} \frac{\mathrm{P}}{1-\mathrm{P}(\mathrm{rmod}\frac{1}{\mathrm{p}})} & \text{if n Ɛ G} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

In Eq.1, variable "P" is desired percentage of the CH (for example .05), variable "r" means current round whereas variable "G" means a set of those nodes which are not selected as CH from last 1/P rounds.

The authors claimed that LEACH reduced 8 times energy as compared to direct communication (DC) and minimum transmission energy (MTE) protocol. LEACH, being a single hop routing protocol, possesses some deployment limits in larger networks where every individual node directly communicates to cluster head and sink.

B. TEEN

In 2001, Arati Manjeshwar and Dharma P. Agarwal [19] proposed "Threshold Sensitive Energy Efficient Sensor Network" (TEEN). According to authors TEEN was the first protocols of its time developed for "Reactive Networks". In reactive networks nodes immediately react to drastic and sudden changes in value of sensed attributes.

In this protocol the CH broadcasts to its member nodes, at change time of every cluster, in addition to its attributes. TEEN was a combination of data-centric and hierarchical approach. Number of transmissions from member sensor nodes to CH are reduced in this protocol.

Two kinds of threshold values to ordinary member sensor nodes work in TEEN, i.e., (i) when a cluster is formed, and (ii) when CH broadcasts. The first value is called hard threshold (Ht) and second is called soft threshold (St).

Hard threshold (Ht) is sensed attribute's absolute/ minimum value, at which sensing node should turn on its transmitter for reporting to its CH. Soft threshold (St) is small change occurred in sensed attributes, which causes to dictate the node to switch on the transmitter for the transmission purpose.

TEEN saves large amount of energy by reducing number of transmissions between cluster head and member sensor nodes.

The main shortcoming of TEEN is detection of dead nodes. Another limitation of TEEN is that it is difficult to forecast the reason if node is not sending the data. This can be happened because of two main reasons (i) node is unable to meet threshold value, and (ii) node may be dead. TEEN is suitable for time critical applications and suitable for energy consumption and response time [19].

C. APTEEN

In 2002, Arati Manjeshwar and Dharma P. Agarwal [20] proposed an improvement to overcome deficiencies of TEEN named as "Adaptive Periodic Threshold Sensitive Energy Efficient Sensor Network" (APTEEN). LEACH was considered suitable for proactive networks and TEEN was suitable for reactive networks. In APTEEN, the authors used hybrid network approach which had combined best features of both LEACH and TEEN. APTEEN was suitable for time critical events as well as to obtain data periodically. Simulation results showed that network lifetime and energy dissipation of APTEEN existed between TEEN and LEACH.

D. PEGASIS

In 2002, Stephanie Lindsey et al. [21] proposed a chain based protocol namely "Power-Efficient Gathering in Sensor Information Systems" (PEGASIS) which was an improvement over LEACH. In PEGASIS each node sends and receives data to only nearby neighbor nodes. Data reaches to the base station in turns, due to which energy consumption per round is reduced. All member sensor nodes are connected with each other in such a way that they make a chain. Using greedy algorithm, chain computation may be initialized by broadcasting data from a node or base station to all member sensor nodes, as shown in Fig.7.

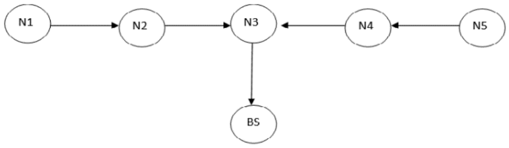


Fig. 7. Chaining in PEGASIS.

E. HEED

In 2004, Ossama Younis et al. [22] proposed protocol named as Hybrid, Energy-Efficient Distributed Clustering (HEED) which was very excellent clustering based protocol.

In HEED node's residual energy is parameter to elect the CH.Node's degree or density is used as a metric, in selection of a cluster to get power balancing. HEED was mainly an improvement over LEACH.

F. SEP

In 2004, G. Smaragdakis et al. [23] proposed a protocol named as Stable Election Protocol (SEP) for cluster-based Hetero-genius wireless sensor networks. SEP is an improved version of LEACH and works like it. SEP prolongs time interval known as "stability period" of the first node before its death, which is very crucial for such kind of applications where feedback from sensor network is considered very reliable. In SEP cluster head (CH) is elected on the basis of energy as a "parameter". In this protocol a node independently selects itself as CH on the base of its own initial energy. SEP depends upon each node's weighted "election probability" to make CH with respect to each node's "remaining energy". From simulation results, it can be concluded that SEP had longer period stability and greater average throughput as compared to existing clustering based heterogeneous oblivious protocols. The study also showed that SEP is more resilient in advance node's energy efficiency.

G. BCDCP

In 2005, Siva D. Muruganathan et al. [24] proposed Base-Station Controlled Dynamic Clustering Protocol (BCDCP). BCDCP maintains clusters and routing paths by utilizing high energy base station. In this protocol cluster head rotations are performed randomly. Intensive energy tasks are performed in this protocol. In BCDCP main concept is formation of balance clusters where every cluster head serves an equal number of member sensor nodes by avoiding cluster head overload. In overall sensor field there is uniform placement of cluster heads. Data is transferred to the base station by utilizing CH-to- CH routing.

H. DWEHC

In 2005, Ping Ding et al. [25] proposed a protocol namely Distributed Weight based Energy Efficient Clustering (DWEHC), in which every node in its enclosure region, first of all, locates its neighbor, calculates weight of itself that depends upon two factors namely, (i) its distance from the neighbor node and (ii) residual energy. In that enclosed region, a node having maximum weight is selected as a CH. Other neighbor nodes become member node under this CH hierarchy. This clustering process finally terminates after seven iterations. This clustering process has no dependency on the size and topology of the network. The authors showed through simulation that this protocol performed well. The performance of this protocol was also analyzed from the perspective of Inter-cluster and intracluster communication. Finally, authors compared DWEHC performance with HEED-AMRP algorithm and concluded that it outperforms HEED-AMRP in "energy consumption" and better "cluster generation" perspectives.

I. EECS

In 2005 and 2006, Mao Ye et al. [26, 27] proposed Energy Efficient Clustering Scheme (EECS) for WSNs. EECS is a clustering algorithm that is more suitable for periodical data gathering applications. EECS is like a LEACH. In this scheme sensor network is divided into many clusters and data communication between CH and BS is single-hop. In this scheme, a node candidate to become a CH competes for a given round to gain the ability to elevate the CH. During this competition, CH candidate nodes broadcast residual energy among their neighboring CH candidates. In this duration, if a node under consideration could not get another node with more residual energy than it, then itself becomes the CH. Cluster formation of EECS is different from LEACH. EECS improves in the capability to LEACH by introducing dynamic size in clusters which are based on "distance" of the cluster from BS.

J. CCS

In 2007, Sung-Min Jung et al. [28] proposed a protocol as Concentric Clustering Scheme (CCS), to improve the performance of PEGASIS. As in PEGASIS, data transmission is redundant because when one node is selected as a head node regardless of the base station it takes data from both side nodes to convey it to BS, so data transmission from head node to BS become redundant. To cope with this issue of redundancy, CCS protocol was proposed. The main concept of CCS is to consider the location of the base station so that the lifetime and performance of WSNs can be prolonged. This became possible to achieve, using CCS. In CCS, WSNs are divided into concentric shaped clusters to give data transmission flow. CCS enhanced performance by using four processes, namely (i) level assignment to each node relevant to base station, (ii) constructing chain in level area using greedy algorithm, (iii) constructing chain between head nodes of each level, and (iv) transferring of data from higher level head node to other lower level head node. CCS saved 35 % energy as compare to PEGASIS.

K. HGMR

In 2008, Dimitrios Koutsonikolas et al. [29] proposed a protocol named as Hierarchical geographic multicast routing (HGMR). HGMR is basically location-based and multicast protocols for WSNs. Impeccably it incorporates innovations in the "locations based" & "multicast" and it optimizes them for WSNs. HGMR performs this by concurrently providing scalability and energy efficiency to the networks of large size. It can be concluded from simulation results that HGMR has combined the strength of two protocols namely Hierarchal Rendezvous Point Multicast (HRPM) and Geographic Multicast Routing (GMR). HGMR protocols decompose multicast groups into the subgroups. It uses GMR's local multicast scheme to forward the data packets with multi branches of a Multicast tree in the single transmission. In HGMR, using mobile Geographic Hashing mechanism, multicast groups can be divided into their subgroups. The deployed area is divided into partitions of various equal sized square shaped subdomains called "Cells" and each cell is consisted of subgroups of members having managable size. There exists an access point (AP) in each cell which is responsible for all members of that cell. A Rendezvous Point (RP) manages all the APs.

L. PANEL

In 2007 and 2010, Levente Buttyan and Peter Schaffer [30, 31] proposed a protocol named as Position-based Aggregator Node ELection (PANEL) for WSNs. There exist some other aggregator node election protocols but PANEL has a novelty from them in a sense that it supports such kind of sensor network applications which are asynchronous. The sensor collects the reading information through the base station after some delay. Main motivational factor in the design of PANEL was its support for reliable and consistent application of data storage like TinyPEDS. PANEL deals with load balancing and also supportx intra and intercluster routing by allowing communication between sensor and aggregator, aggregators itself, an aggregator to BS and BS to aggregators. Cluster formation and energy consumption capabilities of PANEL are better than HEED. Following are key merits of PANEL: (1) PANEL is energy efficient ensuring load balancing due to the election of each node as an aggregator (2) Beside synchronous scenes this protocol also supports the asynchronous application.

Following are key limitations of PANEL: (i) The supposition that cluster formation is found/ determined before deployment thus cannot be applied upon WSN dynamics, (ii) it has information about the geographical position of the nodes, that is used to find which node must be an aggregator. In WSNs there is a constraint that geographical position is not always available except some special conditions like hardware and software having GPS feature, and (iii) an assumption about PANEL described by the authors is that within cluster nodes form a subnetwork, due to this there may occur such a situation that nodes within the cluster could not hear the announcement of nodes closest to reference point, and they may elect aggregator to another node.

IV. Comparative Analysis and Research Challenges

Table 1 shows a comparative analysis of various cluster-based protocols (discussed earlier in Section III) on the basis of different performance metrics, namely (i) energy efficiency, (ii) algorithm complexity, (iii) delay in data delivery, (iv) scalability, and (v) clustering approach. A tradeoff was observed in terms of energy efficiency and data delivery delay, i.e., BCDCP is very poor regarding energy efficiency but offers small delay. It was also observed that some protocols perform much better in terms of scalability; however, their performance is lower if other metrics are taken into account i.e., scalability of HGMR is very high while having very poor energy efficiency. It is worth mentioning that almost all selected protocols in this review follow the distributed clustering approach. Algorithm complexity is noticed from very low to very high.

The following research challenges require attention from research community

- The design of energy efficient cluster based protocols for wireless body area networks for the purpose of improving overall energy efficiency is an interesting domain to explore.

- Further investigation is required considering integration of these protocols with technologies such as “Internet of things”, “Vehicular Ad hoc Networks” and many others.
- Sensor nodes are deployed on vehicles in order to monitor events. Data aggregation is an important issue in VANTEs keeping in view high mobility of vehicular nodes.
- Security is one of the main concerns in WSNs due to its operation in open environment which requires serious efforts. For secure data transmission, the existing security approaches cannot be applied in present form due to limited resources of WSNs. Thus, there is need of mechanisms which provide secure data transmission by using less energy resources.
- The design of routing protocols in the context of Internet of Things requires attention from research community, an overview of the same is provided in [32].

TABLE I. COMPARISON BETWEEN PROMINENT CLUSTERING BASED ROUTING PROTOCOLS IN WSNS

Prominent Cluster based Routing Protocols	**Energy Efficiency**	**Algorithm Complexity**	**Delay in data delivery**	**Scalability**	**Clustering Approach**
LEACH	Very Poor	Low	Very small	Very Low	Distributed
TEEN	Very High	High	Small	Low	Distributed
APTEEN	Medium	Very High	Small	Low	Distributed
PEGASIS	Poor	High	Very Large	Very Low	Distributed
HEED	Medium	Medium	Medium	Medium	Distributed
SEP	Medium	Very Low	Very Small	Medium	Distributed
BCDCP	Very Poor	Very High	Small	Very Low	Centerlize
DWEHC	Very High	Medium	Medium	Medium	Distributed
EECS	Medium	Very High	Small	Low	Distributed
CCS	Poor	Medium	Large	Low	Distributed
HGMR	Poor	Low	Medium	Very High	Distributed
PANEL	Medium	High	Medium	Low	Distributed

V. CONCLUSION

Wireless sensor networks (WSNs) remained an emerging area of research for the last two decades. There are various applications of WSNs like industrial machine performance, environmental monitoring, health monitoring, and military battlefield. Major short fall of WSNs is energy dissipation of nodes deployed in the field of these application areas. Optimal and effective routing protocols and approaches play a key role in energy utilization, so the importance of energy efficient routing protocols in WSNs is significant. In this paper, we reviewed well-known cluster-based energy efficient routing protocols in WSNs. This work highlighted a few clustering approaches and characteristics considered in energy efficient routing protocols. On the basis of primary performance metrics i.e. energy efficiency, algorithm complexity, delay in data delivery, scalability and clustering approach, a comparative analysis has been done among prominent cluster based energy efficient routing protocols used in WSNs. This study concludes that there is not any single protocol which has the capability to perform excellently considering all metrics. If one protocol is good in energy dissipation, it may have more delivery delay or its algorithm may be complex, on the other hand if a protocol offers less delay or low complexity in algorithm then it may be less energy efficient.

REFERENCES

[1] Feeney, Laura Marie, and Martin Nilsson. "Investigating the energy consumption of a wireless network interface in an ad hoc networking environment." In INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 3, pp. 1548-1557. IEEE, 2001.

[2] F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” Computer networks, vol. 38, no. 4, pp. 393-422, 2002.

[3] Guoe, Menting, Jiadue Shan, and Yicheo Yong. "Evaluation of sensor network capability in a practical problem." INTERNATIONAL JOURNAL OF ADVANCED AND APPLIED SCIENCES 1, no. 7 (2014): 1-9.

[4] J. N. Al-Karaki, and A. E. Kamal, “Routing techniques in wireless sensor networks: a survey,” IEEE wireless communications, vol. 11, no. 6, pp. 6-28, 2004.

[5] K. Akkaya, and M. Younis, “A survey on routing protocols for wireless sensor networks,” Ad hoc networks, vol. 3, no. 3, pp. 325-349, 2005.

[6] A. A. Abbasi, and M. Younis, “A survey on clustering algorithms for wireless sensor networks,” Computer Communications, vol. 30, no. 14, pp. 2826-2841, 2007.

[7] Deosarkar, Bhaskar P., Narendra Singh Yadav, and R. P. Yadav. "Clusterhead selection in clustering algorithms for wireless sensor networks: A survey." In Computing, Communication and Networking, 2008. ICCCn 2008. International Conference on, pp. 1-8. IEEE, 2008.

[8] S. K. Singh, M. Singh, and D. Singh, “A survey of energy-efficient hierarchical cluster-based routing in wireless sensor networks,” International Journal of Advanced Networking and Application (IJANA), vol. 2, no. 02, pp. 570-580, 2010.

[9] X. Liu, “A survey on clustering routing protocols in wireless sensor networks,” sensors, vol. 12, no. 8, pp. 11113-11153, 2012.

[10] N. A. Pantazis, S. A. Nikolidakis, and D. D. Vergados, “Energy-efficient routing protocols in wireless sensor networks: A survey,” IEEE Communications surveys & tutorials, vol. 15, no. 2, pp. 551-591, 2013.

[11] A. Nayyar, and A. Gupta, “A comprehensive review of cluster-based energy efficient routing protocols in wireless sensor networks,” IJRCCT, vol. 3, no. 1, pp. 104-110, 2014.

[12] S. P. Singh, and S. Sharma, “A survey on cluster-based routing protocols in wireless sensor networks,” Procedia computer science, vol. 45, pp. 687-695, 2015.

[13] P. Sharma, and I. Kaur, “A Comparative Study on Energy Efficient Routing Protocols in Wireless Sensor Networks,” International Journal of Computer Science Issues (IJCSI), vol. 12, no. 4, pp. 98, 2015.

[14] Ouafaa, Ibrihich, Laassiri Jalal, Krit Salah-ddine, and El Hajji Said. "The comparison study of hierarchical routing protocols for ad-hoc and wireless sensor networks: A literature survey." In Proceedings of the

The International Conference on Engineering & MIS 2015, p. 32. ACM, 2015.

[15] J. Yan, M. Zhou, and Z. Ding, "Recent advances in energy-efficient routing protocols for wireless sensor networks: A review," IEEE Access, vol. 4, pp. 5673-5686, 2016.

[16] Shah, Syed Bilal Hussian, Yin Fuliang, Inam Ullah Khan, Chen Zhe, and Muhammad Zakarya. "Collating and Analysing State-of-the-Art Hierarchical Routing Protocols in WSN to Increase Network Lifetime and Conserve Energy." In Proceedings of the International Conference on Future Networks and Distributed Systems, p. 31. ACM, 2017.

[17] Beni, G., and C. Seldev Christopher. "Analysis of Energy Efficient Routing Protocols in Wireless Sensor Networks." vol 14 Taga journal, 2018.

[18] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on wireless communications, vol. 1, no. 4, pp. 660-670, 2002.

[19] Manjeshwar, Arati, and Dharma P. Agrawal. "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks." In null, p. 30189a. IEEE, 2001.

[20] Manjeshwar, Arati, and Dharma P. Agrawal. "APTEEN: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks." In ipdps, p. 0195b. IEEE, 2002.

[21] Lindsey, Stephanie, and Cauligi S. Raghavendra. "PEGASIS: Power-efficient gathering in sensor information systems." In Aerospace conference proceedings, 2002. IEEE, vol. 3, pp. 3-3. IEEE, 2002.

[22] 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, no. 4, pp. 366-379, 2004.

[23] G. Smaragdakis, I. Matta, and A. Bestavros, SEP: A stable election protocol for clustered heterogeneous wireless sensor networks, Boston University Computer Science Department, 2004.

[24] S. D. Muruganathan, D. C. Ma, R. I. Bhasin, and A. O. Fapojuwo, "A centralized energy-efficient routing protocol for wireless sensor networks," IEEE Communications Magazine, vol. 43, no. 3, pp. S8-13, 2005.

[25] P. Ding, J. Holliday, and A. Celik, "Distributed energy-efficient hierarchical clustering for wireless sensor networks," Distributed computing in sensor systems, pp. 466-467, 2005.

[26] Ye, Mao, Chengfa Li, Guihai Chen, and Jie Wu. "EECS: an energy efficient clustering scheme in wireless sensor networks." In Performance, Computing, and Communications Conference, 2005. IPCCC 2005. 24th IEEE International, pp. 535-540. IEEE, 2005.

[27] M. YE, C. LI, G. CHEN, and J. WU, "An Energy Efficient Clustering Scheme in Wireless Sensor Networks," Ad Hoc & Sensor Wireless Networks, vol. 17, pp. 33, 2006.

[28] Jung, Sung-Min, Young-Ju Han, and Tai-Myoung Chung. "The concentric clustering scheme for efficient energy consumption in the PEGASIS." In Advanced Communication Technology, The 9th International Conference on, vol. 1, pp. 260-265. IEEE, 2007.

[29] D. Koutsonikolas, S. M. Das, Y. C. Hu, and I. Stojmenovic, "Hierarchical geographic multicast routing for wireless sensor networks," Wireless networks, vol. 16, no. 2, pp. 449-466, 2010.

[30] Buttyan, Leventa, and Peter Schaffer. "PANEL: Position-based Aggregator Node Election in Wireless Sensor Networks." In Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference on, pp. 1-9. IEEE, 2007.

[31] L. Buttyán, and P. Schaffer, "Position-based aggregator node election in wireless sensor networks," International Journal of Distributed Sensor Networks, vol. 6, no. 1, pp. 679205, 2010.

[32] Belhaj, Salem, and Sofian Hamad. "Routing protocols from wireless sensor networks to the internet of things: An overview." INTERNATIONAL JOURNAL OF ADVANCED AND APPLIED SCIENCES 5, no. 9 (2018): 47-63.

An Empirical Investigation on a Tool-Based Boilerplate Technique to Improve Software Requirement Specification Quality

Umairah Anuar[1], Sabrina Ahmad[2], Nurul Akmar Emran[3]
Centre for Advanced ComputingTechnology (C-ACT)
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka (UTeM)
Malacca, Malaysia

***Abstract*—The process of producing software requirements specification (SRS) is known to be challenging due to the amount of effort, skills and experience needed in writing good quality SRS. A tool-based boilerplate technique is introduced to provide assistance in identifying essential requirements for a generic information management system and translating them into standard requirements statements in the SRS. This paper presents an empirical investigation to evaluate the usability of the prototype. Results showed that the tool-based boilerplate technique has high usability, usefulness and ease of use.**

***Keywords*—Empirical investigation; usability; software requirements**

I. Introduction

SRS play a vital role in software development. It is a fundamental document which consists of a set of requirements that forms the foundation of software development process. The quality of the SRS is crucially important because requirement is a basic of a system the molds the shape of the system the need to be developed. Poor quality SRS does not only lead to poor quality software but also increase in development and sustainment costs which cause major schedule overruns [1].

Based on the importance of the SRS, we reviewed literature from year 2000 to 2014 on what efforts that has been done to reduce the problem arise in SRS and we found out boilerplate is one of the method used to improving the SRS quality. According to the research, a boilerplate technique was adopted as one of the semiformal language because boilerplate is proposed as a bridge between formal and informal specification [2].

Boilerplate is a section of text that can be included in many places with little or no alteration. It is a text that can be reused in new contexts or applications without greatly changed from the original. The same understanding is adapted to the software engineering specialization to particularly produce an SRS document to meet some standard. The boilerplate technique ability to produce requirements statements in some sort of controlled environment is seen beneficial to reduce the possibility of defects. The control is made to handle the flexibility of natural language (NL) which usually leads to ambiguity and inconsistency problems.

The purpose of this paper is to present the findings from an empirical investigation on the usability of the tool-based boilerplate technique. Following Introduction, Section II elaborates on the Tool-based Boilerplate Technique. This is followed by Section III which explains the empirical investigation protocol. Next, Section IV presents the results and Section V elaborates on further analysis. Finally, Section VI concludes the paper.

II. Tool-Based Boilerplate Technique

The term boilerplate is used to refer to the sheet steel used to make boilers in the field of printing [3]. Nevertheless, in computer programming, the term boilerplate is used to represent the section of code. Boilerplate is similar to a template that holds the layout and style information. The main advantage of boilerplate is the reusability, where it can be used in several places in a computer program with little (or no) alteration [4]. Boilerplate also used as a preliminary basis for requirement checking. The usefulness of boilerplate is also claimed by Arora et al. [5] who stated that boilerplate provides a simple yet effective way for increasing the quality of requirement by avoiding complex structure, ambiguity and inconsistency in requirements. Our boilerplate technique is an adaptation from Rupp's boilerplate as shown in Fig. 1 [6]. Any requirement may easily be mapped to the Rupp's boilerplate because Rupp's boilerplate structure has a very wide usability and able to fit most of the requirement.

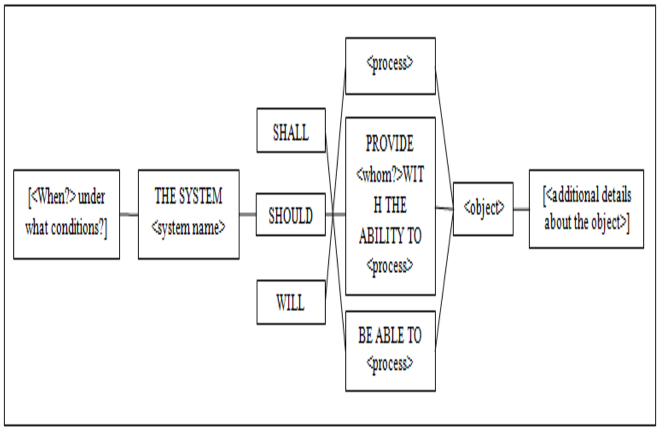

[<When?> under what conditions?]
THE SYSTEM <system name>
SHALL
SHOULD
WILL
<process>
PROVIDE <whom?>WITH THE ABILITY TO <process>
BE ABLE TO <process>
<object>
[<additional details about the object>]

Fig. 1. Rupp's Boilerplate (Taken From [6]).

The Rupp's boilerplate starts with optional condition followed by the system name. This is then followed by the three modal verbs which are shall, should or will. Following modal verbs will be the statement of process which able to come directly after the modal verb without further keyword or and information. The final component is an object with optional additional details about the project.

Derived from Rupp's boilerplate idea, we are motivated to develop a tool-based boilerplate that can improve SRS qualities. By using Rupp's guideline, we develop a prototype of tool-based boilerplate and use it to produce an SRS for information management system. This prototype is based on IEEE template [7] and only covers Section 3.2 of SRS which describes functional requirements. We applied the three modal verbs which are SHALL, SHOULD and WILL to indicate the importance of the requirements. Shall is used to indicate a mandatory requirement. Should is used for non-mandatory requirement and Will is used to indicate a statement which is not subjected to verification.

The aim for this tool is to guide the requirements engineer to write the SRS with consistency. This research scopes the tool prototype to cover Section 3.2 of software requirement specification (SRS) only which covers the elaboration of functional requirements. The tool-based boilerplate is designed for a generic information system management which usually has similar essential requirements. The essential requirements or also known as basic functionalities cover registration, login, searching, adding information and maintaining information with edit, delete and update capabilities. These basic functionalities can be utilized by many information system management and further elaborated to make a complete system. Besides, in order to increase the correctness quality of the SRS, a wireframe interface is auto generated to portray the requirements statements stated in the SRS.

III. Empirical Investigation Protocol

This section discusses empirical investigation method. An empirical investigation is conducted to evaluate the usability of the tool-based boilerplate from the perspective of novice user.

A. *Identifying Participants*

The participants for the empirical investigation were identified among senior undergraduate students enrolling Computer Science Degree at Universiti Teknikal Malaysia Melaka. Seventy third year computer science students majoring in system development from Software Engineering Department were carefully selected. However, only sixty-three of them managed to showed up during the evaluation. The participants were purposely selected based on their background knowledge and familiarity with requirements. They took subjects related to software engineering namely Software Engineering (BITP 2213), Software Requirements Engineering (BITP 2233), Software Architecture and Design (BITP 3243) and Software Verification and Validation (BITP 3253). They were also trained to produce software development documents such as Software Development Plan (SDP), Software Requirements Specification (SRS), Software Design Document (SDD) and Software Test Description (STD).

B. *Instruments*

In order to deploy the investigation, the participants were provided with a case study, a tool-based boilerplate technique prototype and a questionnaire. The prototype was developed for the usage of producing a Software Requirement Specification (SRS) for a generic Information Management System. Basic functions of an information management system are mainly the same and can simply adapted to another system. In this research, a library system is referred to as a study case. The feedback is obtained by a questionnaire. The questionnaire is designed to measure the participants' opinion in order to know how far the evaluation meets the objectives. The qualities we measured in this investigation are Usability, Usefulness and Ease of Use. These can further be broken into sub-attributes. We select this three quality attributes because these three qualities are the main qualities that lead to a good SRS [8]. Under the usability attributes, the quality we measured is correctness, consistency, learnability, efficiency and simple. We measure these quality based on TAM [9]. Under usefulness, we measure three sub-attributes which is 1) Accomplish task more quickly and correctly, 2) Easier to do job and 3) Enhance the effectiveness of the output. These sub-attributes that we measure, is the quality that contributes to usefulness. Lastly, the ease of use, we measure three sub-attributes which is 1) Find it easy to use the system, 2) Interaction with the system is clear and understandable and 3) I would find the system easy to use. Under ease of use, we measure the experience that participants gain from using this tool.

The definition of the quality and the sub-quality attributes are as listed below:

- Usability test stated by [10] is measured based on user perception of the tool. In order to evaluate the usability, there are five quality attributes which are Correctness, Consistency, Learnability, Efficiency and Simple.
- Correctness [11] is defined when every requirement stated in the SRS is achieved.
- Consistency [12] is defined as a set of requirements contain no internal contradictions.
- Learnability [13] is defined as the ease and speed with which the users get familiar with the use of a new product.
- Efficiency [13] is defined as the ability to do or produce something without wasting materials, time or energy.
- Simple [14] is defined as not hard to understand, not complex or fancy.
- The usefulness is assessed by the tool ability to accomplish the task more quickly and correctly, easier to do the job with and whether the tool can enhance the effectiveness of the output [13].

- The ease of use [14] is assessed by the interaction of participants with the system is clear or not.

C. The Protocol

The empirical investigation started with preparing the environment for the investigation. The investigation was allocated during lab session and the time was limited within 2 hours. A briefing on the research purpose and explanation on the tool-based boilerplate was made to the participants. A demonstration of the tool-based boilerplate was given too. Then, 15 minutes were given to the participants to read the case study. The case study represented an input to the requirements engineering process as if information gathered earlier in order to write a requirement specification. It basically describes stakeholders' needs for the system to be developed. Then, the tool-based boilerplate is given to the participants to write requirements specification based on the case study given. Once done, the participants were required to provide feedback based on the questionnaire. For each question, the participants need to score each attribute based on the 5 points Likert scale whereby 1 is strongly disagree and 5 is strongly agree. The entire process will take no longer than 2 hours as the software project scope is small.

IV. Empirical Investigation Results

All sixty-three feedbacks were gathered and analysed. This empirical investigation was conducted with aims to get a feedback for the usability, usefulness and ease of use of the tool-based boilerplate technique. In this section, we will present the evaluation results in descriptive and inferential analysis.

A. Results

Based on Fig. 2, the results show positive results with respect to the usability of the tool. The result shows that most of the participants agree that the tool is in high usability (49.8%) and only few participants who disagree (5.2%).

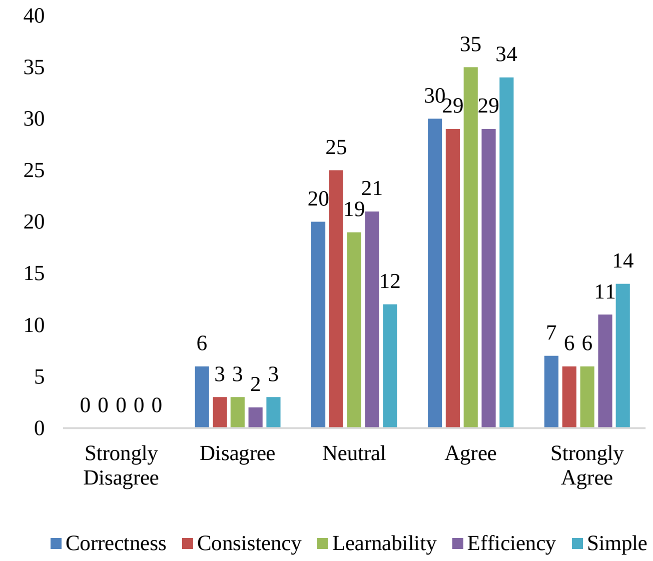


Fig. 2. Results for Usability.

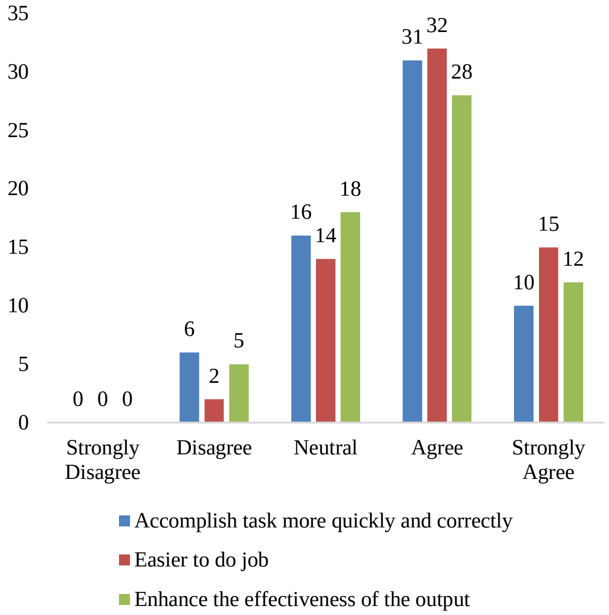


Fig. 3. Results for Usefulness.

Based on Fig. 3, most of the participants agreed that the tool is useful (48%) based on three attributes which are accomplish task more quickly and correctly, easier to do job and enhance the effectiveness of the output. This result is promising as most of participants gave positive feedbacks.

Fig. 4 shows the evaluation results for Ease of Use. The highest percentage among those five scale is 'Agree' (45.5%). Only one participant strongly disagrees that the interaction with the system is clear and understandable. In summary, the evaluation conducted showed that the tool improves requirement qualities and also can help requirement engineer to capture the requirements with ease.

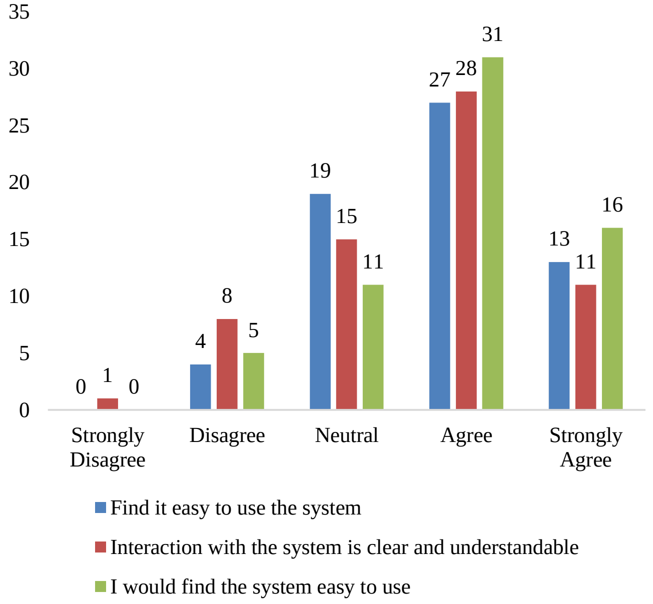


Fig. 4. Results for Ease of Use.

V. Analysis

This section presents further analysis on the results derived from the empirical investigation. In addition, results from second evaluation based on expert judgment is included here to show the correlation among the qualities achieved from both evaluation methods. The protocol and details on the expert judgment evaluation can be referred in [15].

A. *Descriptive Statistic*

We used descriptive statistic to summarize the basic feature data. Measurement includes the mean, median and standard deviation.

Referring to Tables I and II, take a look at mean row. We can see that the mean values for both correctness and consistency from experts' judgements are 4.000 and we can compare that mean values for correctness and consistency from empirical investigation are both 3.603. The mean values show that there is not much difference in correctness and consistency qualities achieved in both evaluations. In empirical investigation, the attributes learnability and simple can be translated into comprehensibility in expert judgment. Table II shows mean values of 3.698 and 3.937 for learnability and simple respectively which has insignificant difference with mean value of comprehensibility attribute (3.900) presented in Table II. The results derived from two evaluations with two different methods were supporting each other and shows that the tool-based boilerplate technique improves SRS quality.

TABLE I. Descriptive Statistic of Quality Attributes Derived from Empirical Investigation

	Empirical Investigation			
	Correct-ness	**Consistency**	**Learn-ability**	**Simple**
N	63	63	63	63
Mean	3.603	3.603	3.698	3.937
Median	4.000	4.000	4.000	4.000
Std. Dev	0.8140	0.8140	0.8140	0.7803

TABLE II. Descriptive Statistic of Quality Attributes Derived from Expert Judgement

	Expert Judgement		
	Correct-ness	**Consistency**	**Comprehen-sibility**
N	10	10	10
Mean	4.000	4.000	3.900
Median	4.000	4.000	4.000
Std. Dev	0.6667	0.4714	0.3162

B. *Inferential Statistic*

Inferential statistic is prepared to show the relationship between quality attributes derived from both empirical investigation and expert judgement. We run a Pearson's Correlation to measure the relationship of the quality attributes and to indicate if the relationship is either strong or weak.

In this analysis, we were using nominal and ratio for the measurement scale and using bivariate correlation for the scale of measurement. We were using bivariate correlation because it is the suitable statistical test in comparing two sample group and can measure the strength and direction of linear relationship between the two variables which are the quality attributes derived from experts' judgement and empirical investigation.

Table III shows that the Pearson's correlation between Expert Judgement Correctness and Empirical Investigation Correctness is 0.645. This means that there is a strong relationship between results derived from Expert Judgement and Empirical Investigation. Next, we take a look at Sig. (2-Tailed). "Sig." stands for significance level. The value explains if there is a statistically significant correlation between the two variables. In our table, Sig. (2-tailed) value is 0.044. That means, there is a significant difference between results derived from Expert Judgement and Empirical Investigation.

Table IV shows Pearson's Correlation between Expert Judgement Consistency and Empirical Investigation Consistency. We can see that Pearson's r value is 0.488. From the result, we can conclude there is a relationship between Expert Judgement and Empirical Investigation but the relationship is weak. This relationship is not as strong as Correctness. The Sig. (2-tailed) value is 0.153 which is larger than 0.05. That means, there is no significant difference between results derived from Expert Judgement and Empirical Investigation.

TABLE III. Pearson's Correlation between Expert Judgement Correctness and Empirical Investigation Correctness

		Correctness EJ	**Correctness EI**
Correctness EJ	Pearson Correlation	1	.645*
	Sig.(2-tailed)		.044
	N	10	10
Correctness EI	Pearson Correlation	.645*	1
	Sig.(2-tailed	.044	
	N	10	63

TABLE IV. Pearson's Correlation between Expert Judgement Consistency and Empirical Investigation Consistency

		Consistency EJ	**Consistency EI**
Consistency EJ	Pearson Correlation	1	.488*
	Sig.(2-tailed		.153
	N	10	10
Consistency EI	Pearson Correlation	.488*	1
	Sig.(2-tailed)	.153	
	N	10	63

TABLE V. PEARSON'S CORRELATION BETWEEN EXPERT JUDGEMENT COMPREHENSIBILITY AND EMPIRICAL INVESTIGATION LEARNABILITY

		Comprehensibility EJ	**Learnability EI**
Comprehensibility EJ	Pearson Correlation	1	.509
	Sig.(2- tailed)		. 133
	N	10	10
Learnability EI	Pearson Correlation	.509	1
	Sig.(2- tailed)	.133	
	N	10	63

TABLE VI. PEARSON'S CORRELATION BETWEEN EXPERT JUDGEMENT COMPREHENSIBILITY AND EMPIRICAL INVESTIGATION SIMPLE

		Comprehensibility EJ	Simple EI
Comprehensibility EJ	Pearson Correlation	1	.509
	Sig.(2-tailed)		.133
	N	10	10
Simple EI	Pearson Correlation	.509	1
	Sig.(2-tailed)	.133	
	N	10	63

Table V shows Pearson's Correlation between Expert Judgement Comprehensibility and Empirical Investigation Learnability. We compare comprehensibility with learnability because they carry the same definition and purpose. The results show the Pearson's r value is 0.506. As the Pearson's r value is close to 1, we can conclude that there is a relationship between comprehensibility and learnability. The Sig. (2 – tailed) value is large than 0.05 which is 0.133 shows that there is no significant difference between results derived from Expert Judgement and Empirical Investigation.

Lastly, Table VI shows Pearson's Correlation between Expert Judgement Comprehensibility and Empirical Investigation Simple. Result shows that learnability (Table V) and simple share the same results. This means that there is a relationship between comprehensibility and simple but there is no statistically significant correlation between results derived from Expert Judgement and Empirical Investigation.

The insignificant difference in the relationship between the quality attributes derived from the two evaluation methods showed that the results are supporting each other to confirm that the tool-based boilerplate technique indeed improve the quality of SRS.

VI. CONCLUSION

This paper presents an empirical investigation to evaluate the usability, usefulness and ease of use of a tool-based boilerplate in order to improve the SRS quality. The art behind the boilerplate technique is explained and the protocol for the empirical investigation is carefully presented. The empirical investigation results showed that the tool-based boilerplate technique does improve the SRS quality in terms of the usability, usefulness and ease of use. In addition, this paper also elaborate on further analysis to compare the results derived from the empirical investigation and another evaluation method which is an expert judgment. The Pearson r value shows that there are relationships between quality attributes achieved from the two evaluation methods. The insignificant correlation between results derived from Expert Judgement and Empirical Investigation show that the results from both methods are confirming each other.

ACKNOWLEDGMENT

The research and the publication of this paper are funded by Universiti Teknikal Malaysia Melaka through a research grant PJP/2017/FTMK-CACT/S01573.

REFERENCES

[1] S. Ahmad, I.E.A. Jalil, S.S.S. Ahmad, "An Enhancement of Software Requirements Negotiation with Rule-based Reasoning: A Conceptual Model," Journal of Telecommunication, Electronic and Computer Engineering, vol. 8, no.10, pp. 193-198, 2016.

[2] M. Ortel, M. Malot, A. Baumgart, J.S. Becker, R. Bogusch, S. Farfeleder, N. Gerber, O. Haugen, S. Häusler, B. Josko, J. Mansell, "Requirements Engineering" in CESAR-Cost-efficient Methods and Processes for Safety-relevant Embedded Systems, A. Ajitha, T. Wahl, eds., Springer, pp. 69-143, 2013.

[3] M, Frank Luther, "*A History of American Magazines", v.4*. Cambridge (MA): The Belknap Press of the Harvard University Press. pp. 53–54, 1957.

[4] S. Farfeleder, T. Moser, A. Krall, H. Zojer, and C. Panis, "DODT: Increasing Requirements Formalism using Domain Ontologies for Improved Embedded Systems Development,", 2011.

[5] C. Arora, M. Sabetzadeh, L. C. Briand, F. Zimmer and R. Gnaga, "RUBRIC: A flexible tool for automated checking of conformance to requirement boilerplates." In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pp. 599-602. ACM, 2013.

[6] K. Pohl and C. Rupp, "Requirements Engineering Fundamentals", 1st ed. Rocky Nook, 2011.

[7] IEEE Recommended Practice for Software Requirements Specifications. IEEE Standard 830-1998. IEEE Computer Society, 1998.

[8] D. Zowghi, & V. Gervasi, "The Three Cs of Requirements: Consistency, Completeness, and Correctness," Proceedings of 8th International Workshop on Requirements Engineering: Foundation for Software Quality, pp. 155–164, 2002.

[9] F. D. Davis, R. P. Bogozzi, & P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models. Management Science", 35, 982-1003, 1989.

[10] N. Bevan, & M. Macleod, "Usability measurement in context. *Behaviour and Information Technology", 13*(1–2), 132–145, 1994.

[11] A. Davis, "Software Requirements: Objects", Function and States (Second Edition), Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[12] H. Rombach, "Software Specifications: A Framework", Pennsylvania, Pittsburgh:Software Engineering Instituts, January 1990.

[13] B. Boehm, J. R. Brown, J. R Kaspar, M. Lipow, G.J MacLoed & M. J. Merritt, "Characteristics of Software Quality". TRW Series of Software Technology, Amsterdam, 1978.

[14] D. Leffingwell & D. Widrig, "Managing Software Requirements. A Unified Approach", Addison-Wesley, 2000.

[15] S. Ahmad, U. Anuar, N.A. Emran, "A Tool-based Boilerplate Technique to Improve SRS Quality: An Evaluation". Journal of Telecommunication, Electronic and Computer Engineering, vol. 10, no.2-7, pp. 111-114, 2018.

New 3D Objects Retrieval Approach using Multi Agent Systems and Artificial Neural Network

Basma Sirbal[1], Mohcine Bouksim[2], Khadija Arhid[3], Fatima Rafii Zakani[4], Taoufiq Gadi[5]
Laboratory of Informatics, Imaging, and Modelling of Complex Systems (LIIMSC)
Faculty of Sciences and Techniques, Hassan 1st University
Settat, Morocco

***Abstract*—Content-based 3D object retrieval is a substantial research area that has drawn a significant number of scientists in last couple of decades. Due to the rapid advancement of technology, 3D models are more and more accessible yet it is hard to find, the models we are searching for. This created the need for efficient and robust retrieval methods, allowing the extraction of relevant matches from the human perspective. Hence, in this paper we are proposing a new framework for 3D object retrieval that starts with a pre-treatment consisting of an Artificial Neural Network (ANN) algorithm with Histogram of features, allowing us to extract a representative value for each category of the database. These values are used for the Multi Agents System (MAS). In this phase, we are classifying these categories according to their relevance to the request object. This sets a distinguishing weight for each object of the database allowing us to extract the right matches. Experiments have proven the stringent of this approach.**

Keywords—3D object retrieval; 3D image processing; distributed artificial intelligence; multi-agent systems; artificial neural network (ANN)

I. Introduction

The rapid development of computer techniques, 3D sensors and imaging devices has led to the rapid growth of rich information contained 3D models. Hence, they are more accessible in our daily lives. This induces the urgent need for efficient retrieval and recognition technologies. An excellent retrieval algorithm implies that the matches extracted belong to the same category, and are relevant from the human perception. This involves the representation of the 3D model by its geometrical, topological or other properties into a compact descriptor. The process of extracting the right matches requires two main steps:

Offline: also is the indexing phase. Where the proprieties of the models are exploited to represent it, a signature is computed for every object of the database, and stored for further usage. Any pre-treatment needed is executed in this phase.

Online: this second phase necessitate that the retrieval system takes a 3D object as input, then obtain the closest and more relevant matches for this query, its signature is computed using the chosen method, then distances between the signature of the request and those of the objects of the database are calculated and compared.

In this work, we propose a new content-based retrieval framework that exceeds the performance of well-known ones. This approach is composed of three major phases; the first one is a pre-treatment employing an Artificial Neural Network (ANN), followed by a Multi agent system and finally the matches' extraction.

The agent notion has been introduced to ease the development of complex software and bring new solutions for unsolved issues. Still it has not been fully exploited [1]. The multi agent phase is where we are using the results generated by the ANN algorithm to classify the classes of the used database in order to extract values that are going to be used afterwards to refine the results of this retrieval process.

In this paper, we are answering following questions: How can we improve the quality and the relevance of the matches given by existing retrieval frameworks? Is it possible to exploit existing methods to achieve the aim?

This paper is structured as follows: Section 2 briefly reviews the related work and interesting work to mention, followed by the background in Section 3. In Section 4, we describe the proposed proposed approach. Experimental results and analysis are provided in Section 5. Finally, in the last section, conclusion and some perspectives are covered.

II. Related Work

In the past decade, a number of content-based 3D model retrieval techniques have been developed. According to Johan, D. Tangelder et al. [2], these approaches can be sorted into 3 main categories, they also indicated that these categories are fusible since many of them can fit into more than one group. For more details about retrieval methods readers can refer to these surveys [3], [4] Each category of methods has advantages and disadvantages, hence why we decide to exploit methods from the two different categories we are discussing next.

Feature based methods: are based on geometric and topological features, extracted directly from the 3D model. Many scientists dedicated their work to this category. We address some interesting ones. To deal with queries of different modalities, Shah et al. [5] proposed using a different way of representing the 3D Model surface, Keypoints-based Surface Representation (KSR) technique involving the geometrical relationship between the detected 3D keypoints for local surface representation. Bouksim et al. [6] introduced a new approach, the heart of it is a multi-criteria method that generates a compact descriptor, using the Data envelopment analysis method (DEA) [7]. Tabia et al. [8] proposed a 3D shape descriptor based on local CNN features encoded using vectors of locally aggregated descriptors instead of conventional global CNN, using Convolutional Neural Networks (CNN) that includes all the entities from all modalities into a common space. Finally another interesting work is the one of Zeng and all [9], proposed a convolutional neural network based multi-feature fusion learning method for no rigid, using both the heat kernel signature (HKS) descriptor and the wave kernel signature (WKS) descriptor.

Geometry based methods (view based): The perception behind this category of methods is that two 3D models are matching, if they look similar from all viewing angles. This is accomplished by representing the 3D model using representative 2D captures. Many methods were published in the literature, a lot of them uses representations like binary images, projection or depth images. We are listing some of the methods existing in the literature. In their work, Wang et al. [10] introduced a boosting approach, where view's discriminative ability is analysed using the proposed reverse distance metric, then an algorithm introduced by the authors is employed to boost the multi-model graph learning based retrieval method. Another interesting work to mention, Lee and al [11] proposed a feature aggregation method, Cross-View Convolution (CVC), which models a 3D shape as a sequence of rendered views. Then used a Cross-Domain Triplet Neural Network (CDTNN) that incorporates an adaptation layer to match the features from different domains better and can be trained end-to-end. In another hand, Liu and al [12] propose a discriminative multi-view latent variable model (MVLVM) for 3D models retrieval. The MVLVM allows to have an undirected graph structure in which the view set of a given 3D object is treated as the observations from which to discover the latent visual and spatial contexts. Then, they use a learning and inference process of MVLVM for view-based 3D object retrieval.

III. Background

In this section, we are discussing the methods and technologies that supports this research paper. Initially, we are starting with a technology that has proven its strength in many fields, it is Artificial Neural Network (ANN). We are exploiting it for the pre-treatment

A. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) has been used for the purpose of 3D object retrieval for a couple of decades now. Here we are mentioning some interesting works, starting with Qayyum and al [13], where they are using deep convolutional neural network (CNN) that is trained for classification of medical images in order to upgrade existing content based medical image retrieval (CBMIR) systems. Furthermore Zhu and al [14] trained Convolutional Neural Network (CNN), and used the extracted features for 3D object representation. Bouksim and al [15] used a histogram of features extracted directly from the 3D along with an artificial neural network (ANN) algorithm for the training, the results of the hidden layers are then used as a descriptor in the retrieval system. For a detailed description of the technology we invite the readers to visit this e-book realised by Nielsen [16].

Next we are giving a brief introduction of the core technology used for this research, Multi Agents System (MAS).

B. Multi Agent System (MAS)

Multi-agent systems (MAS) have been the interest of more and more authors, since it provides adequate solutions for many complicated issues in several domains. It has been explored in a large number of software related domains as robotics, sustainable energy distribution [17], [18], but also in more general fields such as psychology [19] or biology [20]. The agent paradigm shifts different operative implementations. The more common ones are management agents and simulation agents; for in depth details you can consult the work of [21], [22].

Moreover, we are listing some commonly used methods in the literature that consolidates the proposed work. We adopted three methods that are based around strong mathematical Models.

C. Featured Methods

We experimented with different methods before agreeing on this composition that serves the purpose of refining the results existing in the literature. We are giving an overview of the approaches, we invite the readers to read the papers referred for more details.

DEA: Bouksim et al. [6], provided an approach for retrieval the core of it is a multi-criteria method that generates a compact descriptor, which represents the signature for each 3D model. The main intention behind this approach is to exploit the best out of each criterion (i.e., measure) by extracting a combined score using the Data envelopment analysis method (DEA), also known as frontier analysis introduced by Charnes, Cooper [7], and Rhodes in 1978. It is a linear programming method, which hypothetically measures the efficiency of the decision-making units (or DMUs) when this later present multiple inputs and/or outputs.

PANORAMA: Introduced by Papadakis and Al [23] is a 3D shape descriptor, initially utilizes a set of panoramic views of the 3D object, this allows to describe the position and orientation of the object's surface in 3 dimensional space. They acquire a panoramic view of the 3D object by projecting it to the lateral surface of a cylinder parallel to one of its three principal axes also situated at the centroid of the object. Later the object is projected to three perpendicular cylinders, each one of them is aligned with one of its principal axes in order to capture the global shape of the 3D object. For every projection they calculate the corresponding 2D Discrete Fourier Transform as well as 2D Discrete Wavelet Transform. They further increase the retrieval performance by employing a local (unsupervised) relevance feedback technique that shifts the descriptor of an object closer to its cluster centroid in feature space.

Light Field: Chen et al. [24] provided a visual similarity-based 3D Object retrieval system, it calculates the similarity between 3D objects by visual similarity. The primary idea is that if two 3D Objects are similar, they should look similar from all viewing angles. A hundred orthogonal projections of each object, disregarding symmetry, are coded both using

Zernike moments and Fourier descriptors as features for the retrieval process.

In the following section we are giving a visual overview of this framework. , followed by a detailed description of the proposed approach.

IV. Proposed Approach

Our intention is to get the best matches for a 3D Object request. Most of the existing databases are categorised into different classes. Therefore the request object must fit into one of them. This inspired us to believe that weighting the elements of each class of the database with a favouring value can optimize the retrieval process. In this work we are elaborating a new framework for 3D objects retrieval, composed of two main phases an offline pre-treatment and an online classification using a Multi Agent System (MAS) which will carry out the weighting, at a final step we are extracting the matches with the Panorama method [23].

Before giving more details about this approach, the first figure Fig. 1, represents the architecture of this approach.

As described in Fig. 1, this framework is composed of two main phases. An offline phase where the pre-treatment takes place, allowing to extract representative values for the database using an artificial neural network (ANN) algorithm. Afterwards comes the online phase with two steps. The first step is a classification of the categories of the database with a multi agent system, therefore extracting the favouring weight for every object. The second and last step is using these values to extract the matches using Panorama method [23].

A. The Offline Phase

Instead of using the whole data base we propose in this approach to extract a representative Object for each class of the used database. The first phase of the pre-treatment, consists of extracting a histogram of features directly from the 3D objects. Followed by a training phase using an artificial neural network (ANN) algorithm; this last point helps to train the ANN fast and with consistent data. Once trained it allows us to extract a representative object for each class of the database [15], [16].

Since we are using three methods in the classification phase we are computing the signatures for the representative objects in the database using two of the methods, Light Field [24] and DEA [6]. Then we are computing the signatures for all the database objects using the Panorama [23] since we are going to be using it both in the classification and in the extraction phase.

B. The Online Phase

In this phase the request 3D object is given by the user. How can we extract the best matches for this later?

First we need to order the classes of the database according to their relevance. This will allow us to get more accurate results.

Each retrieval method has advantages and disadvantages, hence why we are using the combination three different methods for an efficient classification. Thanks to the pre-treatment we now get to use only the most representative object for each class. The best way is to parallelize the process, hence why we are implementing a Multi Agent System (MAS). Distributed artificial intelligence in the form of MAS allows the classification part of the framework to be faster and reliable.

This Multi Agent System's architecture is composed of three layers, each one involves agents with different tasks. We are describing it in what follows.

The first layer is composed of 19 agents for each method. Each one of them computes the distance between the signatures of the representative element computed offline using the method and the signature of the request object. This information is then communicated to the elements of the second layer.

The second layer involves three agents, one for each method. In this agent, the distances representing each class are ordered, giving a classification of nineteen classes according to their relevance in comparison to the request objects 'signature. Each of the classifications is transferred to the final layer consisting of one agent.

Since each method has a different way of indexing the 3D objects, this last agent exploits the results sent to it and normalizes the signatures given by the three methods values. Then it takes the averages of these values for a unified classification, these values representing each class are considered as the weights. Each of the objects in the database takes a weight according to the order given to the class they belong to.

Then comes the final part of the process. The same agent uses the signatures computed by the method Panorama from the offline phase, and computes the distances between each one and the request object. Simultaneously the final representative score is computed using this distance and the weights representing each object as follows.

$$s_i = d_i \ e^{w_i} \tag{1}$$

Where Si is the score for an element i, and di is the distance between the element i and the request object, wi is the weight computed based on the classification. The smaller the score the more relevant the match is, hence why we order them. This gives us the matches relevant for the request Object. This method has proven its strength as showed in the next section.

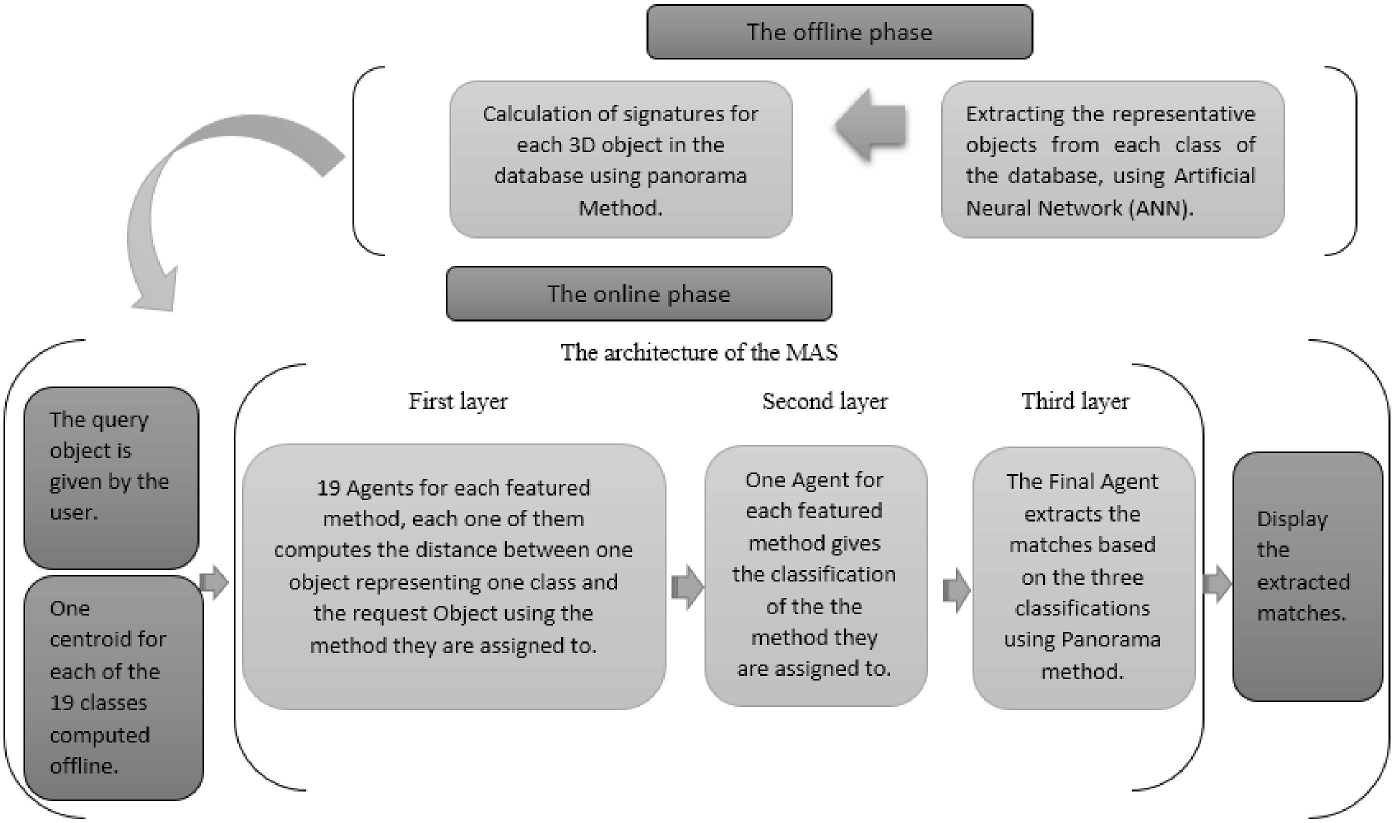

The offline phase
Calculation of signatures for each 3D object in the database using panorama Method.
Extracting the representative objects from each class of the database, using Artificial Neural Network (ANN).
The online phase
The architecture of the MAS
The query object is given by the user.
One centroid for each of the 19 classes computed offline.
First layer
19 Agents for each featured method, each one of them computes the distance between one object representing one class and the request Object using the method they are assigned to.
Second layer
One Agent for each featured method gives the classification of the the method they are assigned to.
Third layer
The Final Agent extracts the matches based on the three classifications using Panorama method.
Display the extracted matches.

Fig. 1. The Architecture of the Proposed Framework.

V. Experimental Results

In this section of this paper we are demonstrating the efficacy and the discriminative capacity of this approach through experimental results. We are comparing the results achieved with those of well-known methods: D2 [25], DEA [6], Harmonics [26], Panorama [23] and Light Field [24].

Foremost, we selected a reliable database to test the method. We choose to use Princeton's segmentation benchmark database [27].The selection of the database have been influenced by many criterions, some of them are the number and the diversity of the models. This database includes 380 3D models portioned into 19 classes (Human, Cup, Glasses, Airplane, Ant, Chair, Octopus, Table, Teddy, Hand, Plier, Fish, Bird, Armadillo, Bust, Mech, Bearing, Vase, and Fourleg).

The first experiment consists in computing the 10 nearest neighbours for each object using different methods, then we record the percentage of those that are right among them, whereas if the result belongs to the same class it is considered as relevant. Finally we obtain the accuracy for each Class by computing the average of all the results for each object of that class. Table 1 presents the results obtained from the proposed approach along with other known methods which are D2 distributions, 3D Harmonics, DEA, Panorama and LightField. We can observe from the results that our method exceeds all other methods, let’s take for example the class Glasses, our approach obtained 98.5% of correct results, which is the highest of all results. Whereas for the class Ant our approach obtained 100% correct matches, same results were obtained for the class Teddy. We can also observe that our approach got the double of the results obtained by the other methods for the class Vase. Overall our method surpasses all the other methods except for the class ARMADILLO where two methods slightly surpasses ours, however our method extracted correctly all 7 neighbours for this class as illustrated in the third figure, Fig. 3.

We are carrying out with a commonly used test, precision-recall diagrams. Recall is the ratio of relevant to the query retrieved models to the total number of relevant models while precision is the ratio of relevant to the query retrieved models to the number of retrieved models. The evaluations were performed by using each model of a dataset as a query on the remaining set of models and computing the average precision-recall performance overall models. , all

$$Recall = \frac{relevant\,correctly\,retrieved}{all\,relevant}, \tag{2}$$

$$Precision = \frac{relevant\,correctly\,retrieved}{all\,retrieved}. \tag{3}$$

The second figure, Fig. 2 illustrates the precision-recall graphs obtained for the proposed method along with DEA [6], Harmonics [26], Panorama [23] and Light Field [24].

We can clearly observe from the curves that the proposed approach surpasses all the other methods, this shows the capacity of this approach.

TABLE I. THE PERFORMANCE OF EACH OF THE METHODS BASED ON THE EXTRACTION OF 10 NEAREST NEIGHBOURS

	Panorama	**Light Field**	**DEA**	**Harmonics**	**D2**	**The proposed approach**
Human	79	56,5	52,5	41	47,5	**100**
CUP	87,5	72,5	56,5	76,5	46	**96**
Glasses	93,5	82	61,5	87	83	**98.5**
Airplane	93	83,5	50,5	79	55	**95**
ANT	97,5	79	98	54,5	42,5	**100**
CHAIR	98,5	97	52,5	94	66	**100**
OCTOPUS	57,5	70,5	54	41	20,5	**95**
TABLE	93	57,5	75,5	45,5	42	**98**
TEDDY	100	95,5	92,5	93	63	**100**
HAND	78,5	37,5	52,5	32,5	29	**94.5**
PLIER	92	94	93,5	63,5	69,5	**100**
FISH	96	82	79,5	76,5	50	**100**
BIRD	68,5	37,5	46,5	42,5	36,5	**72**
ARMADILLO	**95**	63,5	93,5	54	26,5	89
BUST	84,5	60	62	51	32,5	**96**
MECH	77,5	85,5	79,5	82	53,5	**96.5**
BEARING	81,5	54,5	33	40	22	**89**
VASE	39,5	31,5	31,5	20,5	17,5	**68.5**
FOURLEG	89,5	85,5	30,5	73,5	41,5	**95.5**

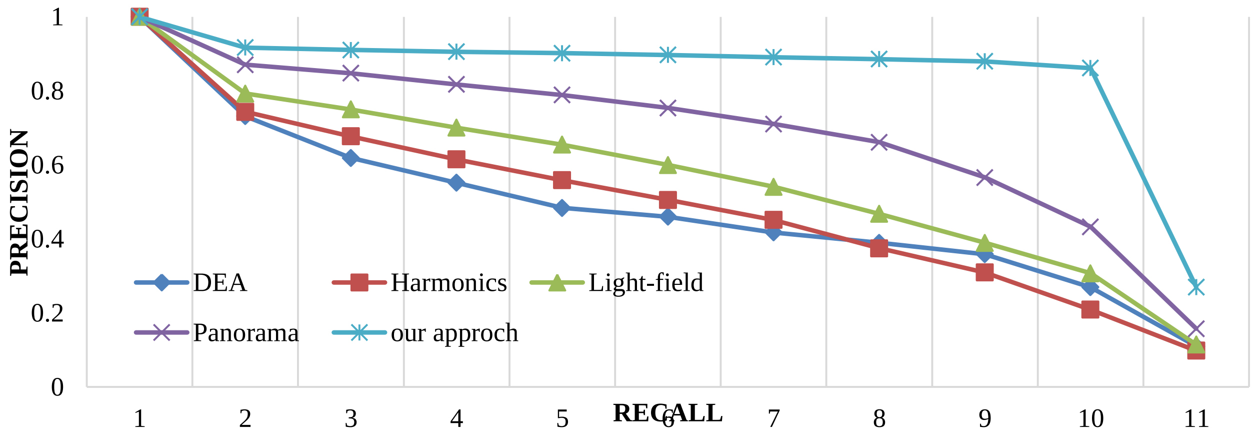


Fig. 2. Precision-Recall Graph using Four Different Descriptors with the Proposed One.

The third test will qualify the proposed method by computing some evaluation metrics, which are:

- Average Precision (AP): It is used to represent the precision performance of an Information Retrieval (IR) method over all relevant items. It is the average of precision values at each ranking position where a relevant item has been retrieved. For example, five relevant items are located at the following ranking positions: 1st, 2nd, 4th, 7th and 10th. Let the precision values at each one of these ranking position are: 1, 1, 0.75, 0.57 and 0.5. Then, AP is the mean of these values (0.76).

- Average Dynamic Recall (ADR): The scalar is used to express the recall performance of an IR method at a given set of ranking positions. It is defined as:

$$ADR = \frac{1}{R}\sum_{i=1}^{R}\frac{R1(i)}{i} \qquad (4)$$

Where R indicates the lower ranking position to be included in the calculation (e.g. 20 first ranking positions), RI (i) represents the number of relevant retrieved items within the first i retrieved items.

- First Tier (FT) and Second Tier (ST): computes the recall for the top C−1 and 2*(C−1) correctly retrieved objects in the result list, where C represents the number of item in each class.

- Discounted Cumulative Gain (DCG): a scalar that focuses on the items that are correctly retrieved and are in the front of the results list, since generally, a low ranking position has a low probability to be discovered by the user.

- F-Measure: The F-Measure simply generates a measure that combines the recall and precision values to express the overall performance of the retrieval system. It is computed as follow:

$$FMeasure = 2 \times \frac{\Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}. \quad (5)$$

You can observe from table 2 that the proposed method obtained the highest values, followed by Panorama[23] then comes respectively Light field [24], Harmonics [26] and finally comes DEA[6]. This is just a confirmation of the results achieved in the previous experiments. Overall, this confirms that our method surpasses all others.

Finally, we select a test that will illustrate the results visually. That is the extraction of the 7 nearest neighbours, for 7 different objects and this using two methods previously used along with our suggested method. Our final figure, Fig. 3 represents the results as follows: the proposed approach (bottom centre), DEA (top left) and Panorama (top right); the illustrations in the left column shows the query models (we choose 7 models randomly among the data base), while the columns on the right displays the closest matches within the used database.

From the visual results we can easily see that the new approach succeed to provide refined matches. Let's take the class human for instance, the methods we are comparing ours to be both giving neighbours that belongs to other classes, for example the class Airplane, Octopus and Armadillo when all the neighbours given by the proposed approach are from the class human. Another example is the Model Octopus where our results belong all to the correct class, whereas the matches given by the other methods contain models from other classes, Fourleg and Glasses. Overall our approach proves once more its strength.

TABLE II. THE SCALAR METRICS FOR EACH METHOD

Descriptors / Scalar Metrics	**AP**	**ADR**	**FT**	**ST**	**DCG**	**F-Measure**
DEA	0.45	0.14	0.40	0.28	5.98	0.27
Harmonics	0.47	0.15	0.48	0.32	5.95	0.30
LightField	0.54	0.16	0.55	0.35	6.21	0.32
PANORAMA	0.66	0.18	0.69	0.41	6.62	0.35
Proposed approach	**0.81**	**0.20**	**0.88**	**0.49**	**7.03**	**0.36**

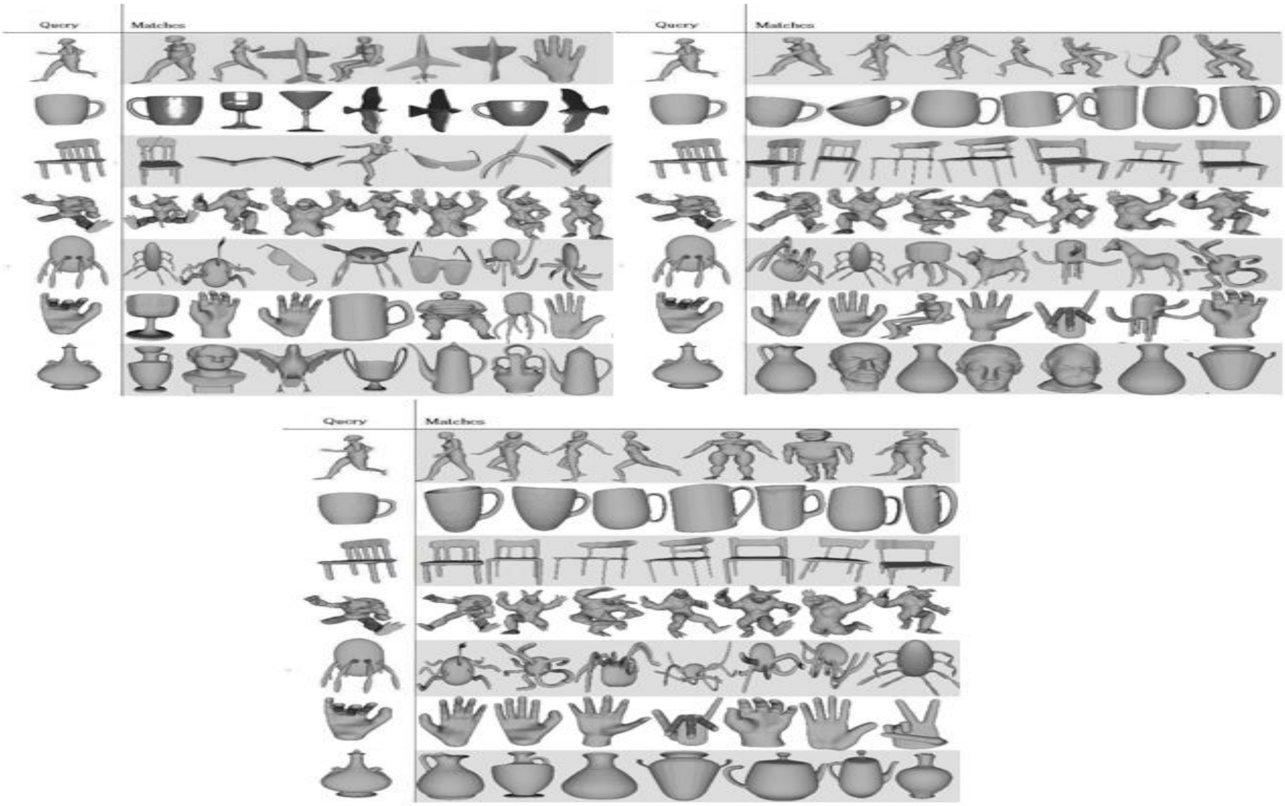


Fig. 3. Top 7 retrieved 3D models using DEA[6] (top left), Panorama[23] (top right) and the proposed approach (bottom centre).

VI. Conclusion

To sum up the most important aspects of our work, it introduces a new framework that utilizes Artificial intelligence to refine the matches retrieved for 3D objects. The process starts offline with a pre-treatment, employing an Artificial Neural Network algorithm, providing a representative value for each class of the used database followed by a Multi Agent System allowing us to classify therefore give a differentiating weight to each object of the database. Finally, we are using these values to extract matches for the request Object. Our method proves its potency in each one of the experiment it has been through. Overall, the proposed method surpasses some well-known methods and gives very satisfactory results. For our future work, we are experimenting with ways to use the same framework for 3D partial matching.

References

[1] G. Jezic, Y.-H. J. Chen-Burger, R. J. Howlett, L. C. Jain, L. Vlacic, and R. Šperka, *Agents and Multi-Agent Systems: Technologies and Applications 2018: Proceedings of the 12th International Conference on Agents and Multi-Agent Systems: Technologies and Applications (KES-AMSTA-18)*. Springer, 2018.

[2] J. W. H. Tangelder and R. C. Veltkamp, 'A survey of content based 3D shape retrieval methods', *Multimed. Tools Appl.*, vol. 39, no. 3, pp. 441–471, Sep. 2008.

[3] K. Srinivasa Reddy, A. R, K. Kalaivani, and P. Swaminathan, 'A comprehensive survey on Content Based Image Retrieval system and its application in medical domain', *Int. J. Eng. Technol.*, vol. 7, pp. 181–185, Jan. 2018.

[4] G. Lara López, A. Peña Pérez Negrón, A. De Antonio Jiménez, J. Ramírez Rodríguez, and R. Imbert Paredes, 'Comparative analysis of shape descriptors for 3D objects', *Multimed. Tools Appl.*, vol. 76, no. 5, pp. 6993–7040, Mar. 2017.

[5] S. A. A. Shah, M. Bennamoun, and F. Boussaid, 'Keypoints-based surface representation for 3D modeling and 3D object recognition', *Pattern Recognit.*, vol. 64, pp. 29–38, Apr. 2017.

[6] Hassan 1st University *et al.*, 'New Approach for 3D Mesh Retrieval Using Data Envelopment Analysis', *Int. J. Intell. Eng. Syst.*, vol. 11, no. 1, pp. 1–10, Feb. 2018.

[7] A. Charnes, W. W. Cooper, and E. Rhodes, 'Measuring the efficiency of decision making units', *European Journal of Operational Research*, p. Vol 2 429–444, 1978.

[8] H. Tabia and H. Laga, 'Learning shape retrieval from different modalities', *Neurocomputing*, vol. 253, pp. 24–33, Aug. 2017.

[9] H. Zeng, Y. Liu, S. Li, J. Che, and X. Wang, 'Convolutional Neural Network Based Multi-feature Fusion for Non-rigid 3D Model Retrieval', p. 15, 2018.

[10] D. Wang, B. Wang, S. Zhao, H. Yao, and H. liu, 'View-based 3D object retrieval with discriminative views', *Neurocomputing*, vol. 252, pp. 58–66, Aug. 2017.

[11] T. Lee, Y.-L. Lin, H. Chiang, M.-W. Chiu, W. Hsu, and P. Huang, 'Cross-Domain Image-Based 3D Shape Retrieval by View Sequence Learning', in *2018 International Conference on 3D Vision (3DV)*, Verona, pp. 258–266, 2018.

[12] A.-A. Liu, W.-Z. Nie, and Y.-T. Su, '3D Object Retrieval Based on Multi-View Latent Variable Model', *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2018.

[13] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, 'Medical image retrieval using deep convolutional neural network', *Neurocomputing*, vol. 266, pp. 8–20, Nov. 2017.

[14] Z. Zhu, C. Rao, S. Bai, and L. J. Latecki, 'Training convolutional neural network from multi-domain contour images for 3D shape retrieval', *Pattern Recognit. Lett.*, Sep. 2017.

[15] M. Bouksim, K. Arhid, F. R. Zakani, M. Aboulfatah, and T. Gadi, 'New Approach for 3D Mesh Retrieval Using Artificial Neural Network and Histogram of Features', p. 11, 2018.

[16] M. Nielsen, 'Neural Networks and Deep Learning', p. 224, 2015.

[17] S. Howell, Y. Rezgui, J.-L. Hippolyte, B. Jayan, and H. Li, 'towards the next generation of smart grids: Semantic and holonic multi-agent management of distributed energy resources', *Renew. Sustain. Energy Rev.*, vol. 77, pp. 193–214, Sep. 2017.

[18] A. K. Mbodji, M. L. Ndiaye, and P. A. Ndiaye, 'Decentralized control of the hybrid electrical system consumption: A multi-agent approach', *Renew. Sustain. Energy Rev.*, vol. 59, pp. 972–978, Jun. 2016.

[19] G. Carslaw, 'Agent based modelling in social psychology', p. 267.

[20] R. Bardini, G. Politano, A. Benso, and S. Di Carlo, 'Multi-level and hybrid modelling approaches for systems biology', *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 396–402, 2017.

[21] D. Kremmydas, I. N. Athanasiadis, and S. Rozakis, 'A review of Agent Based Modeling for agricultural policy evaluation', *Agric. Syst.*, vol. 164, pp. 95–106, Jul. 2018.

[22] S. V. Albrecht and P. Stone, 'Autonomous agents modelling other agents: A comprehensive survey and open problems', *Artif. Intell.* vol. 258, pp. 66–95, May 2018.

[23] P. Papadakis, I. Pratikakis, T. Theoharis, and S. Perantonis, 'PANORAMA: A 3D Shape Descriptor Based on Panoramic Views for Unsupervised 3D Object Retrieval', *Int. J. Comput. Vis.*, vol. 89, no. 2–3, pp. 177–192, Sep. 2010.

[24] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, 'On Visual Similarity Based 3D Model Retrieval', *Comput. Graph. Forum*, vol. 22, no. 3, pp. 223–232, Sep. 2003.

[25] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, 'Shape distributions', *ACM Trans. Graph.*, vol. 21, no. 4, pp. 807–832, Oct. 2002.

[26] T. Funkhouser *et al.*, 'A search engine for 3D models', *ACM Trans. Graph.*, vol. 22, no. 1, pp. 83–105, Jan. 2003.

[27] X. Chen, A. Golovinskiy, and T. Funkhouser, 'A Benchmark for 3D Mesh Segmentation', p. 12, 2009.

Multi-Attributes Web Objects Classification based on Class-Attribute Relation Patterns Learning Approach

Sridhar Mourya[1]
Department of CSE
JNTUH - Hyderabad, India.

Dr. P.V.S. Srinivas[2]
Department of CSE
SNIST - Hyderabad, India.

Dr. M. Seetha[3]
Department of CSE
GNITS - Hyderabad, India.

***Abstract*—The amount of Web data increases with the proliferation of a variety of Web objects, primarily in the form of text, images, video, and music data files. Each of these published objects has some properties that support defining its class properties. Because of their diversity, using these attributes to learn and generate patterns for precise classification is very complicated. Even learning a set of attributes that clearly categorize the categories is very important. Existing attribute learning methods only learn attributes that are closely related to multiple similar objects, but if similar class objects have different attributes, this problem is difficult to learn and classify them. In this paper, a Multi-attributes Web Objects Classification (MA-WOC) based on Class-attribute Relation Patterns Learning Approach is being proposed, which generates a class-attribute with its multi relations patterns. The MA-WOC calculates the relationship probabilities of the attributes and the associated values of the class to understand the degree of association of the construction of classification pattern. To evaluate the effectiveness of the classifier, this will compare to an existing classifier that supports a multi-attribute data set, which shows improvisation of precision with a significant minimum Hamming loss. To evaluate the effectiveness of MA-WOC classification a comparison among the classifiers that are supported to the multi-attribute dataset are being performed to measure the accuracy and hamming loss.**

Keywords*—Classification; multi-attributes; web objects; attribute learning; distinct-class relation

I. Introduction

Web data has become a collection of heterogeneous objects for defining information, for example a particular news topic is being available in text, video and images objects form. Each of these objects can be recognized with the multi-attributes values, and also this classification can be further sub-categorized into various sectors of information such as, "politics", "sports", "education", "entertainments", etc., providing more additional attributes for each individuals [1], [2], [3], [4]. As these information content have multiple attributes, but these can be associated with these multiple classes attributes to recognize them as a distinct class. However, data processing and learning of such multi-attributes data and classifying them is a major challenge for the information providing applications [5], [6], [7]. Many solutions are suggested in the past [1-16] to learn multi-value association and categorize through learning their attributes values. But due to the diversification of these attributes the classification results are not so accurate. This classification methods are needed an accurate learning and association method to provide high classification accuracy.

In the existing classification techniques are majorly assumes that the collection of attributes pattern and being link to one class representation. A dataset collection related to academic information can be classified according to their context for different class attributes in views of the researcher, students or publishers, but these information attributes may have a multiple common or diverse relational attributes network. Thus, network data can have multiple attributes in form of a "vector-based" and multiple attribute relational in form of a "graph-based" illustrations. Moreover, in complex social network data sets such as "Twitter", "Face book", and "LinkedIn" are also typically associated with in excess of one attributes. Here, the attributes are needed to be classified in terms of user interests to the posted text data through a multi-attribute classification [8], [10].

This multi-attribute classification is a challenging task in learning and data mining research, it required to learn the effective relation among the attribute to build an efficient classifier for such heterogeneous datasets. Most of the studies in literature have to make classifier through attribute selection [13], [14] and the association classification Classifiers usually predict data object classes derived from a set of training information. However, the attributes of the classifier configuration is not sufficiently investigated to influence the attributes value of the predicted class, or even in the literature, the problem has not been explored to its extend.

The selection and reduction of feature methods [25] has been used for multi-value classification in the previous proposals. The majority of these recommendations are analyzed by mutual analysis and reduce the features that do not provide critical information for predictive classes. Improvement support can be used to train and organize these lesser or more selective features. But the complexity of the object with multiple values converts the structure of what is appropriate and what is appropriate for the classification. Although some selective methods work well for some classifier [8], [20], but multi-value learning and associating each of the value to its class may not be encouraging to their features. For example, a document with the collection of words term may have some object categories related to "entertainment", "politics", "sports", "economics", and so on, is highly complex to classify to particular class.

Classification is one of the well-liked approaches to associate both the attribute properties and relationship information. This includes node classification techniques that collects the properties of the model properties collectively and indicates the properties of the relevant equipment and the attributes of the properties and combines relationship-based classification in the existing machine learning and repetition process. Researchers have proposed a semi-supervised grouping technique to be partially related to classified networks [10], [16]. The techniques in this group all mean that data points have simply one attribute and one category representation. However, many actual data sets contain further information that can be used to advance performance.

This paper utilizes the additional information and proposes a Multi-attributes Web Objects Classification (MA-WOC) based on Class-attribute Relation Patterns Learning Approach which will be generating a class-attribute with multi its relations patterns. The method of associations of class and attributes are based on a relation probability of attribute and class association value which measure the discreteness of an attributes relation to their comparing class. The proposal is emphasizing to find a discrete class of the web objects having multiple attributes. The MA-WOC addresses the challenges through construction of the multi-attribute learning technique to solve multi-attribute web objects classification.

The following paper is organized as follows. Section-2 describes the related works performed in related to multi-attributes classifications, Section-3 discussed the proposed multi-attributes web objects classification which describes the problem multi-attribute relation learning and class-attribute relation patterns learning approach, Section-4 presents the datasets description, measures and result analysis. Section-5, conclude the conclusion of the paper.

II. Related Works

The exact classification of data is focused on making a deeper analysis of data to provide the necessary information [5], [9], [11], [12]. The classification object is classified by classifier test function and shows a learning set for trained classes [14], [15] For example, a data set containing a collection of records, and every record event has a set of attributes properties that are considered as the set of identity class categories. A classifier performs the classification of the data objects based on the established class knowledge classifier. The purpose of the classifications is to create the perfect classifier, which will provide accurate support for anonymous data classification required for real time. Supervised learning is successfully used in many learning activities to identify relevant objects. A traditional learning system not able to associate appropriately to its class due its complex multiple attributes.

X. Kong et al. [10] apply a grouping technique to handle multi-attribute classification of a single attribute's single relational network data. Converts multiple attribute problems to multiple binary-related issues for each property and captures complex attribute correlations that can exist between properties within the identical instance and correlated instances, and stacking the properties of the similar instance and related instance with the feature set.

F. Charte et al. [9] presents a multi-valued classification scheme for handling multiple data value objects. The proposal is to solve the traditional problem of high-dimensional data classification may suitable for large number of data attributes. The selection of feature selection on the basis of the data transformation and the assessment of association rules transformed based on the attributes dependence. The attribute value identifies the selection feature of the classification algorithm with multiple values. This approach can be successful for linear changes in data objects to indicate the value of addiction, but the results can be inaccurate for highly distributed data in multi-attribute data objects.

The iterative annotation of the "Multiple Relational Social Network (IMR)" [16] is a multiple attribute grouping technique for single attribute multiple relational network data. Multi-attribute problems are treated as multiple-binary related problems by learning the multi-attribute problem from multiple relationships for each attribute classifier to classifiers for each attribute of the feature set that is stacked with the attribute information of the related instance, and this technique does not capture attribute correlations [10].

X. Shi et al. [17] proposed a heterogeneous learning technique as a single attribute classification technique for multi-attribute multi-relational network data. The technique is a bug-driven model that constructs a function on each property view and tries to use two constraints to globally reduce the empirical error function, "consensus across various attribute sources" and "connected instances should have similar prediction".

M. L. Zhang et al. [2] aims at the problem of multi-attribute learning in feature selection. The author utilizes a strategy to learn property-specific features for different class property differentiation. The proposed algorithm name for multi-property learning, LIFT, implements clustering analysis for positive and spoken instances to construct clusters based on attribute by attribute. The classification knowledge base for training and testing is queried in the clustered functional group results. However, while the proposed approach shows promising directions in multi-attribute learning for classification, the importance of feature association for other features should be explored for further optimization.

Multi-attribute learning techniques learn models for each available observation of data and minimize discrepancies between different attributes of non-contributed data. Co-training [18], [19] is a multi-property semi-supervised learning algorithm that learns models for each observation of data and exchanges specific predictions to make use of complementary information available in various properties. However, the multi-attribute learning method does not model network data with relational features.

The "Multi-attribute learning techniques" observe the learning model for each available data and minimize inconsistencies between multiple attributes on unattributed data. The "Co-training methods" [18], [19] is a multi-attribute semi-supervised learning algorithm that learns models and exchanges some predictions each time they observe data, thereby utilizing the additional information available between different attributes. However, multi-attribute learning methods do not use relational features to model network data.

Based on the above reviews and approaches, this shows the importance of multi attribute in classified areas. These areas represent the importance of attribute selection in the accuracy classification. But learning the most characteristic features for categorization is a challenging issue. To overcome these issue and limitation this will propose a new approach to classify the Multi-attributes Web Objects Classification (MA-WOC) based on Class-attribute Relation Patterns Learning Approach. The MA-WOC calculates the probability of a relationship of attribute and class associations for each feature to learn the extent of the organization to create classification patterns. The details of the proposed procedure are discussed in the following sections.

III. PROPOSED MULTI-ATRRIBUTES WEB OBJECTS CLASSIFICATION

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Problem of of Multi-Attribute Relation Learning

Traditional learning systems are especially studied during the classification of supervised machine learning [21], [23], [24]. In this classification, the data object is associated with the value of the supervised learning system to determine the properties of the attribute to be used for the classification as shown in Fig. 1.

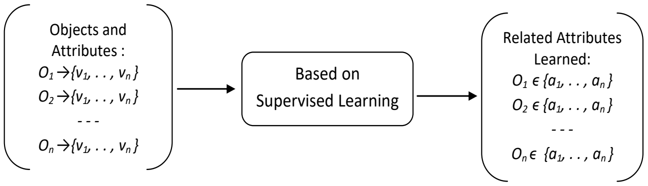

Objects and Attributes :
$O_1 \rightarrow \{v_1, . . , v_n\}$
$O_2 \rightarrow \{v_1, . . , v_n\}$
- - -
$O_n \rightarrow \{v_1, . . , v_n\}$
Based on Supervised Learning
Related Attributes Learned:
$O_1 \in \{a_1, . . , a_n\}$
$O_2 \in \{a_1, . . , a_n\}$
- - -
$O_n \in \{a_1, . . , a_n\}$

Fig. 1. Objects Attributes Learning through Traditional Supervised Learning..

Although this learning is very well suited for a single class attributes term, but complexity arises when an object has multiple attributes. In the existing traditional supervised learning [20], [22], [26] improvisations for applying multiple attribute data objects have been found. However, most of the proposed solutions are based on the functional capabilities of studying the peer dependencies or attributes and calculate the number of common events [27]. However, this type of information attribute may not be appropriate for domains that are not available. In some cases, this is determined by the association rule algorithm and the main dependencies of the correlation attribute, but this does not help to change the various sets of data attributes and other domains. The goal of this work is to create classifiers based on new associations with many attributes that can be implemented in various areas of the multidrug dataset and provide the necessary precise and quick classification.

TABLE I. DISTINCT CLASS-ATTRIBUTES TABLE

Object Class	Attributes Associating Values
Mobile	Display, memory, camera, android, Batteries, Weight, Colour, Dual SIM, Bluetooth, etc
Scenery	Clouds, Landscapes, Lakes, waterfalls, Beach, Sunset, Fall Foliage, Fields, Mountain, Urban, forest, tree, bridges, etc.
Birds	Wings, fly, Brown Creeper, Pacific Wren, Pacific-slope Flycatcher, Red-breasted Nuthatch, Dark-eyed, etc.
Book	Information, article, book, children, story, comic, computer, dynamics, education, learning, games, social, universe, etc
Vehicle	Car, model, colour, mileage, bike, speed, make, engine, power, Displacement, fuel capacity, etc
Music	Album, sound, lyrics, record, player, melody, stereo, voice, singer, etc.

The classification depends on the choice of the object and identification its attributes. It was noted that in a particular object there are two or more attributes of data, indicating a certain level of organization among them. This mult-attribute association can be very useful for multilevel data classification. The offered approach realizes the mechanism of studying and modelling. In the first step, the probability of a relation between the attribute value and the class association is calculated, which is very suitable for the object class proposal, and in the second stage the approach generates various multi-attributes that are supported to create useful class-attribute relationship models for the different classes required for classification.

To compute the relation probability of attribute and class association value (PAvalue) for an object instance this will relates the association of attributes with a Distinct Class-Attributes Table (DCAT) defined for the objects as given in Table-I.

Let's considered a set of objects "*WO"* consists of *n* objects instances having *k* attributes vectors which represented as, "*WO = {O_1,. . . , O_k}"* and its attributes as a collection of *"A = {a_1, . . , a_k }".* The objects and attributes of DCAT let be represented as, "*D = {C_1, . . . , C_n}"* and its attributes as *"T= {t_1, . . ., t_n}"* . In order to compute a probability association value, PA_{value} of an object attributes in compared with the DCAT, this will need to learn the intersection of each record of " *T* " of DCAT collection is compared to compute the association frequency as "A_{freq}" using (1). The obtained "A_{freq} " is utilized to compute the " PA_{value} " of each individual objects using (2).

$$A_{freq} = \sum\left(\int_{i=1,k=1}^{n,k}(T_n \cap A_k)\right) \tag{1}$$

$$PA_{value} = \left(\frac{A_{freq}}{n}\right) \tag{2}$$

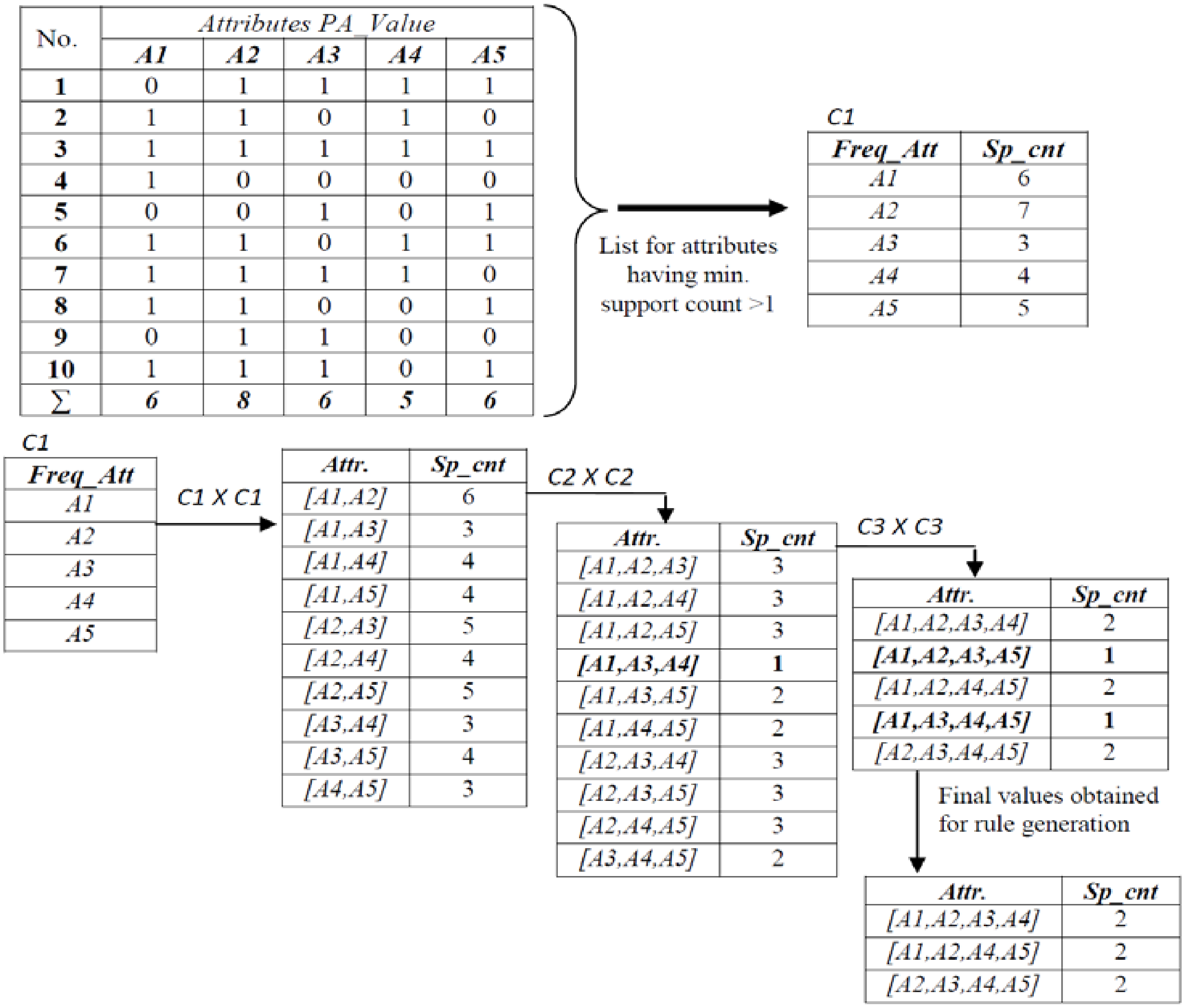
Where, *n* is the number of data objects attributes in "*DCAT*" of each individual objects, *k* is the number of multi-attributes of an object. The value of "PA_{value}" ranges between 0 to 1, the higher the value the closer the association to class. The computed "PA_{value}" of each object are being utilized to construct patterns for the classification. The algorithm-1 presents the steps of association in detail.

Algorithim-1: Finding *Associated* Class for an Objects

```
Input :    WO,  a single dimensional learning data
           DCAT, a two dimensional Distinct Class-Attributes Table
Output :  $WO_{Class}$ , Object Class Value
Method :  for i=0, i < number of objects in WO
           {
              $w_i$ = WO[i];
              A[]=getValues($w_i$);
              for t=0, t < number of records in DCAT
              {
                $C_t$ = DCAT[t][0];     // -- class value
                $T_t$ [] = SCT[t][0]; // -- Association value
                $A_{freq}$ = computeAF (A[] , $T_t$ []); -- (Eq-1)
                $PA_{value}$ = $A_{freq}$ / sizeof($T_t$ []);     -- (Eq-2)
                PA_Value[t][] = [$C_t$ ][ $PA_{value}$];
              }
              // Find the highest PA_Value associated to assign a
              class
              $WO_{Class}$ = getClass (PA_Value [ ][ ])
           }
```

Selecting a particular class for a multiple attribute results in a significant loss of information [10], [20]. To overcome this problem, the method will expanding the class that learns *PA_Value* with multiple attribute values to create associative patterns using association rules to minimize Hamming losses in the data classification. The process of creating a pattern is shown in Fig. 2.

Fig. 2 illustrates the value of an instance in the association to find multiple values for the construction of a sample of classification. Let's assume, the training datasets, "$D = \{(d_1,a_1), (d_2,a_2), \ldots , (d_n ,a_k)\}$", where $d_i \epsilon D$, $v_k \subseteq A$. To find the multi-attributes which can be highly relevant to build the classifier class accuracy this approach will consider a binary relevance of each instance attributes, for example, the value "$A = \{a_1, a_2, a_3, a_4, a_5\}$" can have a binary equivalence as, "$D = \{(0,1,1,1,0), \ldots\ldots , (1,0,1,0,1)\}$". This learning mechanism will utilize all binary values, such that the list of attributes set will be generated from *D* which supports the minimum number of support count required.



No.	*Attributes PA_Value*				
	A1	***A2***	***A3***	***A4***	***A5***
1	0	1	1	1	1
2	1	1	0	1	0
3	1	1	1	1	1
4	1	0	0	0	0
5	0	0	1	0	1
6	1	1	0	1	1
7	1	1	1	1	0
8	1	1	0	0	1
9	0	1	1	0	0
10	1	1	1	0	1
∑	***6***	***8***	***6***	***5***	***6***

C1

Freq_Att	***Sp_cnt***
A1	6
A2	7
A3	3
A4	4
A5	5

C1

Freq_Att
A1
A2
A3
A4
A5

Attr.	***Sp_cnt***
[A1,A2]	6
[A1,A3]	3
[A1,A4]	4
[A1,A5]	4
[A2,A3]	5
[A2,A4]	4
[A2,A5]	5
[A3,A4]	3
[A3,A5]	4
[A4,A5]	3

Attr.	***Sp_cnt***
[A1,A2,A3]	3
[A1,A2,A4]	3
[A1,A2,A5]	3
[A1,A3,A4]	**1**
[A1,A3,A5]	2
[A1,A4,A5]	2
[A2,A3,A4]	3
[A2,A3,A5]	3
[A2,A4,A5]	3
[A3,A4,A5]	2

Attr.	***Sp_cnt***
[A1,A2,A3,A4]	2
[A1,A2,A3,A5]	**1**
[A1,A2,A4,A5]	2
[A1,A3,A4,A5]	**1**
[A2,A3,A4,A5]	2

Attr.	***Sp_cnt***
[A1,A2,A3,A4]	2
[A1,A2,A4,A5]	2
[A2,A3,A4,A5]	2

Fig. 2. Multiple-Values Generating from Instance Value Data.

TABLE II. PATTERN GENERATED FOR THROUGH MULTI-ATTRIBUTE ASSOCIATION

Object Class Att.	**Generated Associated Multi-Values**	**Generated Patterns for Classifier**
C_{Att}	*[A1,A2,A3,A4]* *[A1,A2,A4,A5]* *[A2,A3,A4,A5]*	• *{A1},{A2},{A3},{A4},{A5}* • *{A1,A2}, {A1,A3},{A1,A4},{A1,A5}* • *{A1,A2,A3}, {A1,A2,A4},{A1,A2,A5}* • *{A1,A2,A3,A4}, {A1,A2,A4,A5},{A1,A2,A4,A5}*

In this case, since the absolute rating of support is *2*, and the minimum relative support will be "*2/10 = 20%*". The list received from "C1" is configured as an item that matches the minimum support, while the rest of the items are ignored. In addition, to determine the most common and associated attribute in the resulting "*C1*", combine the outcome with "*C1* ⋈ *C1*" to build the "*C2*" attribute with two attributes and continue until you get one value of the pattern. This iteration continues until there are several attributes that satisfy the minimum support. The final multi-attribute is considered the most relevant attributes. Now through utilizing *PA_Value Class* as C and multi-attribute ingress to create classification rules for classifiers, as shown in Table II.

The generated rules will be used for accurate classification of multiple attribute objects. Classification accuracy is also supported by effective clustering of data objects. In the next section, this work will be experimentally evaluated using multiple attribute data sets to analyze the accuracy and Hamming loss compared to conventional multi-attribute classification method.

IV. EXPERIMENT EVALUATION

A. *Multi-Value Datasets*

The complexity of multi-attribute classification stems from a variety of real-world environments and domain applications. For the data sets related to the experimental setup, the three main application areas of multi-attribute data "multimedia classification", "text classification" and "bioinformatics" are often observed. All data sets are primarily obtained from the "MULAN" [22] data store, consist of "number of instance", "attributes", "Values" and "L_{Card}" as shown in Table III.

The "L_{Card}", represents the attribute which establish the average number of attributes per test data. The "L_{Card}" measured are discussed in [3], [17] for each dataset as, "$D = \{(d_n, A_k) \mid 1 \leq n \leq k\}$" and N is the total data records are denoted as,

TABLE III. EXPERIMENT DATASETS DISTRIBUTION

Datasets	**Objects**	**No. of Instances**	**No. of distinct Attributes**	**No. of distinct Values**	**L_{Card} Value**
Bibtext	text	7395	1836	159	2.402
Scenes	images	2407	294	6	1.074
Birds	audio	645	260	19	1.014

$$L_{Card} = \frac{1}{N}\sum_{i=1}^{n} | A_k | \tag{3}$$

B. *Evaluation Measures*

1) Hamming Loss (HL): This is the most accepted measure of further attribution of errors in the misclassification of data attributes. The measure evaluates the incorrect classification of instances and pairs based on attributes that are independent of expected and related attributes. The performance is considered perfect if "*HL = 0*". Here,"*δ* is the symmetrical dissimilarity between two datasets instances", "*h* is the literal for the hamming loss", "*N* is considered as number of test datasets", "*d* is the each individual test data attributes" and "*A* is the class values that are applied to the dataset".

$$Hamming\ Loss(HL) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{A} | h(d_i)\delta a_i | \tag{4}$$

2) Accuracy (ACC): This measures the percentages of the test attributes of *d* correctly utilized for the object classification using the *A* attributes measurements in a given data set.

$$Accuracy(ACC) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{A} \left| \frac{h(d_i)\cap a_i}{h(d_i)\cup a_i} \right| \tag{5}$$

The proposed Multi-attributes Web Objects Classification (MA-WOC) is evaluated over the popular "*Weka Tool*" using the datasets of MULAN [22] as discussed above. This will be comparing the learned patterns with the standard multi-attribute classification methods as, "*BR - Binary Relevance*", "*LP - Attribute Powerset*", "*CLR - Calibration Attribute Ranking*" and "*RAkEL - Random-k-Attributeset*" [28], [29] to understand the precision of improvisation of the attribute selection and accuracy of the classification for different datasets.

C. *Result Analysis*

This section describes the experimental results analysis obtained on executing the MA-WOC and other classifiers. In utilizing the "Class-Attribute Relation learning", the approach constructs the required patterns for the classification initially. Later the learned knowledge of the MA-WOC classifier is compared with the traditional multi-label classifier methods. The results obtained for each of the data sets are presented in Table IV below.

TABLE IV. GENERATED PATTERN PAIRS FOR DATASETS

Datasets	**Labels**	**Associated multiple-values**	**Non-Associated**	**MA-WOC Classification Pairs**
Scenes	6	3	3	8
Birds	19	13	6	38
Bibtext	159	114	45	386

TABLE V. HAMMING LOSS MEASURE COMPARISON

Datasets	**MA-WOC**	**BR**	**MA-WOC**	**LP**	**MA-WOC**	**CLR**	**MA-WOC**	**RAkEL**
Bibtext	*0.0125*	0.0151	***0.0117***	0.0161	*0.0098*	0.0144	*0.0132*	0.0151
Scenes	***0.0841***	0.0973	*0.0951*	0.1437	*0.0994*	0.1121	*0.1012*	0.0962
Birds	*0.0462*	0.0561	*0.0599*	0.0735	*0.0452*	0.0506	***0.0437***	0.0489

TABLE VI. ACCURACY MEASURE COMPARISON OF CLASSIFIERS

Datasets	**MA-WOC**	**BR**	**MA-WOC**	**LP**	**MA-WOC**	**CLR**	**MA-WOC**	**RAkEL**
Bibtext	***0.7204***	0.4187	*0.6437*	0.3869	*0.5015*	0.4089	*0.3854*	0.3657
Scenes	*0.799*	0.553	***0.839***	0.5893	*0.7918*	0.5265	*0.6247*	0.6841
Birds	*0.6708*	0.4666	*0.7189*	0.5295	***0.7319***	0.528	*0.727*	0.5452

Based on the Table-IV generated patterns an experimental run of a 10-fold validation classification for the test datasets, and measures the *ACC* and *HL* in compare the traditional multi-label classifiers. The Table-V presents the classifiers Hamming Loss in comparison with the "*BR,LP*, *CLR*, and *RAkEL*". In terms of *HL*, "the lower the loss of attributes the better the accuracy". The Table VI present the *ACC*, here "the higher the accuracy the better the improvisation".

Based on the HL and accuracy assessment results as shown in Fig. 3 and Fig. 4, the evaluation analysis found that MA-WOC showed improvement compared to the three traditional classifiers, accepted in the case of the "RAkEL" and "scene" data sets. The differences observed in both cases are very small. This difference can be considered uncertain, but it is very significant in performance compared to other situations. The improvisation of classification accuracy helps to effectively classify objects and effectively support non-categorical object clustering.

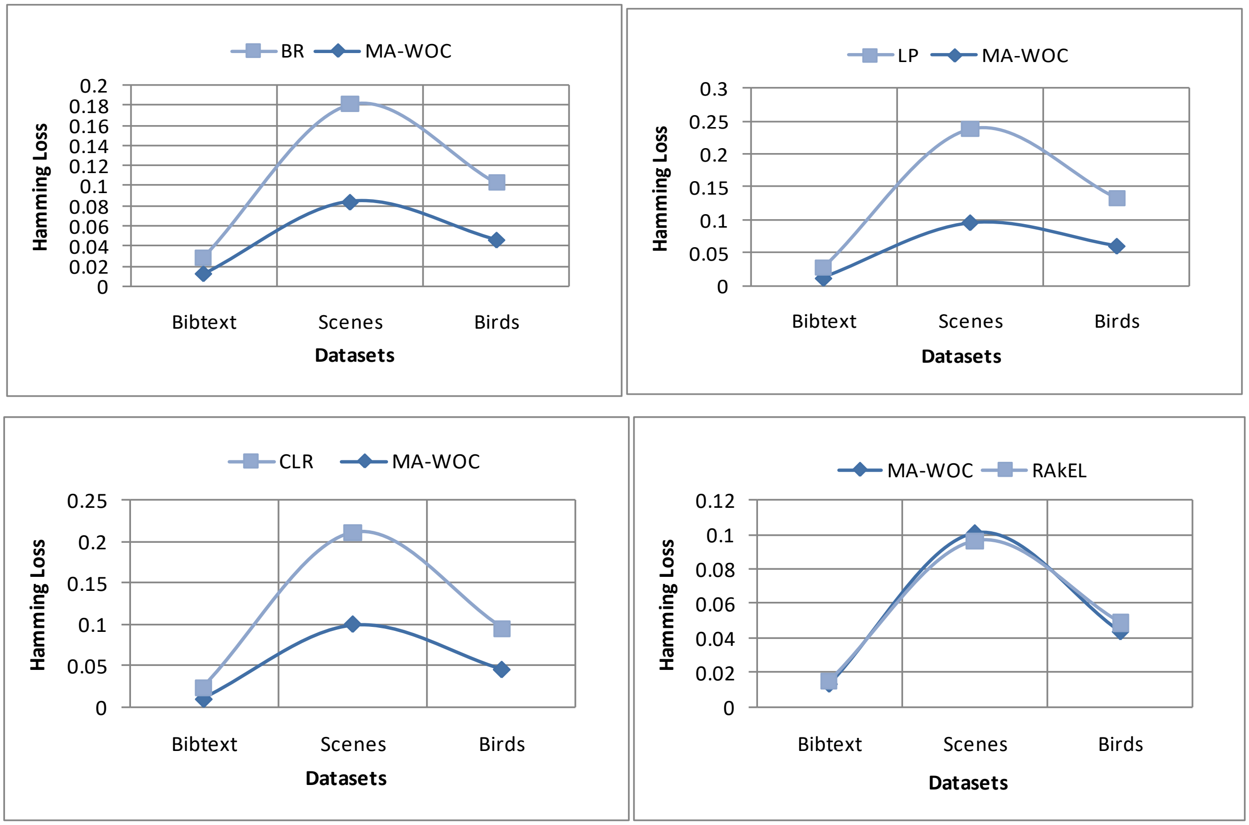


Fig. 3. Hamming Loss Comparison.

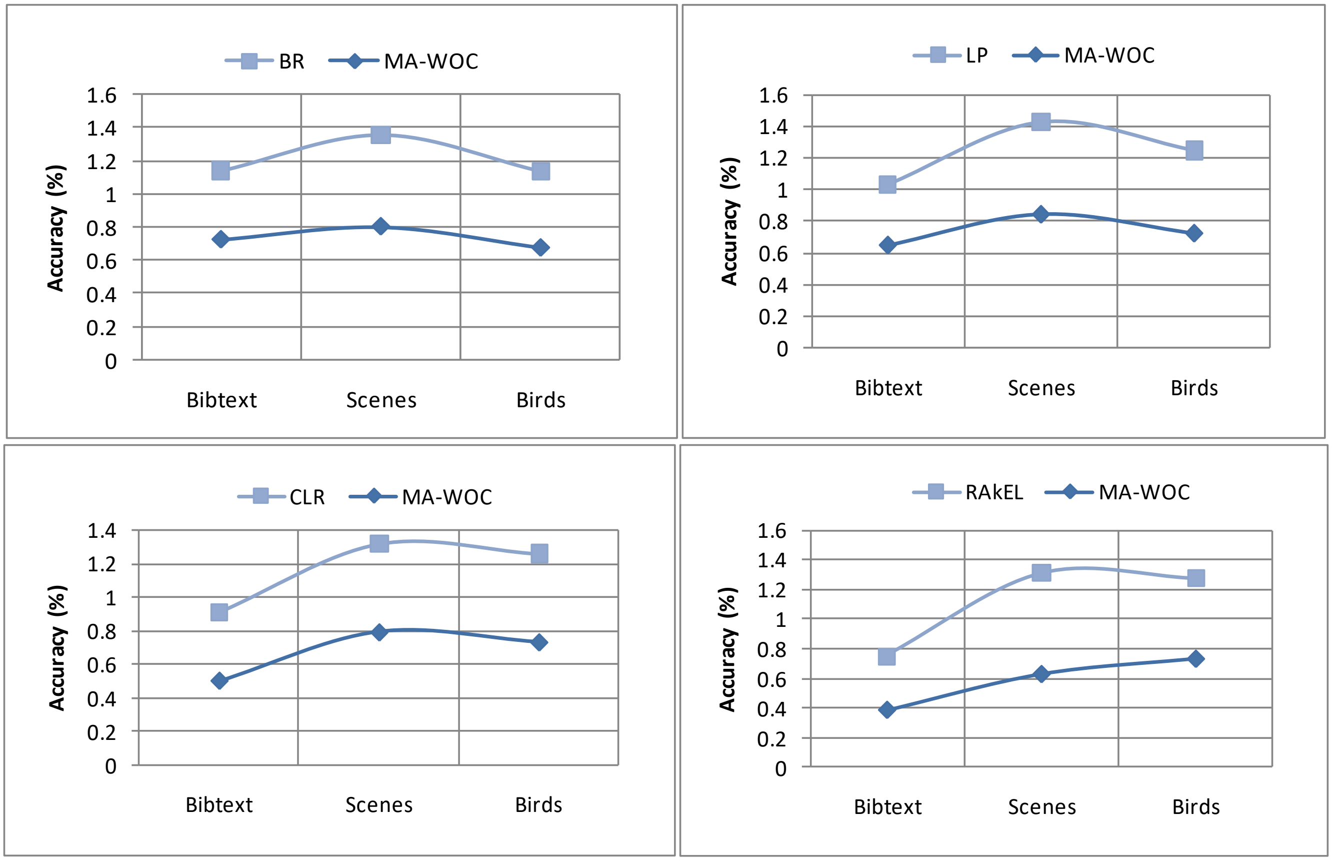


Fig. 4. Accuracy Percentage Comparison.

V. Conclusion

In this paper, a multi-attribute Web object classification (MA-WOC) based on the class attribute relationship pattern learning method is being proposed, which utilizes attribute association between multiple attributes. The learning process initially identifies the probabilities that are well-suited for the class's suggested attributes and class-associated values, and finds another plurality of attributes that support the associated probabilities in the second step to build a class pattern that is useful for different object classifications. The proposed MA-WOC calculates the associated attribute frequency and probability associations and in compare to DCAT to understand the relationship between the attributes of the object instance and the relationship of the class association. The contribution of this proposal will be used to learn various multi-attribute data sets. Experimental evaluations show possible ways to learn multiple attributes for efficient classification using different algorithms. Statistical properties terminate availability and enhancements in multi-attribute classifications. In the future, this can be further studied for the utilization of the association properties with fuzzy and Bayesian elements to accelerate and improve multi-attribute classification.

References

[1] M. Elkano, M. Galar, J. Antonio Sanz, A. Fernandez, E. Barrenechea,F.o Herrera and H. Bustince, "Enhancing Multiclass Classification in FARC-HD Fuzzy Classifier: On the Synergy Between n-Dimensional Overlap Functions and Decomposition Strategies", IEEE Transactions On Fuzzy Systems, Vol. 23, No. 5, October 2015.

[2] M. L. Zhang and Lei Wu, "LIFT: Multi-Label Learning with Label-Specific Features", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 37, No. 1, January 2015.

[3] F. Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera, "LI-MLC: A Label Inference Methodology for Addressing High Dimensionality in the Label Space for MultiLabel Classification", IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, No. 10, October 2014.

[4] J. Shen, E. Zheng, Z. Cheng, C. Deng, "Assisting Attraction Classification by Harvesting Web Data", IEEE Access Volume: 5 Pages: 1600 - 1608, 2017.

[5] T. -Y. Chan, Y.-S. Chang, "Enhancing Classification Effectiveness of Chinese News Based on Term Frequency", IEEE 7th International Symposium on Cloud and Service Computing (SC2), Pages: 124 - 131,2017.

[6] J. Ruohonen, "Classifying Web Exploits with Topic Modeling", 28th International Workshop on Database and Expert Systems Applications (DEXA) Pages: 93 - 97, 2017.

[7] M. P. El-Kafrawy, M. Amr Sauber, Awad Khalil, "Multi-Label classification for Mining Big Data", International Conference on Advances in Big Data Analytics, 2015.

[8] M. L. Zhang and Zhi-Hua Zhou, "A Review on Multi-Label Learning Algorithms", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 8, August 2014.

[9] F. Charte, A. J. Rivera, María J. del Jesus, and Francisco Herrera, "LI-MLC: A Label Inference Methodology for Addressing High Dimensionality in the Label Space for Multilabel Classification", IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, No. 10, October 2014.

[10] X. Kong, B. Cao, and P. S. Yu, "Multi-Label classification by mining Label and instance correlations from heterogeneous information networks", in Proceedings of the 19th ACM SIGKDD KDD'13. New York, NY, USA: ACM, pp. 614-622, 2013.

[11] P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic is a knowledge", In Proceedings of the 22Nd ACM International Conference on Conference on Information &#38; Knowledge Management, ser. CIKM '13, New York, NY, USA, pp. 1401-1410, 2013.

[12] L. Chekina, D. Gutfreund, A. Kontorovich, L. Rokach, and B. Shapira, "Exploiting Label dependencies for improved sample complexity", Machine Learning, vol. 91, no. 1, pp. 1-42, 2013.

[13] N. Spolaor, E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-Label feature selection methods using the problem transformation approach", Electron. Notes Theoretical Comput. Sci., vol. 292, pp. 135-151, Mar. 2013.

[14] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity", IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 3, pp. 301-312, Mar. 2002.

[15] C. -G. Li, X. Mei, and B.-G. Hu, "Unsupervised Ranking of Multi-Attribute Objects Based on Principal Curves", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 12, 2015.

[16] S. Peters et al, "Iterative annotation of multi-relational social networks", in Proc. 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, pp. 96-103, 2010.

[17] X. Shi et al, "Learning from Heterogeneous Sources via Gradient Boosting Consensus", in SIAM International Conference on Data Mining(SDM), pp. 224-235, 2012.

[18] L. Chekina, D. Gutfreund, A. Kontorovich, L. Rokach, and B. Shapira, "Exploiting label dependencies for improved sample complexity", Machine Learning, vol. 91, no. 1, pp. 1-42, 2013.

[19] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-Label image classification", in Advances in Neural Information Processing Systems USA: MIT Press, pp. 190-198, 2011.

[20] M. R. Boutell, X. Shen, J. Luo and C.M. Brown, "Learning multilabel scene classification", Pattern Recognit., vol. 37, no. 9, pp. 1757-1771, 2004.

[21] C. S. Ferng and Hsuan-Tien Lin, "Multi-Label Classification with Error-correcting Codes", 20th Asian Conference on Machine Learning, Journal of Machine Learning Research, 281-295, 2011.

[22] G. Tsoumakas, E. S.-Xioufis, J. Vilcek, and I. Vlahavas, "MULAN: A java library for multi-Label learning", J. Mach. Learn. Res., vol. 12, no. Jul, pp. 2411-2414, 2011.

[23] M. Wang, X. Zhou, and T.-S. Chua, "Automatic image annotation via local multi-Label classification", in Proc. 7th ACM Int. Conf. Image Video Retrieval, Niagara Falls, Canada, 2008, pp. 17-26, 2008.

[24] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hullermeier, "On Label dependence in multi-Label classification", in Workshop proceedings of learning from multi-Label data. Citeseer, pp. 5-12, 2010.

[25] N. Spolaor, E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-label feature selection methods using the problem transformation approach", Electron. Notes Theoretical Comput. Sci., vol. 292, pp. 135-151, Mar. 2013.

[26] G. Tsoumakas, M.-L. Zhang, and Z.-H. Zhou, "Tutorial on learning from multi-label data", in Proc. Eur. Conf. Mach. Learn. Principles Practice Knowl. Discov. Databases, Bled, Slovenia, 2009.

[27] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hullermeier, "On label dependence in multi-label classification", in Workshop proceedings of learning from multi-label data. Citeseer, pp. 5-12, 2010.

[28] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-Label classification", In Springer Machine Learning, vol.85, no. 3, pp. 333-359, 2011.

[29] G. Tsoumakas and I. Katakis, "Multi Label classification: An overview", International Journal of Data Warehouse and Mining, vol. 3, pp. 1-13, 2007.

Improved Design of an Adaptive Massive MIMO Spherical Antenna Array

Mouloud Kamali[1]
Electrical engineering department, National Engineering
School of Carthage, Tunis Carthage-Tunisia
Analysis and Processing of Electrical and Energic Systems

Adnen Cherif[2]
Analysis and Processing of Electric and Energetic Systems
Faculty of Sciences of Tunis
ELMANAR II, PB.1060, Tunisia

Abstract—Massive capacity and connectivity are the main boundaries towards standing the Internet of Everything (IoE) basis and defining modern wireless generation requirements. These needs cannot be achieved by already deployed phased array antenna in terms of distributed and oriented geometry, dimensions and design. We propose in the present paper an innovating massive multiple input multiple output (MIMO) spherical array network aiming to draw a new three-dimensional configuration to enhance the beam steering, improve bandwidth, total capacity and the scan flexibility. Resolved issues in concordance with 5G requirements are adaptive massive MIMO by using millimeter-wave antenna arrays, small cell design and definition of recommended operational frequency considering the International Telecommunication Union (ITU) norms and directives. The new geometric forms of spherical smart antenna could easily scan all 3D space, ensure higher capacity and reach tens of Giga bit per second (Gbps) value besides eradicating energy wastage aspect of Beam Division Multiple Access (BDMA) in base stations. Mathematical design is detailed and performed simulation results are presented using MATLAB software Tool.

Keywords—Adaptive spherical antenna; beam division multiple access BDMA; massive multiple input multiple output MIMO; millimeter-wave mm-wave

I. Introduction

The transported information via mobile networks, counting data, voice and internet are continuously growing with a drastic quality and quantity. The main polemic topic is that older generations of wireless communication from 0G to 4G doesn't support the new information flux.

Unceasingly growing bit rate bulk either at transmission or at reception's end, Besides the substantial strengthening of interrelated networks fitting the base of a new architype developing from the significant objects link such as robots, machines and vehicles. This paradigm is well known in IoE included in 5G.

The most attractive key ascertainments resolved by 5G are: high mobility, slight latency besides the massive connectivity and capacity.

The next expected imminent event of mobile communications is going to be unrelated to nowadays usage. Even though, Request for mobile broadband will remain to increase, fundamentally determined by the new usage of ultra-high definition video and K screen modes, we have already seen the expanding influence of cars, robots and devices that must able to attain and outline themselves to achieve the human necessities. There will be surely many not yet perceived, several performs appear in horizon and engender new materials that communications networks must pact with the demanded cost efficiency.

Essentially, there is tendencies concentrating on how to support the vicissitudes of real-world mobility, virtual mobility, 4th industrial revolution, Solid-performance mobile infrastructure and numerous other related domains such as healthcare, Rail network and automotive, Manufacturing and logistics… . Consequently, the overall annexed industry will be affected by this ongoing migration to the 5G telecom mobile.

As a result, A comprehensive mobile data traffic is increasingly growing in term of new connected everything number [1], including pedestrians, vehicles, healthcare [2] and machines. To upgrade massive connectivity capacity and insignificant latency we should guarantee; 1000X higher mobile data volume per area 10-100X higher [3], Number of connected devices for Internet of Things IoT [4], 10-100X higher typical user data rate [5], 10X longer battery life for low power Machine to Machine M2M Communications, 5X reduced End to End E2E [6] latency.

To accommodate this new trend, the wireless network infrastructure as we said needs considerable upgrades beyond the already installed capabilities. Afterward, many researchers are focusing on several aspects like operable carrier frequencies and service guidelines. The gigantic accessible bandwidth for highly multidirectional wireless links must be in accordance with future generations of wireless mobile networks.

A special interest is imposed to be considered for the channel characteristics, which, in many situations is not well recognized because of the needed bandwidth range. Obviously, at these higher frequencies, GHz signals are working as well as the distance of serving others are nearby some hundred meters.

In literature, researchers are mainly focusing on the modeling and design aspects of millimeter-wave, channel model, energy efficiency, new antenna geometries, MIMO models, Mathematical architecture modeling, Multi-Beam antenna and coders. Hereunder we cite some of them were:

Authors in [7] focus on the impact of blockage of mm-Wave signals in spatial reduced coverage owing penetration through the human hand, body and vehicle at 28 GHz. They use statistical blockage models aiming to reproduce last cited impacts (hand, human body and vehicle). Their results show that the time-scales corresponding to blockage on the order of some hundreds of milliseconds. Also, a robust of mm-Wave beamforming to handle blockage studies have been presented.

In [8] authors propose a 3D spatial channel model controlling, aiming to evaluate practically conceivable boundaries of massive MIMO base stations standardized by the third Generation of mobile networks 3G. They consider the Base Stations BSs completely loaded and different configurations of active User Ends UEs per cell besides, they present the established statistical approach to reduce to half the compliance distance in comparison to the traditional technique.

While in [9] authors investigate the problem of designing a control channel in a 5G system. Their control chain includes the transmission, under the severe latency and reliability of a short data packet comprising a trivial information payload, over a propagation channel that offers limited frequency diversity and no time diversity. They present an achievability bound, built upon the random-coding union bound which trusts on quadrature phase-shift FSK keying modulation and pilot-assisted transmission to estimate the diminishing channel, and scaled nearest-neighbor decoding at the receiver. They determine the number of pilot symbols that should be transmitted to optimally occupy between channel-estimation errors and rate loss related to pilot overhead. Besides, they underline the importance of using multiple antennas at the transmitter and/or the receiver.

In the same axis [10] presents energy-efficient resource allocation in multiple antenna wiretap channels is investigated, focusing on maximum power and minimum privacy capacity/rate constraints. Two energy-efficiency metrics were optimized, knowing the secrecy energy efficiency and the secret-key energy efficiency. If the valid receiver and the listener have a single antenna, and the transmitter has multiple antennas, the global solution provided by the authors is expressed by a simple formula resolved by iterative algorithms which can determine the global maximum of the secret-key energy efficiency and candidate solutions related to secrecy energy efficiency maximization problem.

Though in [11] they present a novel geometry-based statistical model for small-scale non-wide-sense stationary uncorrelated scattering mobile-to-mobile Rayleigh fading channels. Their model is based on the plane wave propagation to capture the temporal evolution of the propagation delay and Doppler shift of the received multipath signal which is different from spherical wave propagation statistical model method, which yield more realistic case. They consider an arbitrary geometric configuration of the propagation zone. They derive general expressions for the most important statistical quantities of non-stationary channels including the frequency correlation function, the local scattering function, the frequency dependent time and the Doppler profiles.

In [12], the author presents 3D vehicle MIMO antenna array model for vehicle-to-vehicle (V2V) communication environments. A spherical wave front is used in the proposed model as an alternative of the plane wave front supposition used in the conventional MIMO channel model. Using the proposed V2V channel model, they first derive the closed-form expressions for the joint and marginal probability density functions of the angle of departure at the transmitter and angle of arrival at the receiver in the azimuth and elevation planes. They analyse in addition the time and frequency cross-correlation functions for different propagation paths. An expression of Doppler spectrum related to the relative motion between the mobile transmitter and mobile receiver has been derived from the proposed model.

Unlike [13] they reflect the possibility of locating a system where antenna arrays are organized as a Large, Intelligent Surface (LIS) electronically active with integrated electronics and wireless communication making the entire situation “intelligent”. They extensively discuss the impact of centralized and distributed deployments of LIS and show that a distributed deployment of LIS could enlarge the coverage and improve the global positioning performance.

In [14], A spatially correlated large antenna array operating at mm-Wave frequencies are considered. The properties of the Power Elevation Spectra (PES) on the meeting massive MIMO properties are then demonstrated by defining and deriving a diagonal dominance metric. Statistically, the properties of the antenna element Mutual Coupling (MC) are exposed on the active spatial correlation (SC), mm-Wave user rate, and eigenvalue structure for different antenna topologies. It is determined that even if MC could meaningfully decrease SC for side-by-side dipole antenna elements, the modification in antenna real gain and consequently, signal-to-noise ratio (SNR) caused by MC becomes a governing effect and eventually limits the antenna array performance. The mm-Wave user rating system with hybrid beamforming, using an orthogonal matching pursuit algorithm, is then shown for different antenna topologies with dipole and cross polarized (x-pol) antenna elements.

While [15] they establish a full 3D channel model to support the performance analysis and evaluation of active antenna array based on wireless communication systems. they analyse and compare the impact of three different down-tilt methods employed in active antenna array antennas, electrical down-tilt, mechanical down-tilt, hybrid down-tilt, on the antenna patterns, which would notably impact the performance of mobile wireless communication systems. A comparison of the wireless communication system throughput of 2D and 3D wireless channel models has been developed. The system performance in terms of capacity and coverage with different active antenna arrays under the 3D channel has been investigated. Significant gains in coverage and capacity for individual antennas with a narrow beam of the vertical patterns have been observed using down-tilt optimization. But it may not lead to a distinguished gain for individual antennas with relatively large beam-width of the vertical patterns.

In the case of [16] a mathematical model is provided of a novel hybrid precoding architecture, and an efficient alternating descent algorithm is developed to jointly design the analog and digital precoders for the last proposed hybrid precoding scheme. Numerical results obtained demonstrate that their proposed precoding scheme with multi-feed reflect array antennas achieves much better precoding performance than its sub-connected counterpart with phased array antennas.

While, in [17] a massive MIMO is presented as an extension of traditional MIMO technology, it subsequently improves the throughput rate, energy efficiency, the link reliability and data transmission rate. It also improves through a lot of increasing the antenna communication number, by means of very duplex communication mode, which correspond to a high spectrum efficiency system.

In [18] the achievable sum-rate of the proposed beam-domain full-duplex (FD) massive MIMO transmission scheme is analysed and the joint user and BS power allocation scheme are proposed to optimize the system achievable sum-rate. Their simulation results show that the proposed beam-domain FD transmission scheme outperforms existing time division duplex and frequency division duplex massive MIMO and FD massive MIMO transmission schemes on the spectral efficiency performance.

Whereas [19] presents interference moderation for multi-beam antenna subsection modulation via side lobe level reduction is introduced. A method for designing thinned arrays with minimum side lobe levels for antenna subset modulation is introduced and generalized for multi-beam antenna subset modulation. A new variable constraint is applied to the optimization problem to control the localization of optimum solution within the antenna array. Two solutions are introduced, convex optimization combined with local search and local search assisted genetic algorithm.

Although [20] is a formalization of hybrid precoders design as a block-sparse reconstruction problem and a minor complexity algorithm for finding precoders is proposed based on the greedy sequence clustering. To assure the performance of this algorithm, an estimation method for the number of the blocks and radio frequency chains is proposed. Under this outline, for the analog precoder the number of the phases needed to be quantized is only two times minus one of that of all blocks, which remains unchanged and even the number of chains increases. They demonstrate that their proposed algorithm could achieve the near-optimal performance and save the feedback overhead by utilizing the block sparsity information.

All these cited works have some drawbacks existing in:

- Studies don't match mm-Waves according to the atmospheric and molecular absorption at mm-Wave frequencies but more exactly cm-waves.
- 4G multi-cell and multi-layer optimisation doesn't match the case of 5G there is a new central universal resource management.
- The capacity offered by PCM modulation is insignificant regarding QAM, which ought to be not less than 256 bits constellation.
- The energy efficiency increases when the distance between BS and UE is important. Thus, the capacity diminishes. Nevertheless, in 5G there are small cells when the capacity is higher and consumed energy is slighter.

Our novelty presented in this paper compared to already cited references was resumed in three main originalities:

First, a new spherical topology in which the number of elements is significantly boosted, embedded in a massive MIMO. Second, the frequency is elevated. Third, cell dimensions become smaller. All this, knowing that in urban zones, already installed antenna array geometries are limited to direct visibility and don't deal in the case of high buildings connection requests. Spherical antenna array with Massive MIMO ensures as a result in dense and vertical buildings agglomerations the coverage quality with the same probability for 3D space.

This paper is divided into 3 sections. The first section is the introduction, the second section treats the mathematical modelling of the spherical antenna, the third purposes to present performed simulation results and its related discussion and finally the conclusion and the main perspectives have been done.

II. Mathematical Modeling of The Spherical Antenna

$$E_{total} = E_{single\ element\ at\ reference\ point} \ *\mathrm{AF} \qquad (1)$$

The total electrical field is the multiplication of the single element reference point field by the array factor.

$$F_{total}(\theta,\phi) = F_{element}(\theta,\phi) * AF_{array}(\theta,\phi) \qquad (2)$$

θ is the elevation angle and φ is the azimuth angle.

A. Single Element at Reference Point of the Electrical Field

Represented by Fig. 1 hereunder the Microstrip element model dedicated for spherical Antenna.

The free-space wavenumber $K = \frac{2\pi}{\lambda}$ (3)

The patch Microstrip: length *L*, width *W*,

where $W = L = \frac{\lambda}{2}$

The magnitude of the fields, given by:

$$f(\theta,\varphi) = \sqrt{E_\theta^2 + E_\varphi^2} \qquad (4)$$

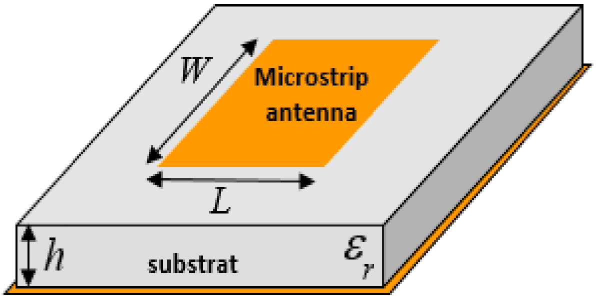


Fig. 1. Microstrip element model dedicated for spherical Antenna

$$E_\theta = \frac{\sin(\frac{KW\sin\theta\cos\varphi}{2})}{\frac{KW\sin\theta\cos\varphi}{2}}\cos(\frac{KL\sin\theta\cos\varphi}{2})\cos\varphi \quad (5)$$

$$E_\varphi = \frac{\sin(\frac{KW\sin\theta\cos\varphi}{2})}{\frac{KW\sin\theta\cos\varphi}{2}}\cos(\frac{KL\sin\theta\cos\varphi}{2})\cos\theta\sin\varphi \quad (6)$$

B. Array Factor

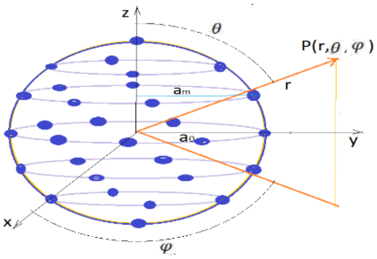


Fig. 2. Antenna topology and disposition for spherical antenna array

Conformal arrays are specifically conceived in a manner to imitate a surface. The radiating elements of the spherical array as shown in Fig. 2 could be easily embedded in the tight curved surface, thus automatically reduce disagreements by the installed components. The spherical shaped antenna array could be preserved as one of the greatest elements of the conformal array. The most striking feature of spherical array is as follows:

- All points in the far-field will perceive the same environment with a spherical array because of its symmetrical elements disposition due to the radiation pattern that will persist like the considered far-field point is enthused over the space.
- These arrays could be handled to reach multiple-beams and adaptive patterns redesigning based on its electronic beam routing and signal processing capabilities.

$$AF_{Spherical}(\theta,\varphi) = \sum_{n=-N}^{N}\sum_{m=1}^{M}[I_{mn}e^{(j\mathrm{Ka}_n\sin\theta\cos(\varphi-\varphi_{mn})+j\varphi_m)+\mathrm{jKd}_n\cos(\theta+\beta_n)}] \quad (7)$$
$$+e^{\mathrm{jka}_0\ \cos\theta} + e^{\text{-}\mathrm{jka}_0\ \cos\theta}$$

$$\mathrm{a_n} = \sqrt{\mathrm{a}_0{}^2 - d_n{}^2} \quad (8)$$

I_{mn} is the current excitation for mth antenna element of nth circular array.
K is the propagation constant,
θ is the elevation angle,
ϕ is the azimuth angle,

ϕ_{mn} is the azimuth position of mth antenna element on nth circular array.
$\mathrm{a_n}$ is the radius for nth circle of spherical array.

$\mathrm{a_0}$ is the radius of spherical array.

ϕ_m is the beam steering phase angle in azimuth direction.

d_n is the distance of nth circular array from reference circular array at the origin.

β_n is the progressive phase shift between nth and reference circular array.

To design the spherical geometry, as circular arrays must be settled in a linear manner. Fig. 3 describes the disposition of the antenna array.

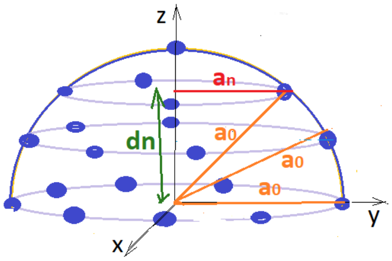


Fig. 3. Antenna topology and disposition

C. Total Electrical Field Magnitude

$$E_{total} = \sqrt{E_\theta^2 + E_\varphi^2} * [\sum_{n=-N}^{N}\sum_{m=1}^{M}[I_{mn}e^{(j\mathrm{Ka}_n\sin\theta\cos(\varphi-\varphi_{mn})+j\varphi_m)+\mathrm{jKd}_n\cos(\theta+\beta_n)}] + e^{\mathrm{jka}_0\ \cos\theta} + e^{\text{-}\mathrm{jka}_0\ \cos\theta}]$$

III. Simulation Results and Discussion of Massive MIMO and Mmwave Spherical Antenna Arrays

Filling a spherical array arranged one above the other, with the same separation distance between them, such as the cylindrical array wherein the antenna elements form a circular surface area equivalent to a disk array of antenna settled in multiple layers as shown in Fig. 3 and Fig. 4.

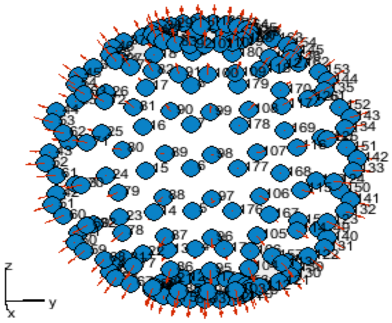


Fig. 4. Antenna topology and disposition of spherical Antenna

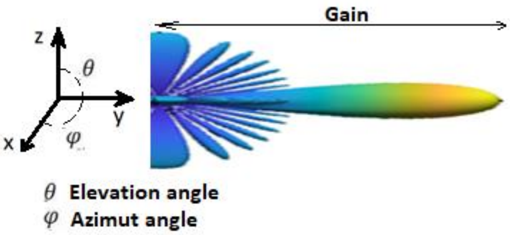


Fig. 5. Narrow beam width power for spherical Antenna topology

In Fig. 5 the beam direction is commanded by azimuth and elevation angle. The power is defined by the distance between the user and the antenna demonstrated by the beam width.

In Fig. 6 a three-dimensional sight of array directivity at 90° azimuth angle and 0° elevation working at 73Ghz frequency indicating that the beams divide between requesting users. The yellow color matches 20 dBi directivity where stated support users.

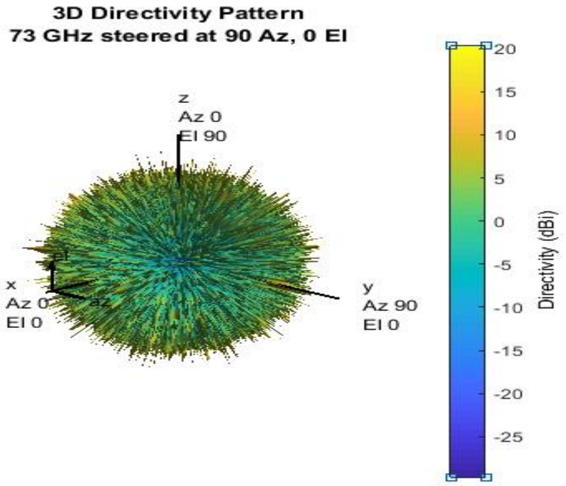


Fig. 6. Pattern power in 3D of the spherical Anenna

TABLE I. ARRAY CHARACTERISTICS

Array directivity	**22.42 dBi at 0° Az,0° El**
Spherical Array Span	**Rsphere=4.5cm**
Number of elements	**Nantenna=182(massive MIMO)**

Table 1 illustrates the array directivity, array span and the number of deployed antennas on our stated topology.

Fig. 7 labelled a cut section of the overall directivity pattern at the same azimuth associated to 0°elevation angle.

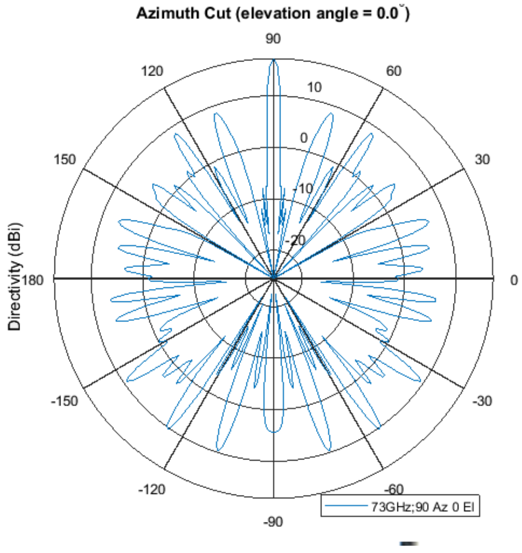


Fig. 7. Azimuth cut elevation angle of spherical Anenna

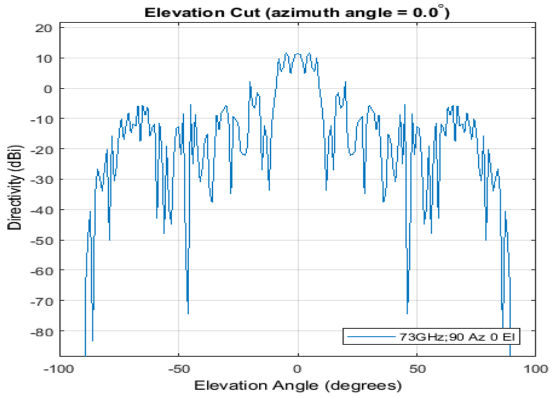


Fig. 8. Elevation cut angle of spherical Anenna

Fig. 8 shows the spherical system antenna Elevation cut angle as given in the design; the excitation coefficients, inter element arrangement, magnitude and phase, and angular partings are reformed in a way to acquire the required pattern in term of directivity strengthening, side lobe reduction, pattern nulling and shaping.

Fig. 9 describes a cut section of the overall directivity at elevation cut related to azimuth angle 0°. The main supported user is at 90° azimuth as we visualize the maximum received power is confined to the cibled point in concordance with spherical coordinates all around the space.

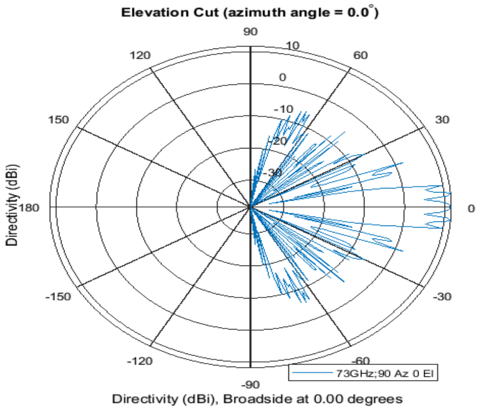


Fig. 9. Elevation cut Azimuth angle of spherical Anenna

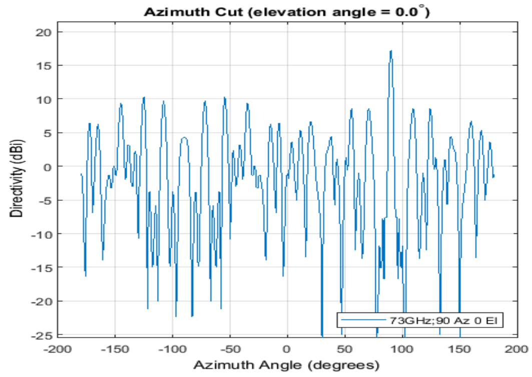


Fig. 10. Azimuth cut angle of spherical Anenna

The last figure, Fig.10 presents a cut section of the overall directivity at 90° azimuth cut associated to elevation angle 0°. The served user is at 90° azimuth as we visualize the maximum received power is axed to the cibled point in concordance with spherical coordinates all over the space.

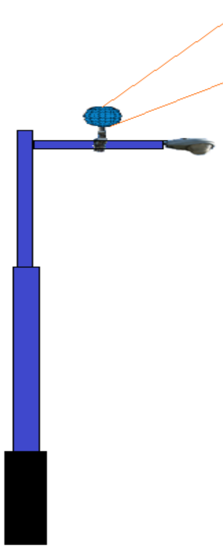


Fig. 11. Example of small cells on Lamposts of spherical Antenna

Figure 11 demonstrates a prototype disposition tied to lampposts of the spherical antenna array.

TABLE II. SUMMARY OF MATLAB VALUES

F=73 GHz	182 antenna (massive MIMO)					
φ (azimut): (-180° to +180°)	0	30	60	90	120	180
θ (elevation): (-90° to 90°)	0	15	30	45	60	90
Pr [dBi]	22.83	22.32	22.46	22.59	22.61	22.51

- Link 73GHz Spherical properties:
- F=73GHz.
- λ =4 x 10^{-3}m=4mm.
- d=λ/2=2 x 10^{-3}m=2mm.
- R_{sphere}=45 x 10^{-3}m = 4.5cm.

Table 2 represents the numerical simulation values in MATLAB. A contingent of the user locates the 3D spherical antenna array in content of users demand with maximal power of 22 dBi thanks to the substantial number of adaptive antenna elements. Hence, beams are most directives and narrow P (r, θ,φ).

An antenna matrix in the base station BS (2x2,3x3,4x4 …) [21] edges the power depending only on the demanding user.

Hence the power is adjusted since exclusively demanding users are attended. Then, Services price will be more redressed for both sides, the operators and customers besides the omission of electromagnetic noise.

IV. CONCLUSION

Along this paper, we have modelled mathematically a spherical antenna array with Massive MIMO, which could offer several benefits for migration into 5G mobile in urban zones, when cylindrical array antenna is considered to be functionally limited to direct visibility and not for high buildings cases. Spherical antenna array with Massive MIMO ensures as a result in dense and high agglomerations with the same probability for 3D space:

a) Largely increase the link reliability and the data rate.

b) Massive MIMO ensures:

- a deterministic channel matrix.
- could be constructed with cheap, low-power units.
- simplify the multiple access layer.
- contain systems with hundreds, even thousands of antennas could enable several-gigabit rate transmissions at high spectral efficiency.

However, numerous technical tasks have to be undertaken in realizing such large MIMO systems. The solution which enables 5G migration is firstly increasing the number of antenna elements using a spherical topology that means a Massive-MIMO system. Use a millimeter wavelength and small cells implicitly a low power consumption due to the adaptive array antenna.

Inspite of the spherical geometry effectiveness there is a little limitation for connectivity in very high speed case. Indeed, Connectivity at 1200 Km / h speed, users inside an airplane probably lose the network connection due to the absence of a a complementary assist as a Low Earth Orbit (LEO) satellite base station. Thus, a satellite support must be taken into account as a future work.

REFERENCES

[1] Marwan A. Al-Namari, Ali Mohammed Mansoor, "A Brief Survey on 5G Wireless Mobile Network", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 11, 2017.

[2] Farah Nasri, Abdellatif Mtibaa, "Smart Mobile Healthcare System based on WBSN and 5G", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[3] Nisha Panwar, Shantanu Sharma, and Awadhesh Kumar Singh, "A Survey on 5G: The Next Generation of Mobile Communication∗", Elsevier Physical Communication, Volume 18, Part 2, pp. 64-84, March 2016.

[4] Mohammad Saeid Mahdavinejad , Mohammad reza Rezvana , Mohammad amin Barekatain, Peyman Adibi, PayamBarnaghi, Amit P.Sheth, "Machine learning for Internet of Things data analysis: A survey", Digital Communications and Networks, In Press, 12 October 2017.

[5] Chun-Nan Liu, "Trend, technology and architecture of small cell in 5G era", IEEE, VLSI Technology, Systems and Application (VLSI-TSA), 2016 International Symposium on, 25-27 April 2016 Hsinchu, Taiwan, pp. 1-2, May 2016.

[6] Ekram Hossain ; Monowar Hasan, "5G cellular: key enabling technologies and research challenges ",IEEE Instrumentation & Measurement Magazine, Volume 18, Issue 3, pp. 11 – 21, June 2015.

[7] Vasanthan Raghavan, Lida Akhoondzadeh-asl, Vladimir Podshivalov, Joakim Hulten, M. Ali Tassoudji, Ozge Hizir Koymen, Ashwin Sampath, Junyi Li, "Statistical Blockage Modeling and Robustness of Beamforming in Millimeter Wave Systems", ARXIV, pp.1-28, Janvier 2018.

[8] Paolo Baracca, Andreas Weber, Thorsten Wild, Christophe Grangeat, “A Statistical Approach for RF Exposure Compliance Boundary Assessment in Massive MIMO Systems”, ARXIV,pp.1-6, Janvier 2018 .

[9] Guido Carlo Ferrante, Johan O stman, Giuseppe Durisi, and Kittipong Kittichokechai, “Pilot-Assisted Short-Packet Transmission over Multiantenna Fading Channels: A 5G Case Study”, pp.1-6, Febraury 2018

[10] Alessio Zappone, Pin-Hsun Lin and Eduard Jorswieck,” Optimal Energy-Efficient Design of Confidential Multiple-Antenna Systems”, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, Volume 13, issue 1, January 2018.

[11] Carlos A. Guti´errez, Jos´e T. Guti´errez-Mena, Jos´e M. Luna-Rivera, Daniel U. Campos-Delgado, Ramiro Vel´azquez, and Matthias P¨atzold, ” Geometry-Based Statistical Modeling of Non-WSSUS Mobile-to-Mobile Rayleigh Fading Channels ”, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, Volume 67, Issue 1, January 2018.

[12] Hao Jiang, Zaichen Zhang, Jian Dang, Member and Liang Wu, “A Novel 3-D Massive MIMO Channel Model for Vehicle-to-Vehicle Communication Environments”, IEEE TRANSACTIONS ON COMMUNICATIONS, Volume. 66, Issue 1, pp. 79 – 90, January 2018.

[13] Sha Hu, Fredrik Rusek, and Ove Edfors, “Beyond Massive-MIMO: The Potential of Positioning with Large Intelligent Surfaces”, IEEE Transactions on Signal Processing, Volume PP Issue 99, pp. 1-14, January 2018.

[14] Callum T. Neil, Adrian Garcia-Rodriguez, , Peter J. Smith, Pawel A. Dmochowski, Christos Masouros and Mansoor Shafi, “On the Performance of Spatially Correlated Large Antenna Arrays for Millimeter-Wave Frequencies”, IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, Volume 66, Issue 1, pp. 132- 148, January 2018.

[15] Guodong Li, Jinsong Wu, Zhixin Chen, Xiong Luo, Taolin Tang and Zhiqiang Xu, “Performance analysis and evaluation for active antenna arrays under three-dimensional wireless channel model”, IEEE Access, Volume PP, Issue 99, PP. 1-10, January 2018.

[16] Zhengyi Zhou, Ning Ge, Member, Zhaocheng Wang, Sheng Chen, “Hardware-Efficient Hybrid Precoding for Millimeter Wave Systems with Multi-Feed Reflectarrays”, IEEE Access, Volume PP, Issue 99, PP. 1-11, January 2018.

[17] Qiang Hu, Meixiang Zhang, and Renzheng Gao, “ Key Technologies in Massive MIMO ”, 4th Annual International Conference on Wireless Communication and Sensor Network (WCSN 2017), Volume 17, Issue ITM Web conference, PP. 1-10 February 2018.

[18] Kui Xu, Zhexian Shen , Yurong Wang, Xiaochen Xia, “Beam-domain full-duplex transmission in massive MIMO system”, Physical Communication , Volume 26, Issue 2018, pp.116–127, December 2017.

[19] Amr AkL, Ahmed Elnakib, Sherif Kishk, “Antenna array thinning for interference mitigation in multi-directional antenna subset modulation”, Physical Communication, Volume 26, Issdue 2018, pp. 31–39, November 2017.

[20] Xuefeng Liu, Weixia Zou, “Block-sparse hybrid precoding and limited feedback for millimeter wave massive MIMO systems”, Physical Communication, Volume 26, Issue 2018 , pp. 81–86, December 2017.

[21] Mouloud Kamali, Adnen Cherif, “Adaptive Cylindrical Antenna Array For Massive MIMO in 5G” , International Journal of Computer Science and Network Security, VOL.18 No.3, March 2018.

Monitoring, Detection and Control Techniques of Agriculture Pests and Diseases using Wireless Sensor Network: A Review

S.Azfar[1]
Department of Computer Science,
Federal Urdu University of Arts, Science & Technology
Karachi, PAKISTAN

A.Nadeem[2], A.B. Alkhodre[*3]
Department of information technology
Faculty of Computer Science and Information System,
Islamic University, Medina, KSA

K.Ahsan[4], N. Mehmood[5]
Department of Computer Science,
Federal Urdu University of Arts, Science & Technology,
and University of Karachi, Karachi, PAKISTAN

T.Alghmdi[6], Y.Alsaawy[7]
Department of Computer Science
Faculty of Computer Science and Information System,
Islamic University, Medina, KSA

***Abstract*—Wireless sensor network technology is widely used in the western world for improving agriculture output. However, in the developing countries, the adaptation of technology is very slow due to various factors such as cost and unawareness of farmers with the technology. There are reports in the literature related to the precision agriculture and hopefully, this paper will add to the knowledge of the use of Wireless sensor network (WSN) for monitoring agriculture fields for pest detection. The literature related to pest monitoring and detection using wireless sensor networking technologies are reviewed. Then, the advanced sensing technologies are currently in use for the detection of a pest has been described. The existing techniques about pest detection and disease monitoring are evaluated on the basis of some key parameters such as the type of sensors used, their cost, processing tools, etc. Finally, the sensing technologies and the possibility of using third generation sensing technology for monitoring and detection of cotton crops are analyzed.**

Keywords—Component; pest monitoring and detection; Wireless Sensor Network; pests; agriculture; sensing technology

I. Introduction

To prevent the crops from pests and their related diseases is a difficult task for farmers. The pests can harm crop, reduce yields and also impact negatively on crop quality. Conventional farmers use a lot of techniques to kill the pests. Identification of pest disease is necessary before treatment. Without identifications, the use of pesticides causes many negative results such as pest can develop immunity to pesticides leading to changed results and stronger pesticides, along with harmful pests it also kills many beneficial pests and natural enemies of pests causing an increase in the insect population. The use of pesticides affects crops that are insect pollinated population resulting in their failure to develop fruits.

Cotton, with its green juicy leaves, its large open flowers, nectarines, and its large fruits attract and support various insects and mites. Over 1000 type of cotton pests has been recorded across the world. Nearly 125 species have been reported to attack cotton crops in India [1] and United States [2]. More than 93 species of insects and mites are reported damaging the cotton crops in Pakistan [3].

In the present paper, the research has focused on developing countries such as India, Pakistan, Bangladesh was the use of wireless sensor network technology is rare. It considers Pakistan as an example where wheat, rice, cotton, and sugarcane are major crops. These four crops produce 33 percent of the total value of agriculture production in Pakistan. Some minor crops contribute 11.1 percent in country's overall agricultural income. Even though agriculture has been playing a pivotal role in the economy of Pakistan since its creation but in the last three decades income, growth and exports of our agri-related products have declined.

The use of wireless sensor network technology is shown in Fig. 1. Monitoring and managing pest could significantly improve the production and quality of crops in developing countries. It has compiled the existing literature to determine the feasibility of using wireless sensor network technology in developing countries.

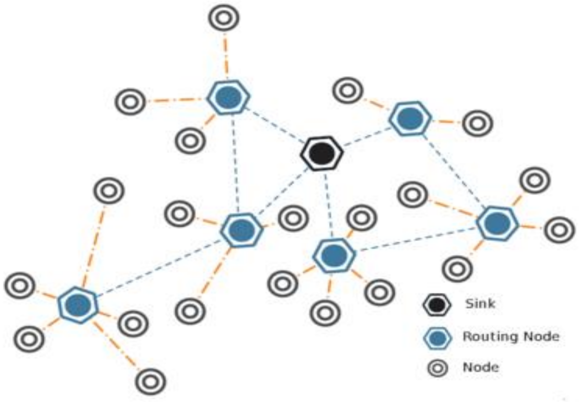


Fig. 1. Structure of Wireless Sensor Network

In the literature, there are only a few reviews on the subject of precision agriculture using WSN where authors have demonstrated the effective use of sensors in agriculture. In [15] authors performed a survey of plant disease detection using WSN and image processing while in [18] authors discussed to pest monitoring and detection techniques. In [25 and 26] authors reviewed sensing technologies available for agriculture and food industry and their future trends. In [31 and 32] authors focused on precision agriculture using WSN and they classified existing remote monitoring and control systems. The last review paper was published four years ago. Moreover, these reviews are crop specific. This paper includes reviews of the WSN application in agriculture and covers many valuable parameters. . For that reason, each and every aspect of this domain which presents several pest detection methods, disease monitoring schemes and current and future sensing technologies for agriculture within single review paper has been covered.

In the rest of the paper: at the outset, in Section II, the existing literature associated to pest monitoring and some disease detection work already done in Europe have reviewed, India, Saudi Arab, and Sri Lanka. Also, the comparison of the characteristics of different sensing technologies with reference to their implementation is presented. In Section III, analysis of the use of sensing technology for pest monitoring and detection is presented. Finally, in Section IV, a summary of the paper and future work plan is presented.

II. Work-Related using WSN in Pest Monitoring and Detection

Mostly the farmers do not like to use chemical on crops. It is for this reason most suitable pest management strategies should be designed based on accurate information about pest and disease. Normally, detection and identification of pests is the farmer's fundamental responsibility for which he relies mostly on his visual judgment randomly. The fields are huge and the farmer cannot cover the overall fields at a time. Farmer often reaches pest infestation too late. Some automatic detecting system is desired for quick assessment of pest infestation in an early stage. There are two different agriculture domains in which we are currently using wireless sensor network.

A. Pest Detection

Pest detection directly through WSN: in which we use an acoustic sensor. We rely heavily on acoustic devices. This is non-destructive, remotely operating and also very useful for automatic detection of hidden insect infestation.

In [4] authors proposed a monitoring system to detect caterpillars of red palm weevil (RPW) through acoustic sensor devices. The authors mentioned that acoustic detection of this pest is the best and most cost-effective solution among all. The proposed system is to monitor the sugarcane crop. The system is based on an acoustic sensor for monitoring the produced sound level of the Pest. Whenever the sound level crosses the defined threshold, it will make the farmer to take the notice of the specific area where the infestation is occurring. By using this technology, the farmer's job to go to each and every part of the crop and perform survey can be reduced significantly.

The acoustic sensor node will be connected to the base station to which each sensor will transmit the noise levels whenever the noise level crosses a predefined threshold level. The base then transmits the information to the control room computer which indicates to the farm where the infestation is occurring so that the necessary action can be undertaken. After successful identification, a farmer can take the necessary measures to spray insecticides over the crops. This detection will also help the farmers to curb infestation at a very early stage and consequently it reduced the high percentage of annual destruction of Sugarcane crops. The proposed monitoring method may cover relatively a big area with low energy consumption.

In [5] the authors proposed an efficient protection mechanism of palms from RPW larvae. The feeding habit of the RPW is concealed, very much like a termite in wood. They can be detected acoustically by the noise emitting from them. Normally, the infestation is detected at the last stage only and when the farmer comes to know the recovery time is almost over and a plant’s one foot in the grave. In the detection phase, the sensor with their propagating modules (transceiver) is attached to a plant and connected to the network by accessing nearby access points. In proposed topology, every access point is connected and receives information from 8 other devices located in its radio range.

In [6], the writers presented a solution for detection of date palm tree (Phoenix dactylifera L) hidden infestation by their designed sensor. They used acoustic sensors to detect the presence of red date palm weevil (RDPW) pest in an early stage; which is usually considered a damaging insect, normally called the red date palm weevil (Rhynchophorus ferruginous) Oliver. They recorded acoustic emission produced by the RDPW that infect date palm trees and then used signal processing method to analyze it. Special probe holding acoustic sensor is injected within the stem of a palm tree to listen and record their voice. It can record the sound of its early stages of life which is known as larvae. In the larval stage of RPW insects perform so many noisy activities like feeding trunk and chewing generating the noise at maximum level. Our recording device near insects works round the clock and records the sound produced by insects easily.

TABLE I. PEST MONITORING AND DETECTION SYSTEMS WITH THE SENSOR NETWORK

Author	***Sensor Generation***	***Major Contributions***	***Treatment Suggested***	***Communication Technologies***	***Price***	***Processing Tool***	***Testing Country***	***Sensing***	***Pests Type***	***Crop***
[4] Srivastava 2013	2nd	Create a pest monitoring & control system using WSN.	Yes	ZigBee / FFT Waveform Monitor	Medium	FFT Waveform Monitor	India	Acoustic	Borer and White flies	Sugar Cane
[5] Srinivas 2013	2nd	Designed a Prototype to prevented Palm plantation from RPW Larvae.	Yes	FM Transmitter	Medium	MATLAB / SP tool and NS2 Simulator	India	Acoustic	Palm Weevil	Coconut Palm
[6] Al-Maine 2007	1st	They used acoustic sensors with some special probe with analysis of the signal processing method.	Yes	Sony DCD T-10	Medium	MATLAB	Saudi Arab	Acoustic	Palm Weevil	Date Palm
[7] Miguel (2013)	2nd	Designing of a self-operative bioacoustics sensor that can be installed within a tree and capable of capturing audio signals, for a long period of time.	Yes	ZigBee with Dongle	Medium	Sensor Deployment software with Laptop along with GPS Support.	Spain	Acoustic	Palm Weevil	Red Palm
[8] Sriwardena (2008)	1st	Develop a portable, smart and efficient acoustic device for monitoring.	Yes	Data Transfer manually	Medium	PC with MATLAB Facility.	Sri Lanka	Acoustic	Palm Weevil Larvae	Red Palm
[9] Ferro (2013)	2nd	Custom made Housing to catch snails with photoelectric sensor operated by solar cells, an Arduino FIO board with ZigBee device for communication	Yes	Arduino board with ZigBee Device	High	PC, Snail Shelter prototype with IR Camera in Field	Spain	Acoustic	Snail pest	Many Crops
[10] BV Laar (2002)	1st	Developed ultrasound gate hard drive recording system that can measure sound activity from 50 Hz to 250 KHz	Yes	Data Transferred Manually	Medium	LAAR/Avisoft SAS Lab pro	Germany	Acoustic	Palm Weevil	Palm Trunks
[11] Chandan K (2018)	3rd	Proposed a system for paddy crop field insects with the help of Drone and an Optical sensor with MATLAB image processing tool.	Yes	Local and Cloud Servers	High	PC with MATLAB / Image Processing Software.	India	Optical	Paddy Field Insects	Paddy Crop Fields
[12] Mankin (2016)	2nd	Acoustic signal collected from trees and orchard to perform analyses on their spectral and temporal pattern of sound impulses and develop their signal analyses to detect them.	Yes	Data Transfer manually	Medium	Signal Analysis Program (DAVIS)	Saudi Arab	Acoustic	Red Palm Weevil and Scarabaeidae.	Date Palm

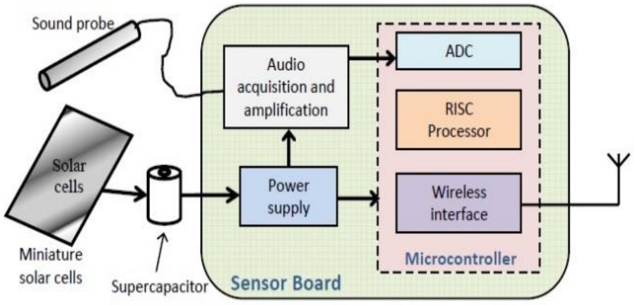


Fig. 2. Block diagram of proposed RPW bioacoustics Sensor. [7]

In [7] entomologists detected Red Palm Weevil (RPW) acoustically in the field. They perform experiments to detect (RPW, *Thnchophorus ferrugineus*) which is critical to detect early (shown in Fig. 2). Authors proposed a system which is able to monitor and record acoustic emissions of the adult Palm (*Thnchophorus ferrugineus*). They used an audio probe, which has an acoustic sensor that is in charge of acquisition of noise generated by RPW. It is also capable of amplifying the captured voice signal and to make it possible to process by an algorithm. Low power processor, which runs the algorithm and a wireless interface to send detected signals remotely. First, the system digitalizes the audio signals and then execute in real time to detect RPW activity or its presence. It is small in size that could be mounted on the appropriate location of any particular palm tree.

In [8] researchers described a handy and, a smart audio device with its possible use in the field for diseased palms. Their device contained a sensor that could be to mount on the palm tree. It has a facility to get the sound of red palm weevil larvae. It comprised of an automatic processing element, a processing unit to process the acquired sound and the earphone set to receive the audio output through recording device and the hearer. This is a battery-driven portable user-friendly device.

In [9, 11] scientists reviewed the current research and recent development states on Red palm weevil (Rhynchophorus ferrugineus). Oliver and monitored them to get the information about their early infestation. Some intensive efforts have been put into the development of detection through visual and acoustic methods. Pros and cons of all methods have been presented here with their comparisons. It is also concluded that considerable efforts are still required to improve the efficiency and sensitivity of existing acoustic methods with another tool.

In [10] authors developed a handheld detector device with some special acoustic sensors to probe. The developed device is specially designed to detect a tiny sound vibration which is a usual activity of RPW. It is tested also on German beetle species living in the wood. Laar [10], invented “ultrasound gate hard disk recording system” and with this device, he measured sound activity from 50 Hz up to 250 kHz.

In [12] researchers developed an acoustic model to capture the sound of fruit flies (*Ceratitis capitata*). They conducted a large number of tests with this acoustic model on the Mediterranean fruit fly. It was tested in high noise as well as in the worst traffic environment. The only male strain of Ceratitis capititata was used in experiments. The first experiment was conducted in a quiet environment with a group of 25 males in 20- by 21- by 22.5 screened cage. A Sony camcorder (DCR-TRV27) was used to get a distant observation of verification flight. Then some tests for flight detection had been conducted in noisy surroundings, which contained more male with relatively big cage. They were visually monitoring their flights in the field cage throughout recording sessions. The recommended flight monitoring system is a handy model apparatus modified from a mobile-pre USB preamplifier along with the audio interface. It may provide 40-70db amplification through two variable adjustment gain control. Signal input provided to an AT 803B unidirectional lavaliere microphone. The signal was transferred to a laptop by a universal serial bus and processed by customized software which runs under MATLAB. The produced noise by insect (C. Capitata) flights through a microphone is naturally concise and easily identifiable.

In [13] scientists detected adult and larva of *Oryctes chinoceros* acoustically in dead and alive palm trees in island territory of Micronesia. They also monitored and detected *Nasutitermes luzonicus* Oshima and some sound generating tiny insects. A large and active *O. rhinoceros* usually generates low frequency, long duration sound impulse trains. There are some soaring frequencies low impulses trains generated by *N. luzonicuz.* Their unique ghostly and sequential pattern of producing noise made it possible to identify suppressing surroundings sound easily.

In this part, some early pest infestation techniques as mentioned in Table has been reviewed 1. It is also compared with respect to their target crops, with those sensing technologies heavily used in agriculture field by means of WSN. Early detection of infestation plays a vital role in the recovery of attack.

B. Disease Detection

In this portion illustrates crop/plant disease detection with the help of WSN and image processing with its analysis that involve color histogram, edge detection, and some other processing tools.

In [14] authors used the internet of things (IoT) technology to design a platform for detecting diseases. Their main focus was on general diseases. They use an IOT to turn it into the key system to acquire data with communication because it is the most important technology among others. In their system; detecting sensor is used to obtain data that is compared and synchronized. After analyzing the collected data they finally carry out an immediate action without any human involvement. They applied this process to the whole data related to pest plant disease and insect pests. They used global information system (GIS) software to manage and present data for linking it to the exact location. WSN and ZigBee with land mobile (GSM) are largely used networks in the precision agriculture field. They designed a platform which includes administrators, experts and common visitors with computers and mobile phones. It also includes an information system for agriculture disease and insect pests' disaster information monitoring system.

TABLE II. DISEASE DETECTION WITH WIRELESS SENSOR NETWORK

Author	Disease	Crop	Technology / Tool	Experiment Country	Contribution
[14] Prevostini (2011)	Flavescence Doree	Grapevine	WSN to take air temperature and other	Germany	Prototype of a small Wireless sensor network as well as on an algorithm that could be used to calculate the spread of the disease vector.
[15] Wang J (2014)	General Plant Disease	General	Multimedia WSN with Image Processing tools.	Korea	Perform a Survey for applications of Image processing in Disease detection and conduct an experiment with multimedia wireless sensor network and other image processing tools to detect specific disease.
[16] 2014Dater S. ()	Downy Mildew	Grapes	WSN with Web Based GPRS	India	Developed a Web application which provide a forecast on the basis of weather parameters like humidity, temperature and wind speed
[17] Tripathy A (2013)	Groundnut Bud Necrosis Virus (BNV)	Groundnut	WSN with Data Mining Technique	India	They tried to understand the hidden relationship among interrelated disease / Pests and weather parameters and develop a web base system for forecasting groundnut disease.

The system is capable to capture an image and send them to a remote control station with specific event stipulated through an application. Basically, these image sensors monitor and count inhabitants (pests) with relatively advanced resolution. However, there seem to be no intelligent image processing activities. During this monitoring process, no human intervention is required. According to the authors, there is a significant reduction in monitoring cost as well.

The main focus of this work is red palm weevil (RPW, Rhynchophorus ferrugineus. Oliver) but it is not limited to that pest. It could be used to monitor many other similar pests. A trap monitoring process which works on unattended mode has some extra benefits such as it reduces the monitoring cost, it is programmable and has high resolution monitoring data. In addition, real-time data can be retrieved at any time by the web portal.

There have been a number of valuable studies to monitor pest insect using latest technologies. However, none of these studies is able to provide a self-sufficient information system totally relied on inexpensive image sensors covering areas with very low energy utilization. High scalability with low power consumption made it possible to deploy this system both in greenhouses and larger plantations.

It is also used for several kinds of insects instead of some specific insects. Using an image recognition algorithm; that is capable to identify RPW insects with a higher success rate up to 95%. The system is smart and its corresponding Metadata, timestamp, GPS coordinates, and results etc. are duly saved in the main monitoring station. Anyone can have access to the data in real time through the internet which is obtainable from the location of the main control.

In [15] authors conducted a survey to detect crop disease using image processing with the wireless multimedia sensor network. Pest related diseases have now become a recent predicament. It is the main cause of the significant decline in product quality and quantity as well. Nowadays the efficient and accurate detection of diseases has become one of the major issues in the agro-industry. Scientists used machine vision approach, with some morphological features like size, shape, texture with pest's color. Also, they applied some location-based attribute to monitor and detect infected plants and leaves.

In [16] authors proposed a solution to detect specific disease (downy mildew) in grapes in India at a very early stage. The current system that detects disease (downy mildew) is based just upon collected update climate information. In the proposed system architecture, there is a remote node which is inexpensive and the user has to keep secure only an isolated node (which is remote) as a replacement for the main node which is located in control station. A central server can be accessed through web applications to get all details of current weather conditions and disease forecast, which depends strongly on a climate of the farm.

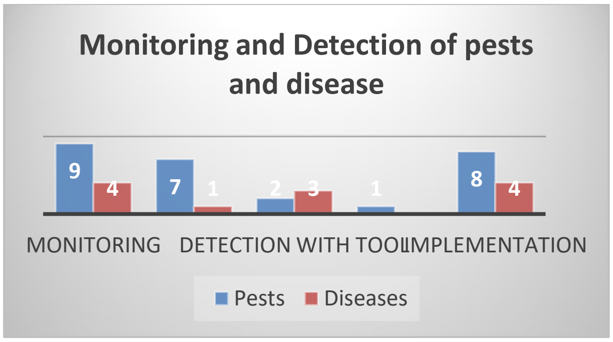
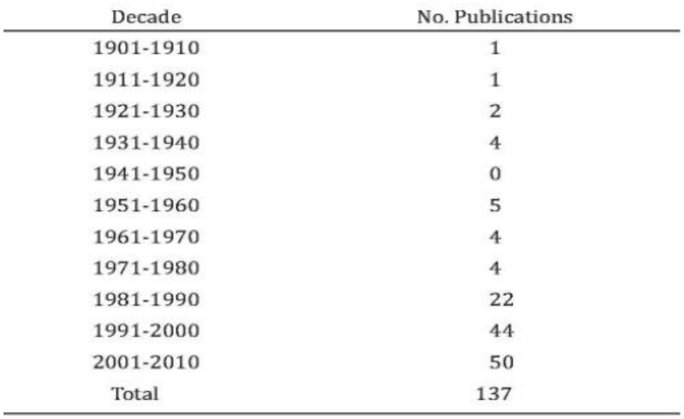


Fig. 3. Available features of reviewed applications and systems.

The existing systems which have been described are mostly implementation based but some are simulation based too. Some authors [6, 8, 11, 12 and 15] performed implementation with different types of data or signal processing tools as shown in figure 3. The system proposed in [4, 7, 9,10,14,16 and 17] are also proposed design with real deployed.

In [17] the authors' reviewed the weather relation with pest and crops along with data mining and WSN. They focused on peanut crops pest and disease interaction in India. They conducted an experiment to examine the crop with climate and insect relation using WSN. They also reviewed independent pest and disease dynamics of peanut crops. To turn the data into useful information they used a smart technique of data mining to draw relation among crop/ pest/disease and climate field. They tried to comprehend the concealed association between interrelated insect and disease with climate parameters. In the end, they developed a collective prediction model, which could help the farmer to improve measures with this prediction model in the future.



Decade	No. Publications
1901-1910	1
1911-1920	1
1921-1930	2
1931-1940	4
1941-1950	0
1951-1960	5
1961-1970	4
1971-1980	4
1981-1990	22
1991-2000	44
2001-2010	50
Total	137

Fig. 4. Insects detection and monitoring acoustically, Publication since 1900. [13]

As mentioned in Table 2, some existing disease detection techniques which are used with the help of a type of WSN are reviewed. Early detection of diseases is as important as pest detection to prevent heavy crop loss.

In [13, 18, 19 and 20] researchers have been highlighted research on acoustic detection of insects with their management and control techniques (see Fig 4). Also, it has been explained how we use olfaction, vision, noise, and hearing capabilities of insects as their destroyers. It is a remote, non-deleterious, automated observing tool for farmers and researchers to find concealed 'insect invasion'. A few milligrams of pheromone in the right context can attract a male moth to its mate. Similarly, a flash of penlight can attract a firefly male from 30 meters.

It is very old technology and we have been using it since the start of the century as mentioned in Table 3. In recent year's various kinds of sound catching equipment are being widely used in the market to monitor crops. The efficacies of acoustic devices are limited. They are only used for sound generating cryptic insects and estimate population density while silent killers are beyond its domain. Success rate depends on the different type of parameters such as sensors, the range of frequencies, substratum structure, the correlation among the substrate, time and duration, crop field with size and behavior of insects, and also the distance between sensor and insect.

We got significant success in the field of passive acoustic devices to monitor grain, palm/wood insets i.e. Red palm weevil. The microphone is useful for airborne signals while vibration sensors are very useful for those signals which are produced in a solid substrate. Ultrasonic sensors are practically very effective to detect wood-boring pests [13, 20 and 21]. Complexities in distinguishing sounds of target insects are major restrictions in using acoustic devices. Currently, some other tools like signal processing and smart sensors have greatly increased the acoustic use and its reliability.

Some early infestation techniques which are heavily used with the help of WSN in the agriculture field are reviewed. Early detection of infestation is important to the recovery of attack. So it, not just WSN, but also so many other techniques such as image processing, canny edge with color histogram [22 and 23] and laser induces breakdown spectroscopy [24] are used.

III. Analysis of Sensing Technologies for Agriculture Sensor Networks

Every remote technique which could be used in wireless sensor network depends on the electromagnetic propagation or acoustic energy between the sensor and a target pest. It must use some motion/vibration sensors, which are highly sensitive as they can capture even tiny pulsation of pests in the field. It is also could use such acoustic sensors which are able to record some special frequencies produced by pests. Although, it is difficult to identify directly by using these sensors yet at least it is possible to point out the presence of some insects in a specific area of our field. Image analysis and processing with other advanced computer technologies also exist to identify pests directly. Some digital automated identification systems (DAISY) are used to identify special pests [25, 26, 27, 28, 29 and 30]. A detailed review of every aspect of the sensing of the insects, geographical terrain is the scope of this paper.

Favorable/unfavorable climate for target insects is out of the scope of this paper..

In the market, there are various sensor products for agricultural monitoring i. e MOTE, Field Server, SUNSPOT [63, 64 and 65]. There are advanced products with wireless IR, ZigBee (IEEE802.15.4), ultrasonic [33-34] technologies are also used.

Our main concern is on the development of locally made, an affordable device which would be a good and adaptable addition in industry. Many technologies mentioned in this paper have been verified in past and will be used in the future as well (see Figure 5). Some other technologies also exist but yet to be applied in entomology. Following are some possible sensing technologies that can be used to detect and monitor cotton pest specially bollworms in Pakistan.

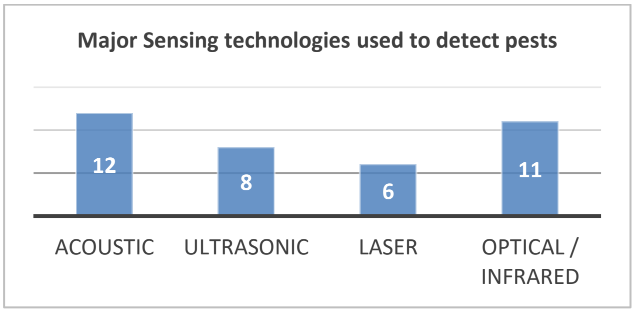


Fig. 5. Major Sensing technologies used in agriculture reviewed by this paper.

A. Acoustic Detection

In theories, there is a rule of thumb that distant monitoring must contain the transmission and reception of electromagnetic waves from the sensor to target and vice-versa. Researches [4, 5, 6, 7, 8, 9, 10, 11 and 12) have done enough work see table 1 to develop passive acoustic devices to detect insects that are out of our site, and are found hidden in stored commodities as well as around plants, timber, and fruits. Insects are detected usually through low strength (0.5- 150 kHz) incidental sounds which they generate during flying, feeding or calling their opponent gender. There are many factors that are involved in the use of acoustic devices such as interference and noise ratio, background noise, distortion and attenuation during travel in the medium and uniqueness of voice example from a non-target and separate organism [35]. Sensing elements include transducers (condenser / piezoelectric microphones), vibration transducers, mobile or fixed in any hard media such as timber, stone and underground. Some atmospheric sound detection and ranging (SODAR) devices are also provided meteriological data to monitor pests migrations [36]. Similarly research has been done in [37] where simple non doppler sodar machine is used for verification and the actual observing height of the night climate data. It also determines the best altitude for netting above ground level accordingly.

TABLE III. SESNING TECHNOLOGIES AND THEIR CHARACTERISTIC'S COMPARISON [4-9,18,19,22,29,35-37,40-43,49-61 AND 62] .

Sensing Technologies	Range	Accuracy	Cost	Outdoor Performance	Sensitivity	Complexity of Support Electronics
Acoustic	Medium	Medium	Low	Medium	Medium	Medium
Optical	Medium	Medium	Medium	Medium	Medium	Low
Laser	High	High	High	High	High	Low
Ultrasonic	High	High	Medium	High	High	Medium

B. Ultrasonic

Another technology called ultrasonic sensing has also been used in the past and can be used in the future as well to monitor crops and pest closely. We have been using ultrasonic technology in crop production since 1988 [38]; where researcher used commercial ultrasonic range transducer to measure some specific parameters. This system was mounted and tested with an air blast sprayer and its results were used to optimize the sprayer in future. Later on the same scientists [63]; investigated spray volume saving using an ultrasonic measurement and results varied greatly depending on target crop morphology.

Group of researchers [39, 40, 41, 42 and 43] conducted some studies in different aspects of ultrasonic sensing, its applications and drew a comparison between a laser and ultrasonic transducers for crop constraints and canopy volume measurements of citrus trees. The laser sensors performed relatively better than ultrasonic since they had a higher resolution. Author invented a model sprayer which could calculate the size of the target and approaching density using ultrasonic sensing [44].

Segment array is used and also suggested to use [45] with the combination of ultrasonic systems which might be very helpful to track small insects in short range (e.g. Aphids and whiteflies) those are in the flying range of a crop shades. In their studies, a web of more than a dozen ultrasonic transmitters emits pulses of over 40 KHz. These pulses are delayed by their phase to make it possible for the system to cover entire volume which needs to be sensed (2.5 long x 1.5 width x 2 hight in meters). Echos those returns after hitting to the bugs are picked by multiple installed observer which also boost their energy and send them to the digital signal processing capable machine. The actual position of insects with movement direction within the tracking space could be predictable and also displayed in real time. Until now the complete system has not been operated below the real condition of the field but a small individual device was tried to determine how it copes with more than one objective (Target) with high background noise.

These types of studies are very important and vital to develop a smart, portable ultrasonic device to monitor insects/pest in the field. Big gap still exists in this domain to predict crop disease and presence of insects in the agricultural fields.

C. Laser

The LASER is used since 1970, in many areas, however, it was late 80's when the laser was used for forest biomass detection and crop production [46], in which scientist implemented an airborne pulsed laser system to access forest biomass and temper volume. The same scanning techniques could be applied to detect pest and tiny bugs. Later on, in the early '90s [47], a laser altimeter was used to quantify vegetation properties and their results showed a variation in the canopy heights between two to six meters. Collected data is connected and compared to similar data which is gained from other methods. This study also exposed to us that a similar application can be applied to detect disease in any type of crop. Some other laser [48], was also used with the combination of Lidar and satellite imagery which was also very useful to monitor changes. In the recent era, a newly developed laser sensing system was applied [49] to citrus crop to measure height and width of trees and covering capacity. The system has been tested to calculate its resolution and high accuracy and with less than 5% error. Further, the same authors in [50], implemented scanning system based on the laser to calculate vegetation thickness. Results also revealed good occurrence and of less than 3% avg. the coefficient of variation (CV). Recently [51] also utilized a laser system with the implementation of rangefinder technology to estimate the location of precise foliage factors, such as the height of the plant, its coverage besides biomass solidity which could be a major factor in optimizing crop harvesting method.

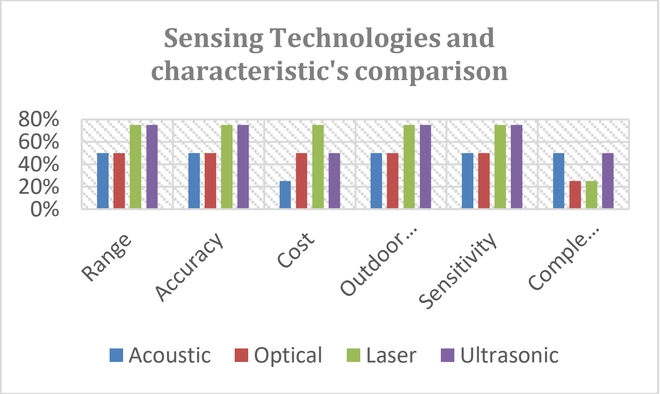


Fig. 6. Sensing Technologies and their unique features [4-9, 18, 19, 22, 29, 33, 35-37, 40-43, 49-61 and 62]

These studies are very useful in future developments. Also, use some new technologies and build a prototype which may help us to protect crops from insect and disease can be used.

D. Optical Sensor with Trapping

Optical sensors are also used to get the exact time of entry of pests into the trap as well as exit recording. Such as in [36, 52 and 53] it was some extra facilities provided by them to assign captured insects to different classes according to their morphology. It is due to different momentary of insects over an irradiated exposure of light scattered. Sometimes this might show good enough to substitute that trap altogether [37].

a) Night Vision Devices

These devices are used to observe and monitor insect movement at night when human vision cannot work adequately. The capabilities of these devices can be amplified by some extra devices such as a telescope. In the night vision device, light consists of photocathode which later releases electrons! The total number of following lights is significantly increased by some form of light voltage, as a result, some electrons are used to reproduce the duplicate on phosphors canopy [38]. These devices are divided into three generation which extends the operational life of this device. Occasional entomological uses of graphic intensification tool mentioned in [54, 55 and 56] i. e. specific Helicoverpa moth low elevated flights are tracked through an observer. The observer uses night vision goggles and inferior red light illuminators. They ride on special 4 wheeled vehicles. Their experiment proved that when observer follows the moth, he is able to see them up to 100-meter heights. [55]

b) Optical-Electronics Devices

Some more specialized devices for insect monitoring and specialized optoelectronics devices are crossed- beam infrared detectors [57 and 58]. When pests move over-designed capture capacity, they are sensed. Frequencies of their wing beat are also recorded. These systems are specifically designed for Spodoptera exempta and Helicoverpa armigera moths monitoring while they fly underneath detector shadow. Farmery's apparatus is not useful in dusk and daylight.

This article showed the wide variety of researches to monitor and detect pests and disease so far. However, some techniques with their pros and cons are mentioned and also an overview of crop insects are given, their, favorable weather parameters and disease indications. The standalone technique can be very useful but research workers also use the combination of sensing technologies. Such as, the pheromone light traps have been used.

We have drawn a comparison of existing technologies as a result of their review (See table 3 and Figure 6.) that are already being used in sensing of pests or diseases in wireless sensor networks. As a result of the present review, we can go further and develop more accurate sensors to target pest or diseases. Our recommendations on the use of sensor technology are presented in the next section.

IV. Conclusion and Future Work Plan

A comprehensive assessment of numerous solutions and efforts has been presented in this paper as regarding wireless sensor network in agriculture pest monitoring and disease detection domain. Since the pest monitoring field is based on the situation and environment so the potential of using sensors and WSN is very high. We conclude the following notable issues which are faced in the developing Asian countries:

- Lack of technology awareness in farmers
- The extra cost is involved
- Prospect solution is complex to adopt or implement and require full technical support in all aspects
- Difficulties in making a general solution for different problems due to variant situations
- Main research works present the solutions in parts: some are focused on data processing and storage while others are data acquisition or context modeling
- Complex or sometimes unavailability interlinks parts reduce the adaptability ratio of solution

Therefore, there is a need to increase the application of WSN and sensors-based solution on an industrial scale. For this purpose, the following drastic steps should be taken

- Development of local, resilient and extreme low-cost sensors
- Crop-based generalized solutions to solve different problems which may include single or combination of sensing technologies
- There is an opportunity for us to build a complete common framework from weather data acquisition for modeling and implementation to conclusion support
- Segment solution must be encouraged but we have to work on comprehensive and detailed compatible interlink procedures which can make our planned solution comprehensive and accurate.

In our future work, it is planned to investigate some more solutions specifically for developing countries. We are working on a special locally made wireless sensor mote to monitor and detect bollworm in a cotton crop, as it is mentioned above. Digital image processing is one of the options that can be used with the above-mentioned prototype module in order to identify pests more accurately. Also, it is intended to develop a high-level collaboration among agriculture industry, technology stakeholders and academia that will be beneficial for farmers of the underdeveloped agriculture industry and overall the country.

References

[1] Vennila S., Cotton pests, predators and parasitoids: Descriptions and seasonal dynamics, Central Institute of Cotton Research, Research Notes. India, 2013.

[2] Robert E. P. Fundamental of Applied Entomology. 6th Ed. The Macmillan Company; Printed in the USA. pp. 343, USA, 1971.

[3] Agriculture Knowledgebase, accessed in 2014, http://www.seetoptens.com/cottoninsectspakistan

[4] Srivastav N, G. Chopra, P. Jain, B. Khatter. Pest Monitor and control system using WSN with special reference to Acoustic Device; ICEEE, India, 27th Jan. 2013.

[5] Srinivas S, Harsha K.S, Sujatha A and N. Kumar, Efficient protection of palms from RPW Larvae using WSN: IJCSI Vol 10, Issue 3,.India, 2 May. 2013.

[6] Al-Manie M. A., M. I. Al-Khanhal. Acoustic detection of the Red Date Palm Weevil, International Journal of Electrical, Robotics, Electronics and communication Engineering Vol: 1 No. 2, KSA, 2007.

[7] Miguel M. R, H. M. Gomis, O. L. Granado, M. Perez, A. Marti and J. Jose Serrano. , Journal of Economic Entomology, 2013. On the design of a Bioacoustic Sensor for the early detection of the Red Palm Weevil, 1706-1729, Sensors 2013

[8] Siriwardena K.A.P, L.C.P. Fernando. Portable acoustic device for detection of coconut palms infested by Rynchophorus ferruginous, Crop Protection 25-29., UK, 2010

[9] E. Ferro, V.M. Brea, D. Cabello, Pl Lopez, J. Iglesias, J. Castillejo, "Wireless Sensor Mote for Snail Pest Detection – IEEE. 2014

[10] Laar B. V, The bioacoustics detection of the Red Palm Weevil, Gut Klein Gornow, Germany. 2002

[11] Chandan K. S. P. K. Sethy and S. K. Behera, " Sensing technology for detecting insects in a paddy crop Field Using Optical Sensor, Springer Nature Singopore, 2018.

[12] Mankin, RW and Al-Ayedh, HY and Aldryhim, Y and Rohde, B. "Acoustic detection of Rhynchophorus ferrugineus (Coleoptera: Dryophthoridae) and Oryctes elegans (Coleoptera: Scarabaeidae) in Phoenix dactylifera (Arecales: Arecacae) trees and offshoots in Saudi Arabian orchards", Journal of economic entomology, vol 109-2, Pages 622-628, Oxford University Press, 2016

[13] Mankin R. W., D. W. Hagstrum, M. T. Smith. Perspective and Promise: A century of Insect Acoustic Detection and Monitoring, American Enologist Springer. FL, USA, 2011

[14] Mauro P. "Wireless Sensor Network for Pest Control", Commission for Technology and Innovation (CTI), 2011

[15] Jinpeng W, Yibo C. Jean-Pierre C. An integrated Survey in Plant Disease Detection for precision agriculture using Image Processing and wireless multimedia sensor network (ICACEEE 2014) France, 2014.

[16] Datir S., Sanjeev W. Monitoring and detection of agriculture disease using WSN. IJCA (0975 -8887) Vol. 87. India, 2014.

[17] Tripathy A.K, J. Adinarayana, D. Sudharsan, K. Vijayalakshmi, S. N. Merchant, U. B. Desai. Data Mining and Wireless Sensor Network for Groundnut Pest / Disease Interaction and Prediction –A Preliminary Study ISSN 2150-7988 Volume 5, India, 2013.

[18] Azfar S., A. Nadeem. Pest detection and Control Techniques using wireless sensor networks, Journal of Entomology and Zoology studies, JEZS; 3 (2) 92-99, INDIA, 2015.

[19] Walker. T. J. Acoustic methods of monitoring and manipulating insect pests and their natural enemies, Entomology and nematology department, University of Florida, USA, 1996.

[20] Alexander R. D. Sound production and associated behavior in Insects, The Ohio Journal of Science 57 (2): 101, USA, 1957

[21] Bohmfalk G. T., R.E. Frisbie. Identification, Biology and Sampling of Cotton Insects, the Texas A&M University System. (Cotton Pest Study), USA, 2011.

[22] Shital B. , Plant Disease Detection techniques using canny edge detection & Color histogram in Image Processing, International Journal of Computer Science and Information technologies Vo. 5 (2), 1165 – 1168, India, 2013

[23] Jongman C. junghyeon C, Mu Q, Automatic Identification of Whiteflies, aphids and thrips in greenhouse base on Image analysis. International Journal of Mathematics and Computers in Simulations. Korea, 2007.

[24] Farooq W. A., Application of Laser Induced Break down Spectrophy in Early infestation of Red Palm Weevil: (Rhynchophorus ferrugineus) Infestation in Date Palm ", Plasma Science and Technology, Vol 17, No. 11, KSA, 2015

[25] Wang N., N. Zhang. Wireless sensors in agriculture and food industry-Recent development and future perspective (Review), Computer and Electronics in Agriculture, 50 1-14, Canada, 2006

[26] Lee W.S., V. Alchanatis. Sensing Technologies for precision specialty crop production, Computer and Electronics in Agriculture 2-33. FL, USA, 2010

[27] Dorge, T., J. Michael. Direct identification of pure Penicillium species using image analysis, Journal of Microbiological Methods, Elsevier, UK, 2000.

[28] Watson A.T, M. A. O'Neill, I.J. Kitching, Automated Identification of Live moths using Automated Identification System (DAISY), Systematic and biodiversity 1 (3), 287-300, UK, 2003.

[29] Faria F. A, F. Perre. Automatic identification of Fruit flies (Diptera: Tephritidae), Journal of Visual Communication and Image representation, SP, Brazil, 2014.

[30] Mayo M., A. T Watson. Automatic Species Identification of Live Moths, Dept. Computer Science, the University of Waikato, Hamilton, pp 58-71, New Zealand,2007.

[31] Gangurde P, Manisha Bhende, A review on precision agriculture using WSN, International Journal of Engineering trends and technologie (IJETT) – volume 23 No. 9, India May 2015

[32] Awasthi A and S.R.N Ready, Monitoring for precision Agriculture using wireless sensor network – A review, Global Journal of computer science and technology network, Web and Security, Vol. 13, No 7, Ver. 1, Global Journals Inc USA, 2013

[33] Aqeel-ur-Rehman, Abu Zafar, Noman I, Zubair Shaikh, A review of wireless sensors and network application in Agriculture, Computer Standards & Interfaces, 36, Elsevier, 2014

[34] [34] Commercial Sensor's web portal, http:// www.zigbee.org , Accessed in May, 2016

[35] [35] Wei J., M. Salyani, Development of a laser scanner for measuring tree canopy characteristics: Phase 2. Foliage density measurement." Trans. ASAE 48 (4), 1595-1601, Michigan, USA, 2005.

[36] Ehlert D., H. Horn. Measuring crop biomass density by laser triangulation", Compute. Electron. Agric. 61, 117-125, Netherland, 2008.

[37] Skatulla U., E. Fiecht. Observations of the flight behavior of Lymantria monacha L. (Lep. Lymantriidae) to pheromone baited traps. Journal of Applied Entomology 119. Freising, Germany, 1995.

[38] [38] Reynolds D. R., J.R. Riley. Remote-sensing, telemetric and computer based technologies for investigating insect movement: A survey of existing and potential techniques", Elsevier Journal of Comp. and elect in agriculture. UK, 2002.

[39] Manijeh K, A. Deljoo, "A wireless sensor network solution for precision agriculture based on Zigbee Technology", Scientific Research Journal/ Wireless Sensor Network, Vol.4, 25-30, 2012.

[40] Giles D. K, M. J. Delwiche, Electronics measurement of tree canopy volume, Trans. ASAE 31 (1), USA, 264-272, 1988.

[41] Giles D. K., M. Delwiche, and M. Dodd. "Sprayer control by sensing orchard crop characteristics: Orchard architecture and spray liquid saving. J. Agriculture Eng. research, 43, 271-289., Davis, USA, 1989.

[42] Molto E., B. Martin, Pesticides loss reduction by automatic adaptation of spraying on globular trees, J. Agri. Eng. Res. 78 (1) 35-41.Valencia, Spain, 2001

[43] Solanelles F., A. Escola. An electronic control system for proportional pesticide application to the canopy volume in tree crops", proceeding of joint congress on IT and Agriculture, EFITA / WCCA July. Spain, 2005

[44] Gil E. A. Escola," Variable rate application of Plant protection products in vineyard using ultrasonic sensors", Crop protect. 26 (8) 1287-1297 2007.

[45] Tumbo S. D, Masoud S. Investigation of Laser and ultrasonic ranging sensors for measurements of Citrus Canopy Volume", Appl. Engineering in Agriculture, Vol 18 (3): 367-372, Florida, USA, 2002.

[46] Balsari P., G. Doruchowski, A system for adjusting the spray application to the target characteristics", Agri. Engi. Intl. Vol. X.,Italy, 2008

[47] Isard S. I., S. H. Gage. Flow of life in the atmosphere: An airspace approach to understanding invasive organisms. Michigan State University Press., USA, 2001

[48] Nelson R., Estimation forest biomass and volume using airborne laser data", Remote Sens. Environ. 24 (2), 247-267, 1998.

[49] Ritchie J. C. Measuring canopy structure with an airborne laser altimeter, Trans. ASAE 36 (4), 1235-1238., Michigan, USA. 1993.

[50] Nilsson M., Estimation of Tree heights and stand volume using an airborne LIDAR system", Remote Sens. Environ. 56 (1), 1-7. 1996.

[51] Wei J., M. Salyani. Development of a laser scanner for measuring tree canopy characteristics: Phase 1. Prototype development. Trans. ASAE 47 (6), 2101-2107, Michigan, USA, 2004.

[52] Mankin R. W., Acoustic detection and identification of insects in soil", Proceeding of the 16th International congress of Acoustic and the 135th annual meeting of the acoustical Society of America, pp 685-686, USA, 1998.

[53] Hendrik D. E, Development of an electronic system for detecting Heloothis sp. Moths (Lepidptera: Nottuidae) and transferring incident information from the field to a computer." Journal of Economics Entomology. USA, 1989.

[54] Waggington K. D., The effects of season, pertaining, and scent on the efficiency of traps for capturing recruited honey bees (Hymenoptera: Apidae). Journal of Insect Behavior 9. USA, 1996.

[55] Schouest L.P. Automated pheromone traps show male pink bollworm (Lepidoptera: Gelechiidae) mating response is dependent on weather conditions". Journal of Econ. Entomology. 87. 965 – 974, USA 1994

[56] Lingreen P. Night vision equipment, reproductive biology, and nocturnal behavior: Importance to studies of insect flight, dispersal and migration", Springer, Germany, 1986.

[57] Lingreen P. D., Flight behavior of corn earworm (Lepidoptera: Noctuidae) moths under low wind speed conditions", Environ, Entomol, 24, USA, 1995.

[58] Fitt G.P., G.S. Boyan. Methods for studying behavior. In: Zalucki, M.P. (Ed). Heliothis: Research method and prospects, Springer, New York., USA, 1991.

[59] Mankin R. W., R. Machan. Field testing of a Prototype acoustic device for detection of Mediterranean Fruit Flies Flying into traps, proceeding of the 7th Intl. Symposium on Fruit Flies of Economic Importance, FL, USA, 2006

[60] Farmery M. J, Optical studies of insect flight at low altitude, Thesis, University of York., UK, 1981.

[61] Schaefer G.W., G.A. Bent. An infra-red remote sensing system for the active detection and automatic determination of insect flight trajectories (IRADIT). Bull. Entomol. Research, pp. 261 – 278, UK, 1984

[62] Kirankumar Y. B, J. D. Mallapur, "Advanced remote monitoring of a crop in agriculture using WSN Topologies", International Journal of Electronics and communication engineering & Technology, Vol, 6, issue 9, pp. 30 – 38, September 2015.

[63] Commercial Sensor's web portal, https:\\www.xbow.com Accessed in May, 2016 ,

[64] Commercial Sensor's web portal http:// www.elab-experience.com Accessed in May, 2016

[65] Sensor's web portal http:// www.tauzero.com /rob-tow/Sun-spots-sensor-networks , Accessed in May, 2016

Improved Discrete Differential Evolution Algorithm in Solving Quadratic Assignment Problem for best Solutions

Asaad Shakir Hameed[1], Burhanuddin Mohd Aboobaider[2], Ngo Hea Choon[3], Modhi Lafta Mutar[4]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka Hang Tuah Jaya
76100, Durian Tunggal, Melaka, Malaysia

***Abstract*—The combinatorial optimization problems are very important in the branch of optimization or in the field of operation research in mathematics. The quadratic assignment problem (QAP) is in the category of facilities location problems and is considered as one of the significant complex's combinatorial optimization problems since it has many applications in the real world. The QAP is involved in allocating *N* facilities to *N* locations with specified distances amid the locations and the flows between the facilities. The modified discrete differential evolution algorithm has been presented in this study based on the crossover called uniform like a crossover (ULX). The proposed algorithm used to enhance the QAP solutions through finding the best distribution of the *N* facilities to *N* locations with the minimized total cost. The employed criteria in this study for the evaluation of the algorithm were dependent on the accuracy of the algorithm by using the relative percent deviation (PRD). The proposed algorithm was applied to 41 different sets of the benchmark QAPLIB, while the obtained results indicated that the proposed algorithm was more efficient and accurate compared with Tabu Search, Differential Evolution, and Genetic algorithm.**

Keywords—Quadratic assignment problem; combinatorial optimization problems; differential evolution algorithm

I. Introduction

There are several specific problems for COPs, such as the quadratic assignment problem (QAP), routing problem (RP), etc. The QAP was introduced by [1] and the model of this problem has been applied in many aspects of life and is famous on campus and in hospital layout QAP is a complex problem that has attracted the attention of researchers since its first formulation [2], [3], [4], [5], and [6]. There are many challenges facing the installation of facilities to location, such as a lack of layout in the buildings, which leads to an increase in computational complexity [7]. The methods that found solutions to the QAP problem were classified into two categories as follows: the category that obtains the exact solution to QAP was called the exact methods, including the bounded dynamic branches and processes, Lagrangian-based relaxation methods, linear and quantitative programming methods. However, in these methods, the size of the problem requires a long calculation period if there are more than 30 methods [8], [9], [10], and [11]. The second category obtains the approximate solution or near the optimal solution with reasonable calculation time and are known as the approximate methods. The approximate methods have been divided into three categories [12]:

- Local Search Algorithm, such as Tabu Search;
- Swarm Intelligence, such as Ant Colony Optimization;
- Evolutionary Algorithm, such as the differential evolution algorithm.

The differential evolution (DE) method is one of the latest evolutionary optimization methods reported by [13]. DE is a global optimizer that relies on population and random space continuously [14]. Due to its efficiency and strength, DE has increasingly become common and has been utilized in numerous fields such as the function of continuous real value and the problems of combinatorial optimization with a discrete decision. In a study, [15] proposed an algorithm regarding the discrete differential evolution (DDE) for computation of the variation of the flow-shop preparation problem. The overall operation of this method was not as efficient as other methods, which could be due to the employment of low mutation probability (0.2). In contrast, the operation of the DDE algorithm was observed to be competitive when using the local search. An earlier work [16] modified the DE to a discrete optimization problem and was applied to solve the QAP.

However, the proposed method, which utilized the property Tabu List, was not able to use the crossover. Hence, the mutant vector directly became a trial vector and can solve nearly all the instances from Nug15 to Nug30 in QAPLIB. Nevertheless, the obtained results were not superior to the max-min ant system hybrid with the random selection and the local search. In another study [17], modified DDE with the local search-based modification using insertion and swap was used. Employment of DDE with local search further improved the results of two types of sparse and dense examples of QAPLIB.

The study aimed to modify the DDE for management of the complex problems while being capable of the exact search space with minimum cost. Moreover, the execution of the proposed algorithm led to enhancement of solving instances of QAP from the benchmark QAPLIB. The remaining of the study has been organized as follows. Section II provides a description of QAP, while section III presents the materials and methods. The computational results have been discussed in section IV, while the conclusion has been provided in section V.

II. Quadratic Assignment Problem (QAP)

QAP is considered as one of the site problems which reduce the momentum within the places of high mobility such as hospitals, campuses, and several facilities to be allocated to these sites by calculating the matrix of distances between location and the flows between facilities. Solving the QAP indicates examining the assignment that reduces the cost of transportation among the facilities. In order to have a QAP instance, visibility and a list of distances of accessible locations and material flow among facilities *(F_{ij})* must be available. Each *N* facility is interchangeable, and there are *N* locations that can only provide for one facility. Moreover, there are *N* facilities set and *N* locations set and for each location pair the specification of distance *(D_{ij})* and for each facility pair, a flow (Fij) is itemized. The difficulty in assigning entire facilities to alternative positions is aimed at minimizing the sum of distances increased by conforming flows [5]. Formally, let and be two *N*N* matrices and let *P* be the set of permutation of $\{1, 2, \ldots, n\}$. Then, the mathematical model of QAP can be written as

$$Min\ f(\pi) = \sum_{i=1}^{n}\sum_{j=1}^{n} F_{ij}\, D_{\pi(i)\pi(j)} \qquad (1)$$

Overall permutations $\pi \in P_n$

A. Mathematical Model Assumptions

- *N* is the dimension of the problem case
- The objective function is Mini Sum
- π signifies a potential permutation over (1, 2,...,n) and π(*i*) relates to the index of the location to which facility i is allocated
- π is an ideal way of representing a solution to a QAP problem
- Each facility is allocated to precisely one location and vice-versa
- The solution space is discrete and finite
- The number of location and facilities are known
- All decision variables of the model are binary (0–1) variables

B. Mathematical Model Outputs (Decision Variables)

$$X_{ij} = \begin{cases} 1 & if\ facility\ i\ assigned\ in\ location\ j \\ 0 & Otherwise \end{cases} \qquad (2)$$

III. Materials and Methods

A. Materials

Since the first formulation of the QAP model, numerous researchers have performed studies in this are to generate algorithms with a capacity of locating practical solutions. Several algorithms were created along with numerous problem instances. Several researchers from the Graz University of Technology created the QAPLIB (http://anjos.mgi. Polymtl. ca/qaplib/) in order to deliver these data and explanations to the scientific community. Then and there, the QAPLIB was an up-to-date source which possessed all the available QAP instances. Majority of the available algorithms for the purpose of solving the QAP were examined on these benchmark instances. In excess of over 100 instances were obtained either from real life applications or randomly produced problem instances. In this study, five categories of instances were solved by the QAPLIB with problem sizes fluctuating from 12 to 80 locations as follows

- Randomly generated instances such as (Tai25a, Tai30a, Tai40a, Tai50a, Tai60b, Tai64c, Lipa70a, Lipa80a).
- Real-life instances such as (Chr12c, Chr15a, Bur26a, Kra30a, Kra30b, Ste36a).
- With grid-based distance matrix such as (Nug12, Nug14, Nug15, Nug16a, Nug16b, Nug17, Nug18, Nug20, Nug25, Sko49).
- The entries in flow matrices of the rectangular distanced problems are pseudorandom numbers (Sko49, Wil50).
- Manhattan distances of rectangular grids such as (Had12, Had14, Had20).

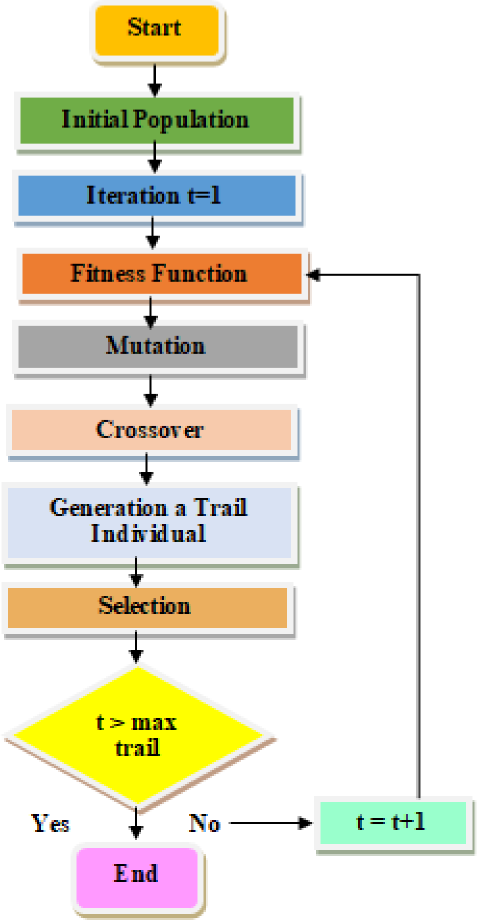


Fig. 1. Flowchart of DDE Algorithm.

B. Methods

Various metaheuristics have been used and developed for finding the best solution of QAP and the likelihood of receiving a satisfying solution value within an acceptable time span. In this study, the modified of discrete differential evolution (DDE) has been proposed, which includes the type of the crossover (ULX) [18] to get the diversity of the search space. Moreover, the algorithm can solve the permutation of the QAP. The DDE algorithm is simple in nature and by mutating the target population it produces the mutant population. Then, a crossover operator is required to incorporate the mutated solution with the target solution in order to produce a trail solution. Finally, the selection was based on the survival of the fittest among the trial and target solutions. Fig. 1. shows the main steps of the proposed approach to solve QAP.

1) Initial population: The DDE algorithm begins with initializing of the primary target population $\pi_i = [\pi_1, \pi_2, .., \pi_{NP}]$ with the dimension of NP. Every individual contains an n-dimensional vector with parameter values randomly and equally established among pre-defined search range. The initial population of DDE algorithm is shown in Table 1.

TABLE I. INITIAL POPULATION

π_1	6	3	1	4	2	5
π_2	2	1	6	5	4	3
. . .	. . .	. . .	. . .	. . .	. . .	. . .
π_n	4	6	2	3	5	1

a) A permutation was directly encoded as an individual vector;

b) The dimension of the individual was equivalent to the size of the QAP problem;

c) In the initialization stage, the population was generated as a random permutation.

2) Mutation individual: The differential deviation was attained in the form of perturbations of the optimum solution from the former generation in the objective population. Since perturbations are stochastically achieved, every individual in the mutant population was anticipated to be characteristic. In order to attain the mutant individual, the subsequent equation could be employed

$$v_i^t = \begin{cases} insert(\pi_b^{t-1}) & if(r < P_m) \\ swap(\pi_b^{t-1}) & otherwise \end{cases} \tag{3}$$

where π_b^{t-1} is the optimum solution from the former generation in the objective population; P_m is the perturbation probability; the insert ( ) and swap ( ) are merely the solo additions and swap shifts. A constant arbitrary number r was produced among [0,1]. If r was lower than P_m, then a sole insertion shift was employed in order to produce the mutant individual v_i^t; else, a sole swap shift was employed.

3) Crossover

The crossover operator with the crossover was proposed by [18] and is known as the uniform like crossover (ULX). The crossover was obtained as the follows First, all items were allocated to a similar location and both parents were copied to this position in the child.

a) Second, the unassigned positions of a permutation were scanned from left to right: for the unassigned position, an item was selected randomly, consistently from those in the parents if they were not yet incorporated in the child.

b) Third, remaining items were randomly allocated.

4) Generation a trail individual: After the perturbation phase, the trial individual was attained as follows:

$$u_i^t = \begin{cases} CR & if(r < P_c) \\ v_i^t & otherwise \end{cases} \tag{4}$$

where CR is the crossover operator; and P_c is the crossover probability. When a uniform random number r was less than the Pc the crossover operator was utilized to produce the trial individual u_i^t . Otherwise, the trial individual was selected as $u_i^t = v_i^t$. Hence, the trial individual was made up of the outcome of the perturbation operators or from the crossover operator.

5) Selection: The selection was based on fitness function and the following equation can be used

$$\pi_i^t = \begin{cases} u_i^t & if\left(f(u_i^t) \leq f(\pi_i^{t-1})\right) \\ \pi_i^{t-1} & otherwise \end{cases} \tag{5}$$

The selection was grounded on the existence of the rightest amongst the trial and target individuals.

6) Verification of the stopping criterion: The stopping criterion is dependent on the finish of the specified number of repetitions. The algorithm could be stopped if the solution was not improved.

IV. COMPUTATIONAL RESULTS

The algorithm which was proposed was encoded in MATLAB on a PC with Intel(R) Core (TM) i7-3770 CPU @ 3.40 GHz and 4.00 GB RAM under MS Windows 10. This section has been presented two stages, the first stage included the parameters which used by the proposed algorithm. Then, the second stage has been presented the discussion of results which obtained by using the proposed algorithm.

TABLE II. PARAMETER SETTING OF DDE ALGORITHM FOR QAP

Parameters	**Value**
Population Size	100
Number of Particles	30
Maximal iterative Number	1000
Probability of Mutation Pm	[0,1]
Probability of Crossover Pc	[0,1]

A. Parameter setting

The parameters to be determined in the DDE has been shown in table 2 as follows:

B. Results and Discussions

The operations of the DDE are tabulated in Table 3,4, and 5. The RPD field denotes the Relative Percent Deviation between the best-found solution S by proposed algorithm and the optimal (or the Best-Known Solution BKS) as a formula:

$$RPD = \frac{(S - S_{optimal(or\,BKS)})}{S_{optimal(or\,BKS)}} * 100\% \qquad (6)$$

The precision of the algorithm was estimated by utilizing the rate of RPD. Smaller values of the average RPD was more robust for the evaluated algorithm. In the cases, (Nug12, Nug14, Nug15, Nug16a, Nug16b, Nug17, Nug18, Nug20, Nug25) the proposed DDE selects the optimal solution with gap 0%. Then, the performance of DDE finds the optimal solution in an instance (Bur26a, Bur26b, Bur26c, Bur26d, Bur26e, Bur26f, Bur26g, Bur26h) with gap 0%. On the other hand, the performance of DDE has been applied in 12 instances in this work (Tai25a, Tai30a, Tai40a, Tai50a, Tai60b, Sko49, Wil50, Lipa70a, Lipa80a, Chr15a, Esc128, Kra30b).

The attained results of these instances were suitable for finding the optimum solution and excellent accuracy. Finally, the different instances of QAP (Tai64c, Lipa40b, Chr12c, Esc16i, Had12, Had14, Had20) were solved by DDE. The results of the DDE were compared with the Tabu Search Algorithm (TS) which belongs to a local search. Other algorithms were compared with DDE such as Genetic Algorithm (GA), and Differential Evolution Algorithm (DE), which belong to evolutionary methods. The Tabu Search which belongs to local search algorithms, was applied in [19] to solve some of instances of QAP from benchmark (QAPLIB) such as (Nug12, Nug14, Nug15, Nug16a, Nug16b, Nug17, Nug18, Nug20, Nug25, Bur26a, Bur26b, Bur26c, Bur26d, Bur26e, Bur26f, Bur26g, Bur26h, Tai25a, Tai30a, Tai40a, Tai50a), while optimum results were obtained in the cases (Nug14, Nug17) with gap 0%. Next, in the size of problem less than 30 (Nug12, Nug15, Nug16a, Nug16b, Nug18, Nug20, Nug25, Bur26a, Bur26b, Bur26c, Bur26d, Bur26e, Bur26f, Bur26g, Bur26h, Tai25a) the TS was unable to access the optimal solution, and the gap of the results between these cases and the results in QAPLIB was between 0% to 5%.

On the other hand, the performance of TS was inferior to the cases (Tai30a, Tai40a, Tai50a) and the gap between these cases were between 4% to 8%. Moreover, the average of relative percent deviation between the solution by TS and the optimal solution or best-known solution in benchmark QAPLIB was 1.41%. Based on the dataset above, the performance of the proposed algorithm was better than the algorithm TS and it located the best solution for 17 out of 21 cases (Nug12, Nug14, Nug15, Nug16a, Nug16b, Nug17, Nug18, Nug20, Nug25, Bur26a, Bur26b, Bur26c, Bur26d, Bur26e, Bur26f, Bur26g, Bur26h) at the gap 0%. The obtained results were better than the algorithm TS in the cases (Tai25a, Tai30a, Tai40a, Tai50a). Finally, the average of relative percent deviation between the solution by DDE and the optimal solution or best-known solution in benchmark QAPLIB was 0.69%. Table 3 illustrates the comparison between the TS and DDE as follows

Fig. 2 displays the relative percentage deviation (relative difference) of the solution quality for various problem sizes for TS and DDE algorithms. The obtained results indicated that the DDE algorithm possessed a suitable solution quality, which was higher than the TS algorithm for solving QAP instances.

TABLE III. SUMMARY OF COMPARISON OF DDE VERSUS TS

Instances	QAPLIB		RPD of TS [19]	RPD of DDE
	Optimal	BKS		
Nug12	578		0.39	0
Nug14	1014		0	0
Nug15	1150		0.39	0
Nug16a	1610		0.87	0
Nug16b	1240		1.37	0
Nug17	1732		0	0
Nug18	1930		0.69	0
Nug20	2570		1.04	0
Nug25	3744		1.55	0
Bur26a	5426670		0.09	0
Bur26b	3817852		0.19	0
Bur26c	5426795		0.26	0
Bur26d	3821225		0.02	0
Bur26e	5386879		0.03	0
Bur26f	3782044		0.05	0
Bur26g	10117172		0.01	0
Bur26h	7098658		0.01	0
Tai25a	1167256		4.26	2.05
Tai30a		1818146	4.75	3.55
Tai40a		3139370	6.12	3.88
Tai50a		4938796	6.94	5.04
Average RPDs			**1.41**	**0.69**

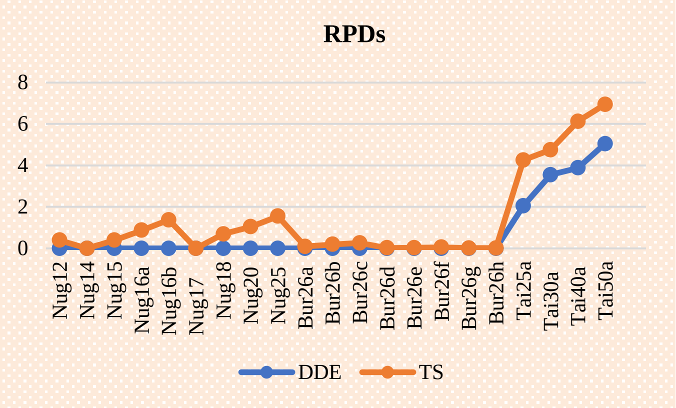


Fig. 2. Comparison of RPDs for DDE Versus TS.

TABLE IV. SUMMARY OF COMPARISON OF DDE VERSUS DE

Instances	QAPLIB		RPD of DE [17]	RPD of DDE
	Optimal	BKS		
Esc32e	2		0	0
Esc32g	6		0	0
Esc32h	438		15.07	0
Esc64a	116		24.14	0
Tai64c		1855928	0.19	0
Lipa40b	476581		24.73	0
Sko49		23558	12.5	3.45
Wil50		49482	7.58	1.36
Tai60b	630211362	630211362	26.45	3.61
Lipa70a	171605		1.76	1.08
Lipa80a	253195		1.59	1.04
Average RPDs			**10.36**	**0.95**

In a study by [17], DE was applied to solve 11 sets of benchmark instances as follows: Esc32e, Esc32g, Esc32h, Esc64a, Tai64c, Lipa40b, Sko49, Wil50, Tai60b, Lipa70a, and Lipa80a. The obtained results in the cases of Esc32e and Esc32g were an optimal solution with gap of 0%. Then, in the cases Esc32h, Esc64a, Lipa40b, Tai60b, Sko49, and Wil50, the results were inferior to the results in benchmark with the gap between 7.58 to 26.45%. Finally, the performance of DE was satisfactory in the cases of Tai64c, Lipa70a, and Lipa80a, while the gap of these results was between 0.19 to 1.76%.

The obtained results showed that the average of relative percent deviation between the solution by DE and the optimal solution or best-known solution in the benchmark was 10.36 %. By means of applying the proposed algorithm to the same cases that were solved by the algorithm DE, the execution of the proposed algorithm was superior to the DE algorithm, where the optimal solution was found for 6 cases out of 11 cases (Esc32e, Esc32g, Esc32h, Esc64a, Tai64c, Lipa40b) and the gap was 0%. On the other hand, the results for the five remaining cases (Sko49, Wil50, Tai60b, Lipa70a, Lipa80a) were noble and were superior to those obtained by DE, with the gap between 1.04 to 3.45%. The obtained results showed that the average of the relative percent deviation between the solution by DDE and the best solution in the benchmark was 0.95 %. Table 5 presents the comparison between the DE and DDE.

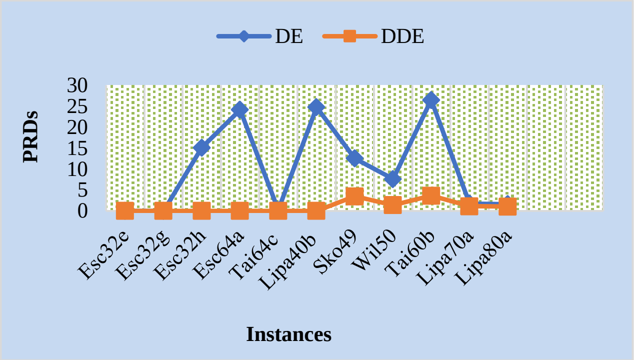


Fig. 3. Comparison of RPDs for DE Versus DDE.

TABLE V. SUMMARY OF COMPARISON OF DDE VERSUS GA

Instances	QAPLIB		RPD of GA [20]	RPD of DDE
	Optimal	BKS		
Bur26h	7098658		0.39	0
Chr12c	11156		0	0
Chr15a	9896		0.39	0
Esc128	64		0.87	0
Esc16i	14		1.37	0
Esc32h	438		0	0
Esc64a	116		0.69	0
Had12	1652		1.04	0
Had14	2724		1.55	0
Had20	6922		0.09	0
Kra30b	91420		0.19	0
Average RPDs			**13.14**	**4.38**

Fig. 3 displays the relative percentage deviation (relative difference) of the solution quality for various problems sizes for DE and DDE algorithms. The obtained results indicated that the DDE algorithm possessed a suitable solution quality, which was higher than the DE algorithm for solving QAP instances.

In a study by [20], GA was applied to solve 11 sets of benchmark instances, as follows: (Bur26h, Chr12c, Chr15a, Esc128, Esc16i, Esc32h, Esc64a, Had12, Had14, Had20, Kra30b). The obtained results showed for the three cases (Chr12c, Esc32h, Kra30b) the gap was between 6.09 to 9%, while the average of relative percent deviation between the solution by GA and the optimal solution in benchmark QAPLIB was 13.14%. The results of the proposed algorithm in this study are shown in Table 6. It was observed that DDE (4.38% grand average RPD) outperformed the GA (13.14% grand average RPD) on all 11 instances. The proposed algorithm was effective in selecting the optimal solution for eight out of 11 cases (Bur26h, Chr12c, Esc128, Esc16i, Esc32h, Esc64a, Had12, Had14, Had20).

Fig. 4 displays the relative percentage deviation (relative difference) of the solution quality for various problems sizes for GA and DDE algorithms. The obtained results indicated that the DDE algorithm possessed a suitable solution quality, which was higher than the GA algorithm for solving QAP instances.

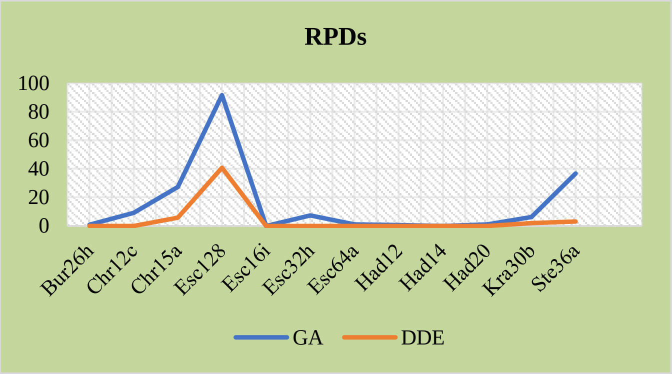


Fig. 4. Comparison of RPDs for DDE Versus GA.

The limitations of this study involve the absence of assurance in obtaining the optimum solutions in a predetermined amount of time. Nevertheless, asymptotic convergence proofs are accessible for problems, which are computationally expensive and require a massive amount of computational resources.

V. Conclusion

In this study, a modified discrete differential evolution (DDE) algorithm has been proposed for obtaining an operational solution to the QAP. Through beginning with an initial population of DDE, the single insertions and swap shift were applied to generate the mutation individually. However, uniform-like crossover (ULX) was employed as a crossover operator in this algorithm in order to obtain diversity in the search space and find the best solution of QAP. The obtained results of five classes of benchmark QAPLIB instances indicated the efficiency of the proposed algorithm. From the 41 instances which were investigated, 31 instances were solved optimally. Then a comparative study between DDE and three algorithms (TS, DE, and GA) were presented for similar instances.

The found results indicated that DDE outperformed TS, DE, and GA on all category instances. Moreover, DDE was found to be better than other algorithms in terms of the solution quality. Further studies regarding the utilization of DDE along with other algorithms could be carried out in order to provide better results.

Acknowledgment

This research has been supported by Universiti Teknikal Malaysia Melaka (UTeM) under UTeM Zamalah Scheme.

References

[1] T. C. Koopmans and M. J. Beckmann, 1957. “Assignment Problems and the Location of Economic Activities Author (s): Tjalling C. Koopmans and Martin Beckmann,” Econometrica, vol. 25, no. 1. pp.53–76.

[2] S. Sahni and T. Gonzalez, 1976. “P-Complete Approximation Problems,” J. ACM, vol. 23, no. 3, pp. 555–565.

[3] E. Duman, M. Uysal, and A. F. Alkaya, 2012. “Migrating Birds Optimization: A new metaheuristic approach and its performance on quadratic assignment problem,” Inf. Sci. (Ny)., vol. 217, pp. 65–77.

[4] U. Benlic and J. K. Hao, 2013. “Breakout local search for the quadratic assignment problem,” Appl. Math. Comput., vol. 219, no. 9, pp. 4800–4815.

[5] B. R. M. M. Y. Mohamad Amin Kaviani, Mehdi Abbasi, 2014. “A hybrid Tabu search-simulated annealing method to solve quadratic assignment problem,” Decis. Sci. Lett., vol. 3, no. 3, pp. 391–396.

[6] Shawky, L.A.E.F., Metwally, M.A.E.B. and Zaied, A.E.N.H., 2015. Quadratic Assignment Problem: A survey and Applications. International Journal of Digital Content Technology and its Applications, 9(2), p.90.

[7] R. K. Ahuja, K. C. Jha, J. B. Orlin and D. Sharma, 2002. “Very large-scale neighborhood search for the quadratic assignment problem”, Working Paper, MIT Sloan School of Management.

[8] Lim, W. L., Wibowo, A., Desa, M. I., and Haron, H. 2016. A biogeography-based optimization algorithm hybridized with tabu search for the quadratic assignment problem. Computational intelligence and neuroscience, 2016, 27.

[9] M. Abdel-Baset, H. Wu, Y. Zhou, and L. Abdel-Fatah, “Elite opposition-flower pollination algorithm for quadratic assignment problem,” J. Intell. Fuzzy Syst., vol. 33, no. 2. 2017, pp. 901–911.

[10] Shylo, P.V., 2017. Solving the quadratic assignment problem by the repeated iterated tabu search method. Cybernetics and Systems Analysis, 53(2), pp.308-311.

[11] Pradeepmon, T., Sridharan, R. and Panicker, V., 2018. Development of modified discrete particle swarm optimization algorithm for quadratic assignment problems. International Journal of Industrial Engineering Computations, 9(4), pp.491-508.

[12] Hameed, Asaad Shakir, Aboobaider, Burhanuddin Mohd, Choon, Ngo Hea, and Mutar, Modhi Lafta, 2018. Review on the Methods to Solve Combinatorial Optimization Problems Particularly: Quadratic Assignment Model. International Journal of Engineering & Technology, 7(3.20), pp.15-20.

[13] R. Storn, K. Price, 1997. “Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces”, Journal of Global Optimization, vol. 11, pp. 341- 359.

[14] Price, K.V., Storn, R. and Lampinen, J. 2005. “Differential Evolution: A Practical Approach to Global Optimization”. Springer-Verlag, London, UK.

[15] Pan, Quan-Ke, Mehmet Fatih Tasgetiren, and Yun-Chia Liang., 2008. "A discrete differential evolution algorithm for the permutation flow shop scheduling problem." Computers & Industrial Engineering 55.4, pp. 795-816.

[16] Kushida, Jun-ichi, et al., 2012. "Solving quadratic assignment problems by differential evolution." Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), Joint 6th International Conference on IEEE.

[17] M. Fatih Tasgetiren, Quan-Ke Pan, P. N. Suganthan, and Ikbal Ece Dizbay, 2013. “Metaheuristic Algorithms for the Quadratic Assignment Problem”. IEEE Symposium on Computational Intelligence in Production and Logistics Systems (CIPLS). pp. 131-137.

[18] D.M. Tate, A.E. Smith, 1995. A genetic approach to the quadratic assignment problem. Computers & Operations Research, Vol.1, pp. 73–83.

[19] Mohamad Amin Kaviania, Mehdi Abbasib, Bentolhoda Rahpeymab and Mohamad Mehdi Yusefib, 2014. A hybrid Tabu search-simulated annealing method to solve quadratic assignment problem. Decision Science Letters 3, pp. 391–396.

[20] Gamal Abd El-Nasser A. Said, Abeer M. Mahmoud, and El-Sayed M. El-Horbaty,2014. “A Comparative Study of Meta-Heuristic Algorithms for Solving Quadratic Assignment Problem”. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1

A Novel Image Encryption Approach for Cloud Computing Applications

Saleh ALTOWAIJRI[1]
Department of Information Systems
Faculty of Computing and Information Technology
Northern Border University
Kingdom of Saudi Arabia

Mohamed AYARI[2]
Department of Information Technology
Faculty of Computing and Information Technology
Northern Border University
Kingdom of Saudi Arabia

Yamen EL TOUATI[3]
Department of Computer Science
Faculty of Computing and Information Technology
Northern Border University
Kingdom of Saudi Arabia

***Abstract*—In this paper, a novel image encryption approach is proposed in the context of cloud computing applications. A fast special transform based on non-equispaced grid technique is introduced and applied as the first time in image encryption applications. By Combining with Fractional Fourier Transform (FRFT) instead of Discrete Fourier Transform (DFT), a good framework for image encryption is opened to enhance data security degree. The both image encipherment and decipherment process are analyzed based on random phase matrix. The time complexity effort of this novel approach is examined and evaluated. Comparative study with traditional encryption algorithms will prove the efficiency and robustness of our proposed technique.**

Keywords—Cloud computing; image encryption; fourier transform; random phase function

I. Introduction

The tremendous advancement of information and communications technology (ICT) enables big documents to be effortlessly diffused online. Data encipherment is commonly applied to guarantee different security branches mainly data confidentiality, data integrity, and data availability. By the way, the majority algorithms used for enciphering are implemented for textual data. Despite these algorithms are efficient with this kind of data, they are too limited with multimedia applications by the reason of imposed constrains such as real time analysis and big data treatment. To confront these limitations, many algorithms have developed in the context of image encipherment and can be classified into two sets. The first one consists in transforming two-dimensional image matrix into unidimensional vector and then apply one of the well-known modern encryption techniques[1-3] like data encryption standard (DES), three-DES and advanced encryption standard (AES) belonging to symmetric key cryptosystems or RSA(Rivest-Shamir-Adelman) cipher, Diffie-Hellman key exchange and Elliptic curve cryptography (ECC) belonging to public-key cryptosystems (PKCS). The second one is to handle 2D- matrix presenting the authentic image and apply one of the available image encryption techniques [4-5] such as wave transmission [6] and two-dimensional discrete Fourier transform (2D-DFT).

In fact, wave transmission technique as its name indicates consists in replacing the value of each image pixel by self-adaptive wave. Taking into consideration the phase matrix keys, this method as well as chaotic system can be combined with DFT [7], fractional Fourier transform (FrFT) or any other image processing methods in order to improve the image encryption algorithms in terms of parallel implementation, speediness and multi-parameter selection.

FrFT is considered as a powerful tool in many sectors like signal processing, quantum mechanics and optics [8-10]. It is nothing but a generalization of the classical FT (Fourier Transform) characterized by the transform order which makes it suppler than FT [8-11], increases the key space and then improves the security degree of systems based on FrFT [12-22].

The discrete version of FrFT so-called discrete fractional Fourier transform (DFrFT) can be obtained from DFT taking into consideration some free parameters [23-24]. In effect, the mathematical computation of DFrFT is achieved by the Eigen decomposition of the 2D-DFT matrix whereas the researchers have studied this decomposition in order to accelerate DfrFT algorithm by proposing different techniques to generate its eigenvectors such as commuting DFT matrix[25], Hermit-Gaussian[23], direct batch evaluation[27] and singular value decomposition method[26].

In addition few years ago, random phase encoding has been applied in the image encryption process to elevate its security level [28]. It is introduced in fractional Fourier domain and combined with fractional order to offer an efficient key for enciphering/deciphering schemes. But, the noise-like aspect presenting in encrypted image becomes a deficiency since it opens a greedy gate for unauthorized user to make more and more attacks.

As mentioned above, the proposed solutions to improve data security in literature are concentrated on either the introduction of intelligent technique to reinforce the different transforms used in encryption scheme, the addition of specific key as random phase or the combination of them. Another research axis as presented in [8] can be much attractive for Scientists by rectifying the image encipherment process where the two Fourier transform operators have been replaced by two FRFT one.

In this context, we propose a novel image encryption by introducing a new robust transform which is proposed for the first time in image encryption/decryption algorithms in order to ameliorate the security degree of data. This approach can be as well convenient to investigate cloud-based applications.

This paper is organized as follows: section II figures out the theoretical background and mathematical foundations building our novel image encryption process. Section III describes the proposed encryption/decryption algorithms. The analysis and evaluation of the proposed approach in the context of cloud-based applications are the subject of subsequent section (Section IV). Section V presents our conclusions and ongoing work.

II. Theoretical Background

Many mathematical transformations are in the core of different disciplines related to image encryption and offer a good workspace to increase the data security. In this section, we will focus only on the aforementioned main important transformations that will be implemented and combined in the context of image encryption.

A. *Two Dimensional-Discrete Fourier Transform (2D-DFT)*

2D-DFT is considered as a very popular signal processing tool to understand the features of periodic, discrete-time signals in the frequency domain. Hence, it is very likely that it will be implemented in our proposed application.

The 2D-DFT of a sequence (x_{kl}) can be defined as

$$X_{mn} = \frac{1}{\sqrt{M*N}} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} x_{kl} * W_N^{nk} * W_M^{ml} \quad (1)$$

So, its inverse can be written as:

$$x_{kl} = \frac{1}{\sqrt{M*N}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X_{mn} * W_N^{-nk} * W_M^{-ml} \quad (2)$$

Where

$$W_N = e^{-j2\pi/N} \quad (3)$$

$$W_M = e^{-j2\pi/M} \quad (4)$$

and (x_{kl}) are a finite duration matrix with length *N*M*.

B. *Fast Fractional Fourier Transform (FrFT)s*

Conventional Fourier transform takes the following expression:

$$F(v) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t)\, e^{-jvt} dt \quad (5)$$

$$F(t) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} F(v)\, e^{-jvt} dv \quad (6)$$

Referred to (5) and (6), the α^{th} order fractional Fourier transform $F^{\propto}(v)$ can be written as follow:

$$F^{\propto}(v) = \begin{cases} f(v) & ,\alpha = 2n\pi\ (n\ integer) \\ f(-v) & ,\alpha = (2n+1)\pi\ (n\ integer) \\ \frac{1}{2\pi}\left((1-jcot\alpha)e^{jv^2cot\alpha}\right)^{\frac{1}{2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2jt^2cot\alpha}} e^{-j\llbracket(sin\rrbracket^{-1}\alpha)vt} f(t)dt & ;\ \alpha \neq n\pi \end{cases} \quad (7)$$

Obviously, FrFT has many properties like the conventional Fourier transform since FrFT is nothing but a general case of FT. The reactions of $F^{\propto}(v)$ applied on a Pi function $\prod(t/4)$ with low order α are given in Fig. 1. By the same way, the responses for high order are shown in Fig. 2.

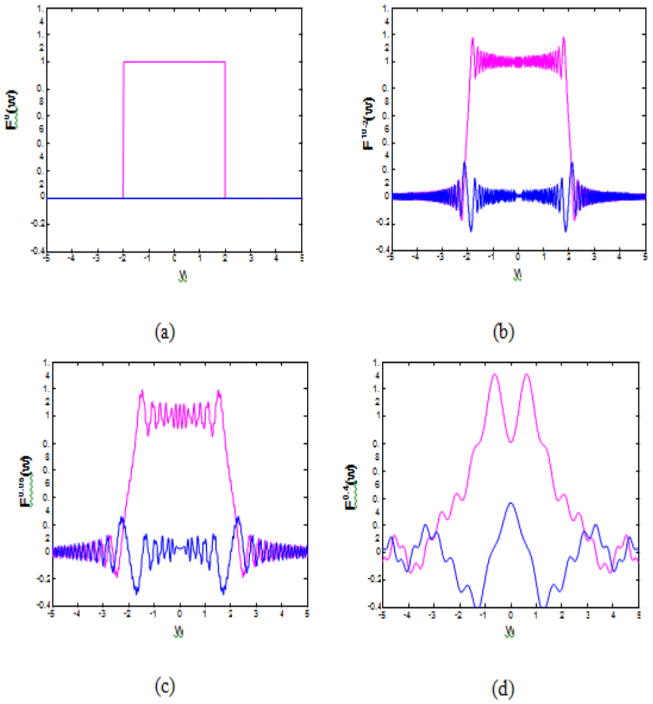


Fig. 1. FrFT with the order (a) α=0 (b) α=1/100 (c) α=1/20 (d) α=2/5 (Blue for imaginary part and pink for real part).

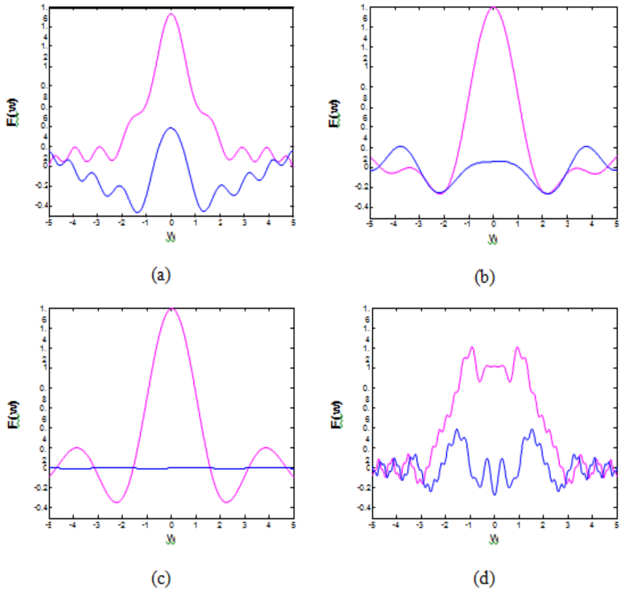


Fig. 2. FrFT with the order (a) α=π/4 (b) α=7π/16 (c) α=π/2 (d) α=15π/16 (Blue for imaginary part and pink for real part).

By optimizing the fundamental range of the fractional order, an efficient and rapid algorithm for numerical computation of FrFT has been investigated in [29] and refereed to [30], a two-dimensional FrFT based can be easily implemented in our work.

C. Two-Dimensional Fast Wavelet Transform (2D-FWT)

2D-FWT is a numerically effective way of the 2D-discrete wavelet transform (2D-DWT) taking into account at adjacent scales the relationship between all 2D-DWT coefficients.

Let consider $\phi(x,y)$ a two-dimensional scaling function, and $\psi^V(x,y)$, $\psi^H(x,y)$ and $\psi^D(x,y)$, three 2-D wavelets defined respectively as follows:

ψ^V: Defines changes according to rows (i.e. vertical edges),

ψ^H: Defines changes according to columns (i.e. horizontal edges),

ψ^D: Defines changes according to diagonals,

Each of the aforementioned functions can be written as a multiplication of a unidimensional scaling function ϕ and its analogous wavelet ψ.

$$\phi(x,y) = \phi(x)\phi(y) \tag{8}$$

$$\psi^V(x,y) = \phi(x)\psi(y) \tag{9}$$

$$\psi^H(x,y) = \psi(x)\phi(x) \tag{10}$$

$$\psi^D(x,y) = \psi(x)\psi(y) \tag{11}$$

By applying numerical filters and its associated down-samplers, 2D-DWT can be easily implemented.

The diagram depicted in Fig.3 describes the two-dimensional FT analysis filter bank. This one-scale filter bank can be repeated by applying the computation output to another filter bank input so as to generate as fractal scale system.

The following diagram as shown in Fig.4 presents two recurrences of the filtering method where the sub-images— W_ϕ, W_ψ^H ,W_ψ^V, and W_ψ^D are produced in the first iteration and the second one shows the two-scale decomposition.

To obtain the reverse problem already described in Fig. 3, the synthesis filter bank should be applied as depicted in Fig. 5:

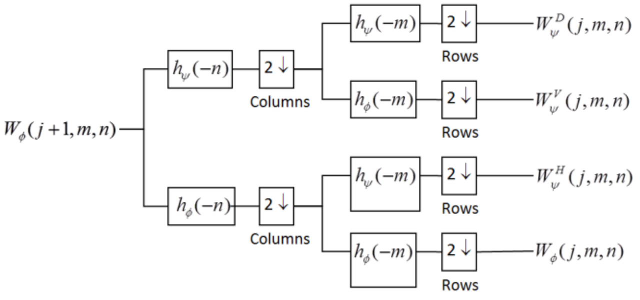


Fig. 3. 2D-FWT – Analysis Filter Bank.

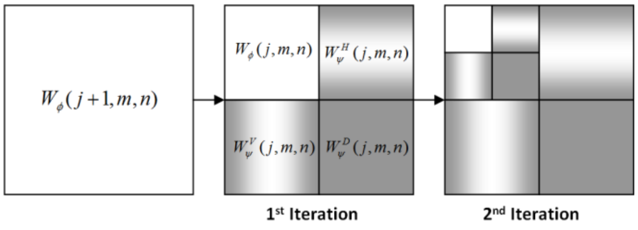


Fig. 4. Two-Scale of 2D-Decomposition.

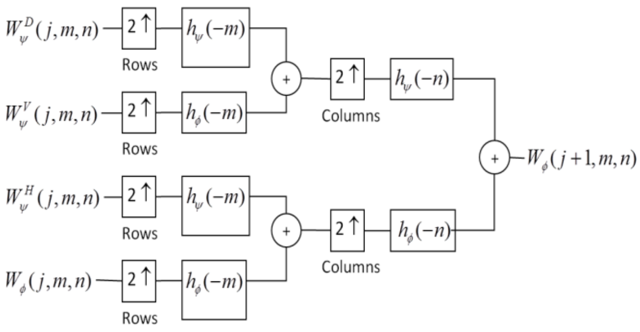


Fig. 5. 2D-IFWT – Synthesis Filter Bank.

D. 2D-Non-Uniform Fast Fourier Transform (2D-NUFFT)

In image encryption applications, the 2D-NUFFT [31] – developed by our research team– can be a good solution to enhance the security of data. But up to now, this technique has not introduced in the encryption domain. The non-uniformity and the speed of this mathematical transform can produce a powerful mechanism for many research applications. Under the frame of the numerical complexity, the conversion between spatial and Fourier domains is ensured at $O(N_T \log N_T)$time (N_T is the total pixel numbers in 2D decomposition) while taking into consideration the non-equispaced of input data.

The efficiency of the 2D-NUFFT algorithm is equivalent to the FFT one. Therefore, this powerful transform will be implemented in this work by combining with aforementioned mathematical transformations. The schemes depicted in the following Figures (Fig.6 and Fig.7) [31] summarize this transformation and its inverse.

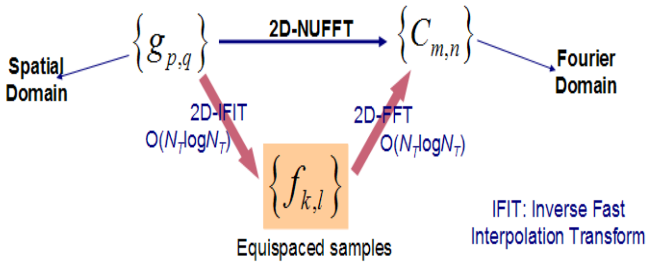


Fig. 6. 2D- NUFFT Scheme.

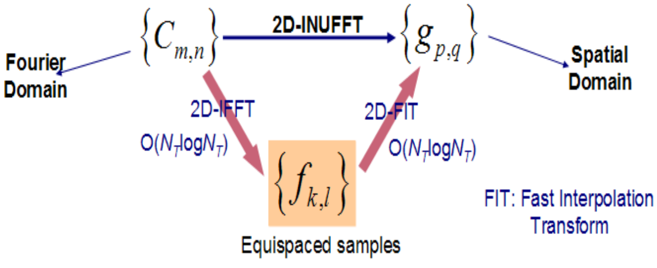


Fig. 7. 2D-Inverse NUFFT Scheme.

III. Proposed Encryption/Decryption Algorithms

In this section, we develop the new encryption and decryption schemes based on the different tools mentioned above in the context of image encryption applications.

Let consider a 2D color image *I* = *F*(*x*,*y*) (x and y are the spatial coordinates of the image) with size *M*N* identifying *RGB* (Red, Green, Blue) channels.

This image can be represented as follows:

$$I = \sum_{x=1}^{M} \sum_{y=1}^{N} F(x,y) \qquad (12)$$

Where

$$F(x,y) = \sum_{1 \le i \le 3} f_i(x,y) \qquad (13)$$

Assume that these channels of the authentic image are separated. So, all different processes already described above will be applied simultaneously to RGB channels.

Initially, the *2D-NUFFT* transform is executed in each channel to produce:

$$2D\text{-}NUFFT\{f_i(x,y)\} \qquad 1 \le i \le 3 \qquad (14)$$

Then, the 2D-FWT is applied straightforwardly for each channel to generate:

$$2D\text{-}FWT\left\{2D\text{-}NUFFT\{f_i(x,y)\}\right\} \qquad 1 \le i \le 3 \qquad (15)$$

The obtained distribution from (15) will be coded by the Chaotic Random Phase Mask (CRPM). This can be expressed arithmetically as follows:

$$CRPM = e^{j\frac{\pi}{2}S(x,y)} \qquad (16)$$

S(x,y) stands for the sequence with random number produced by the chaos function with selected seed value.

The 2-*D FrFT* transform is then implemented in this to produce:

$$2D\text{-}FrFT_{\alpha,\beta}\left\{2D\text{-}FWT\left\{2D\text{-}NUFFT\{f_i(x,y)\}\right\} * e^{j\frac{\pi}{2}S(x,y)}\right\} \qquad 1 \le i \le 3 \qquad (17)$$

where α,β are 2D- *FrFT* fractional orders and * denotes the element-wise multiplication (i.e. Hadamard product) of equivalent matrices.

The inverse *2D- FWT* must be now applied at the resultant matrix obtained from (17).

In the final step, each channel will be operated with the inverse 2D-NUFFT to generate RGB channels associated to encoded image. This encrypted image *G*(*x*, *y*) is generated after the combination of these three channels via the following equation:

$$G(x,y) = \sum_{i=1}^{3} g_i(x,y) =$$

$$\sum_{i=1}^{3}\left\{2D\text{-}INUFFT\left\{2D\text{-}IFWT\left\{2D\text{-}FrFT_{\alpha,\beta}\left\{2D\text{-}FWT\{2D\text{-}NUFF\right.\right.\right.\right.$$

$$\left.\left.\left.\left.e^{j\frac{\pi}{2}S(x,y)}\right\}\right\}\right\}\right\} \qquad (18)$$

These steps can be seen in the following process (fig.8):

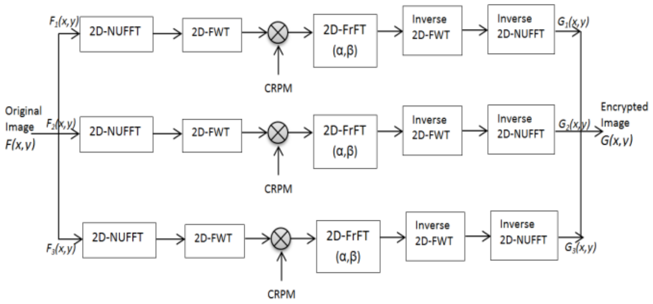


Fig. 8. Encryption Scheme in Proposed Algorithm.

The decipherment process given in the following figure (Fig.9) can be obtained from the reverse of the encipherment scheme as shown in Fig.8.

Indeed, the 2D-NUFFT operation should be applied to G(x,y) for each channel in order to generate:

$$2D\text{-}NUFFT\{g_i(x,y)\} \qquad 1 \le i \le 3 \qquad (19)$$

Then, the 2D-FWT is directly applied on (19). It yields:

$$2D\text{-}FWT\left\{2D\text{-}NUFFT\{g_i(x,y)\}\right\} \qquad 1 \le i \le 3 \qquad (20)$$

The inverse 2D- FrFT (with order -α, -β) will be applied after that taking into consideration the CRPM conjugate already presented in equation (16).

We obtain:

$$2D\text{-}IFrFT_{-\alpha,-\beta}\left\{2D\text{-}FWT\left\{2D\text{-}NUFFT\{g_i(x,y)\}\right\} * Conj\left\{e^{j\frac{\pi}{2}S(x,y)}\right\}\right\} \qquad 1 \le i \le 3 \qquad (21)$$

Finally by applying the combination between inverse 2D-FWT and inverse 2D-NUFFT on the outcome obtained in (21), the decrypted image can be now represented by the following distribution:

$$F'^{(x,y)} = \sum_{i=1}^{3} f'_i(x,y) =$$

$$\sum_{i=1}^{3}\left\{2D\text{-}INUFFT\left\{2D\text{-}IFWT\left\{2D\text{-}IFrFT_{-\alpha,-\beta}\left\{2D\text{-}FWT\{2D\text{-}NUFFT\{f_i(x,y)\}\} * Conj\left\{e^{j\frac{\pi}{2}S(x,y)}\right\}\right\}\right\}\right\}\right\} \qquad (22)$$

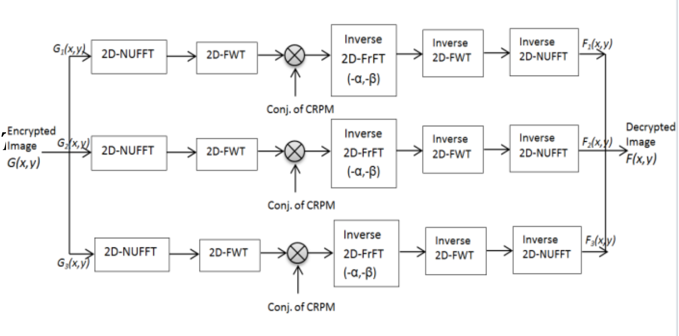


Fig. 9. Decryption Scheme in Proposed Algorithm.

IV. Analysis of the Proposed Algorithm

The proposed algorithm is based on the combination between two powerful mathematical transforms – 2D-NUFFT and 2D-FWT – and also their inverses 2D-INUFFT and 2D-IFWT in the same encryption or decryption process. The random phase function is produced as a 2-D sequence of random numbers and merged with the 2D-FrFT with order (α,β) in encryption process. By the way, the conjugate of the random phase function is used with the 2D-IFrFT with order (-α,-β) in the decryption process.

In the following table, we compute the complexity of this algorithm in both encryption and decryption sides so as to evaluate the efficiency and robustness of our new image encryption scheme. We adopt $N*N$ as the image size of I in order to simplify the problem.

TABLE I. Time Complexity of the Proposed Algorithm

	Step	Description	Complexity
Encryption Phase	1	Application of *2D-NUFFT* on the RGB channels *fi(x, y)* (1<=i<=3) of given authentic image.	$\boldsymbol{O(N^2\, log_2 N)}$
	2	Application of *2D-FWT*	$\boldsymbol{O(N)}$
	3	Encoding by random phase function	$\boldsymbol{O(\frac{N^2}{2})}$
	4	Application of *2D-FrFT* with order (α,β)	$\boldsymbol{O(N^3 + Nlog_2 N}$
	5	Application of *2D-IFWT*	$\boldsymbol{O(N)}$
	6	Application of *2D-INUFFT* on the RGB channels.	$\boldsymbol{O(N^2\, log_2 N)}$
		Total time complexity of encryption	$\boldsymbol{O(N^3)}$
Decryption Phase	1	Application of *2D-NUFFT* on the RGB channels *fi(x, y)* (1<=i<=3) of given authentic image.	$\boldsymbol{O(N^2\, log_2 N)}$
	2	Application of *2D-IFWT*	$\boldsymbol{O(N)}$
	3	Encoding by conjugate random phase function	$\boldsymbol{O(\frac{N^2}{2})}$
	4	Application of *2D-IFrFT* with order (-α,-β)	$\boldsymbol{O(N^3 + Nlog_2 N)}$
	5	Application of *2D-IFWT*	$\boldsymbol{O(N)}$
	6	Application of *2D-INUFFT* on the RGB channels.	$\boldsymbol{O(N^2\, log_2 N)}$
		Total time complexity of decryption	$\boldsymbol{O(N^3)}$
Total computation complexity of the system			$\boldsymbol{O(N^3)}$

The evaluation of proposed encryption/decryption scheme by computing the number of its operations proves the speediness, efficiency and robustness of this algorithm in the context of image encryption applications. Furthermore in comparison with other methods developed in references [16-20], the security degree becomes significant due to the introduction of the non-uniformity feature in the proposed process by the application of the NUFFT technique.

There is no doubt that cloud computing technology [32] is a developing sector of both computer and networking security and information security in large-scale

There are a set of significant strategy issues used to secure data, applications, and the related infrastructure of cloud-based resources, which include image encryption applications. Such proposed encryption/decryption algorithm shall be as accurately tailored as feasible to protect a big data and ensure a good workspace to handle with image having a high complexity problem.

V. Conclusion

In this paper, the theoretical background and mathematical foundations of the novel image encryption approach have been successfully developed and presented. The features of this approach as well as its strong points that set apart from other encryption schemes have been mentioned. Also, the detailed description of the proposed encryption and decryption process has been presented. The efficiency and robustness of our approach in terms of computational complexity have been demonstrated in the context of image encryption and cloud computing applications. The introduction of double chaotic random phase mask associated with 2D- FrFT in our approach can increase the security degree and circumvent the problem of big data met in cloud computing technologies.

Acknowledgment

The authors gratefully acknowledge the approval and the financial support of this research from the Deanship of Scientific Research study by the grant number CIT-2016-1-6-F-5704, Northern Border University, Arar, KSA.

References

[1] National Institute of Standards and Technology, “Data encryption standard (DES),” (1999).

[2] National Institute of Standards and Technology, “Advanced encryption standard (AES),” (2001).

[3] Rivest, R. L., Shamir, A., and Adleman, L., “A method for obtaining digital signatures and public-key cryptosystems,” Communications of the ACM 21(2), 120–126 (1978).

[4] Zhou, Y., Panetta, K., Agaian, S., and Chen, C. L. P., “Image encryption using p-Fibonacci transform and decomposition,” Optics Communications 285(5), 594–608 (2012).

[5] Zhou, Y., Panetta, K., Agaian, S., and Chen, C. L., “(n, k, p)-gray code for image systems,” IEEE Trans Syst Man Cybern B Cybern (2012)

[6] Liao, X., Lai, S., and Zhou, Q., “A novel image encryption algorithm based on self-adaptive wave transmission,” Signal Processing 90, 2714–2722 (2010).

[7] Chang, H. T., Hwang, H.-E., and Lee, C.-L., “Position multiplexing multiple-image encryption using cascaded phase-only masks in fresnel transform domain,” Optics Communications 284(18), 4146–4151 (2011).

[8] LIMA, Juliano B.; NOVAES, L. F. G. Image encryption based on the fractional Fourier transform over finite fields. Signal Processing, 2014, 94: 521-530.

[9] AWAL, Md Abdul, et al. A robust high-resolution time–frequency representation based on the local optimization of the short-time fractional Fourier transform. Digital Signal Processing, 2017, 70: 125-144.

[10] Madrid, Y., Molina, M., & Torres, R., "Quantum Fractional Fourier Transform. In Frontiers in Optics". Optical Society of America. (pp. JTu2A-73, September 2018.

[11] D. Mustard, "The fractional Fourier transform and the Wigner distribution," J. Aust. Math. Soc. B, vol. 38, pp. 209–219, 1996.

[12] R. Tao, B. Deng, and Y. Wang, "Research progress of the fractional Fourier transform in signal processing," Science in China (Ser.F, Information Science), vol. 49, pp. 1–25, Jan. 2006.

[13] G. Unnikrishnan and K. Singh, "Double random fractional Fourier-domain encoding for optical security," Opt. Eng., vol. 39, pp. 2853–2859, 2000.

[14] G. Unnikrishnan, J. Joseph, and K. Singh, "Optical encryption by double random phase encoding in the fractional Fourier domain," Opt. Lett., vol. 25, no. 12, pp. 887–889, 2000.

[15] Zhu B, Liu S, Ran Q: Optical image encryption based on multifractional Fourier transforms. Opt. Lett. 25 (2000),pp 1159–1161

[16] B. M. Hennelly and J. T. Sheridan, "Image encryption based on the fractional Fourier transform," Proc. SPIE, vol. 5202, pp. 76–87, 2003.

[17] R. Tao, Y. Xin, and Y. Wang, "Double image encryption based on random phase encoding in the fractional Fourier domain," Opt. Express, vol. 15, no. 24, pp. 16067–16079, 2007.

[18] R. Tao, X. M. Li, and Y.Wang, "Generalization of the fractional Hilbert transform," IEEE Signal Process. Lett., vol. 15, pp. 365–368, 2008.

[19] Hennelly B, Sheridan JT: Optical image encryption by random shifting in fractional Fourier domains. Opt. Lett. 28 (2003),pp 269–271.

[20] S. C. Pei and W. L. Hsue, "Random discrete fractional Fourier transform," IEEE Signal Process. Lett., vol. 16, no. 12, pp. 1015–1018, Dec.2009.

[21] L. J. Yan and J. S. Pan, "Generalized discrete fractional Hadamard transformation and its application on the image encryption," in Proc. Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing, 2007, pp. 457–460.

[22] H. Al-Qaheri, A. Mustafi, and S. Banerjee, "Digital watermarking using ant colony optimization in fractional Fourier domain," J. Inf. Hiding Multimedia Signal Process., vol. 1, no. 3, pp. 179–189, Jul. 2010.

[23] S. C. Pei and M. H. Yeh, "Improved discrete fractional Fourier transform," Opt. Lett., vol. 22, pp. 1047–1049, 1997.

[24] C. Candan, M. A. Kutay, and H. M. Ozaktas, "The discrete fractional Fourier transform," IEEE Trans. Signal Process., vol. 48, no. 5, pp. 1329–1337, May 2000.

[25] B. W. Dickinson and K. Steiglitz, "Eigenvectors and functions of the discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-30, pp. 25–31, Jan. 1982.

[26] M. T. Hanna, N. P. A. Seif, and W. A. E. M. Ahmed, "Hermite-Gaussian-Like eigenvectors of the discrete Fourier transform matrix based on the singular value decomposition of its orthogonal projection matrices," IEEE Trans. Circuits Syst. I, vol. 51, no. 11, pp. 2245–2254, 2004.

[27] M. T. Hanna, N. P. A. Seif, and W. A. E. M. Ahmed, "Hermite–Gaussian-Like eigenvectors of the discrete Fourier transform matrix based on the direct utilization of the orthogonal projection matrices on its Eigen-spaces," IEEE Trans. Signal Process., vol. 54, no. 7, pp. 2815–2819, Jul. 2006.

[28] Liu, Z., Chen, H., Blondel, W., Shen, Z., & Liu, S. Image security based on iterative random phase encoding in expanded fractional Fourier transform domains. Optics and Lasers in Engineering, 105, 1-5, 2018.

[29] Yang, Xingpeng, Qiaofeng Tan, Xiaofeng Wei, Yong Xiang, Yingbai Yan, and Guofan Jin. "Improved fast fractional-Fourier-transform algorithm." JOSA A 21, no. 9 (2004): 1677-1681.

[30] Zayed, Ahmed. "Two-dimensional fractional Fourier transform and some of its properties." Integral Transforms and Special Functions (2018): 1-18.

[31] Ayari, Mohamed, Taoufik Aguili, and Henri Baudrand. "New version of TWA using two-dimensional non-uniform fast fourier mode transform (2d-nuffmt) for full-wave investigation of microwave integrated circuits." Progress In Electromagnetics Research 15 (2009): 375-400.

[32] Ali, Mazhar, Samee U. Khan, and Athanasios V. Vasilakos. "Security in cloud computing: Opportunities and challenges." Information sciences 305 (2015): 357-383.

Repetitive Control based on Integral Sliding Mode Control of Matched Uncertain Systems

Nizar TOUJENI[1], Chaouki MNASRI[2], Moncef GASMI[3]
Computer Laboratory for Industrial Systems (LISI)
National Institute of Applied Sciences and Technology (INSAT)
Carthage University, Tunisia

***Abstract*—This paper proposed an integral sliding mode control scheme based on repetitive control for uncertain repetitive processes with the presence of matched uncertainties, external disturbances and norm-bounded nonlinearities. A new method based on the combination of repetitive control and sliding mode approach is studied in order to use the robustness sensibility property of the sliding mode control to matched uncertainties and disturbances and to cancel gradually tracking error for periodic processes. A sufficient condition of the existence of sliding mode is studied based on basic repetitive control and a sliding mode controller is synthesized through linear matrix inequalities, which guarantees the stability along the periods of the controlled closed-loop process and the reachability of the sliding surface is ensured. Then, an adaptive integral sliding mode controller is synthesized to improve performances of the proposed control scheme. The effectiveness of the proposed controlled design schemes is proved by the use of a third order uncertain mechanical system and the simulation results using the new approaches give good performances.**

Keywords*—Repetitive control; 2D systems; matched uncertainties; integral sliding mode control; sliding surface; linear matrix inequality*; *reachability

I. Introduction

Repetitive Control (RC) has been applied to many engineering applications, such as robot manipulators, rotary systems, power supply systems, computer disk drives, etc. [1]-[4]. Based on the error signals in previous periods, the basic idea of RC is to improve transient responses on each pass by refining the control inputs in tracking problems for periodic operated dynamic systems. The basic RC is related to learning control [5]-[6] and it is formed of two main parts in order to produce a zero tracking error; a periodic signal is generated by the internal model originated from the idea offered by Wonham and Francis [7] and a proper compensator consists to stabilize the closed-loop feedback system.

Recently, an interesting theme in RC research fields is robust repetitive control design against system uncertainties. In practice, there exist a complex relationship between model and real system. When the controller is applied to real systems, disturbances and uncertainties must be absolutely considered and examined. Thus, they cause instability in the control system [8].

In other hand, many strategies using RC have been developed in order to solve this problem. Authors in [9]-[14] offer some methods of repetitive control system design for a class of linear system. They are based on two-dimensional (2D) continuous/discrete hybrid model. The traditional problem of repetitive controller design is reformed in an equivalent problem for a 2D continuous-discrete system and solved it by 2D Lyapunov theory by means linear matrix inequality (LMI) approach.

Therefore, classical RC has been associated and integrated with many robust control techniques [15]-[21] such as backstepping control, adaptive robust control, H_∞ control, and sliding mode control (SMC). Referring to offered [8], [22], SMC has been looked as a good robust technique, especially for its insensibility to uncertainties satisfying the matching condition. Besides, SMC can offer good transient performance, fast response, and order reduction. These advantages make SMC a very practical and effective way in robust control design. Therefore, robustness against uncertainties matched to the control system can be only reached after the apparition of the sliding mode so-called reaching phase [23]-[24]. To ensure robustness in the overall closed-loop system response and eliminate the reaching phase, Integral Sliding Mode Control (ISMC) was proposed in [25]. The main contribution of this work is to propose a new Repetitive Integral Sliding Mode Control (RISMC) law of matched uncertain linear repetitive processes with external disturbances in order to achieve a zero tracking error.

The rest of this paper is organized as follows. Section II presents major work related to this study. The problem of equivalence 2D system and repetitive control is formulated in Section III. The sliding mode process is analyzed in Section III. Section V presents the reachability analysis. Section VI gives an illustrative example, and Section VII concludes this paper.

II. Related Works

The authors in [26] proposes a sliding mode based repetitive control system for periodic reference tracking in order to reduce transient overshoot, output noise, and chattering. The design method is simple with less restriction in stability conditions. The research work on [27] focused on improving the tracking performance and robust performance for turntable system by using a repetitive control design based on integral sliding mode. In [28], a quasi-sliding mode control of differential linear repetitive processes with unknown input disturbance is proposed.

The main advantage of this work is to design a repetitive controller based on integral sliding mode control of matched uncertain systems with external disturbances by exploiting two major properties; the first is the insensibility to uncertainties satisfying the matching condition and the second is the improving control system performance in a periodic manner by including the learning capacity.

III. Problem Formulation and Preliminaries

A. Problem Statement

Consider this uncertain system defined by:

$$\begin{cases} \dot{x}(t) = (A+\Delta A)x(t) + (B+\Delta B)u(t) + f(x,t) + \varpi(t) \\ y(t) = C\,x(t) \end{cases} \tag{1}$$

Where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, $f(x,t)$ and $\varpi(t) \in \mathbb{R}^m$ are the state vector, the output of the system, the input control, the vector of unmodelled dynamics and nonlinearities, and the vector of external disturbances respectively. Thus, A is the state matrix, B is the input matrix and C is the output matrix with appropriate dimensions. $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta B \in \mathbb{R}^{n\times m}$ represent the system and the input matrix uncertainties. To complete the description of the uncertain dynamical system, the following assumptions are used:

Assumption 1:

(i) The pair (A,B) is stabilizable and the input matrix B has full rank.

(ii) ΔA , ΔB , $f(x,t)$ and $\varpi(t)$ are continuous on their arguments and they are unknown but have a known upper bound for all $(x,t) \in \mathbb{R}^n \times \mathbb{R}$.

(iii) Matching conditions: there exist functions ΔA_m , ΔB_m , $v(t)$ and $g(x,t)$, for all $(x,t) \in \mathbb{R}^n \times \mathbb{R}$, such that

$$\begin{aligned} \Delta A &= B\Delta A_m \\ \Delta B &= B\Delta B_m \\ f(x,t) &= Bg(x,t) \\ \varpi(t) &= Bv(t) \end{aligned} \tag{2}$$

Assumption 2: There exist known positive constants a_m , b_m , α_m and β_m such that

$$\begin{aligned} &\|\Delta A_m\| \le a_m \\ &\|\Delta B_m\| \le b_m < 1 \\ &\|f(x,t)\| \le \alpha_m \|x\| \\ &\|\varpi(t)\| \le \beta_m \end{aligned} \tag{3}$$

Where $\| \,.\, \|$ denotes a signal quadrically-norm.

The tracking error between the periodic reference input and the output is defined by $e(t) = r(t) - y(t)$ where $r(t+T) = r(t)$ and T is the fundamental repetition period. Consider the following function which is describe by

$$\begin{aligned} \xi(t) &= \xi(kT+\tau) := \xi_k(\tau) \\ \Delta\xi(t) &= \xi(t) - \xi(t-T) := \Delta\xi_k(\tau) \end{aligned} \tag{4}$$

Where $k \in \mathbb{N}$ and $0 \le \tau < T$ are two independent parameters. The first describes learning between successive periods and the second characterizes control inside a period.

By applying function which defined in (4), plant model can be transformed in the following state-space equation

$$\begin{cases} \Delta\dot{x}_{k+1}(\tau) = (A+\Delta A)\Delta x_{k+1}(\tau) + (B+\Delta B)\Delta u_{k+1}(\tau) \\ \qquad\qquad + \Delta f_{k+1}(x_{k+1},\tau) + \Delta\varpi_{k+1}(\tau) \\ e_{k+1}(\tau) \;\;\, = \text{-}C\,\Delta x_{k+1}(\tau) + e_k(\tau) \end{cases} \tag{5}$$

Let $\eta_{k+1}(t) = \Delta x_{k+1}(\tau)$, $\tilde{\varpi}_{k+1}(t) = \Delta\varpi_{k+1}(\tau)$, $\tilde{u}_{k+1}(t) = \Delta u_{k+1}(\tau)$, and $\tilde{f}_{k+1}(\eta_{k+1},t) = \Delta f_{k+1}(x_{k+1},\tau)$. Finally, the system (5) can be written as follows

$$\begin{cases} \dot{\eta}_{k+1}(t) = (A+\Delta A)\eta_{k+1}(t) + (B+\Delta B)\,\tilde{u}_{k+1}(t) \\ \qquad\qquad + \tilde{f}_{k+1}(\eta_{k+1},t) + \tilde{\varpi}_{k+1}(t) \\ e_{k+1}(t) = \text{-}C\eta_{k+1}(t) + e_k(t) \end{cases} \tag{6}$$

However, the result system (6) creates a 2D continuous-discrete hybrid model of the repetitive control system within the presence of matched uncertainties in both state and input matrices and external disturbances.

Next, the main objective of this work is to make a sequence of control input functions in order to improve gradually the desired performance with the successive periods. The convergence condition of the tracking error and control input can be described as follows:

$$\lim_{k\to\infty}\|e_k(t)\| = 0,\;\; \lim_{k\to\infty}\|\tilde{u}_{k+1}(t) - \tilde{u}_\infty(t)\| = 0 \tag{7}$$

Where $\tilde{u}_\infty$ is called the learned control.

In general, there are two distinct stability concepts that exist in the stability theory of linear repetitive processes; Asymptotic stability and Stability along the pass. The first concept ensures the existence of a limit profile described by a classical linear system state space model (1D system), the second concept guarantees that the existing limit profile is stable during the successive pass dynamics (2D system) [29]. In order to achieve stability of the controlled system (6), the concept stability along the pass must use.

B. Preliminaries

In order to explain and achieve main results, the following preliminaries are essential.

Lemma 1. [29] (Schur complement): Let E, F and G be given matrices with appropriate dimensions, where E and G are positive definite symmetric matrices. Then, the following inequality

$$F^T GF \text{ - } E < 0 \tag{8}$$

is equivalent to

$$\begin{bmatrix} -E & F^T \\ F & -G^{-1} \end{bmatrix} < 0 \quad or \quad \begin{bmatrix} -G^{-1} & F \\ F^T & -E \end{bmatrix} < 0 \tag{9}$$

Definition 1. [28]: Consider the following linear time invariant repetitive system

$$\begin{cases} \dot{x}_{k+1}(t) = Ax_{k+1}(t) + B\, y_k(t) \\ y_{k+1}(t) = Cx_{k+1}(t) + D\, y_k(t) \end{cases} \tag{10}$$

The system (8) is stable along the pass if and only if three conditions are satisfied:

(i) $\rho(D) < 1$

(ii) $\mathrm{Re}\big(\rho(A)\big) < 0$

(iii) All eigenvalues of $G(s) = C(sI - A)^{-1}B$ with $s = j\omega$, for all real frequencies $\omega \geq 0$, have modulus strictly less than unity.

where the pair $\left[\rho(A), \rho(D)\right]$ represents respectively the spectral radius of a matrices *A* and *D*.

In order to guarantee stability along the pass for system (10), the following lemma based on 2D Lyapunov stability theory is given.

Lemma 2. [30]: For linear time invariant repetitive system in (10), we introduce the 2D Lyapunov function $V_k(t)$ as

$$\begin{cases} V_k(t) \triangleq V_{1,k}(t) + V_{2,k}(t) \\ V_{1,k}(t) \triangleq x_{k+1}^T(t) P x_{k+1}(t) \\ V_{2,k}(t) \triangleq y_k^T(t) Q y_k(t) \end{cases} \tag{11}$$

where P and Q represent two symmetric positive definite matrices to be found. Then, the associated 2D Lyapunov function increment $\Delta V_k(t)$ is given by

$$\begin{cases} \Delta V_k(t) \triangleq \dot{V}_{1,k}(t) + \Delta V_{2,k}(t) \\ \dot{V}_{1,k}(t) \triangleq \dot{x}_{k+1}^T(t) P x_{k+1}(t) + x_{k+1}^T(t) P \dot{x}_{k+1}(t) \\ \Delta V_{2,k}(t) \triangleq y_{k+1}^T(t) Q y_{k+1}(t) - y_k^T(t) Q y_k(t) \end{cases} \tag{12}$$

The linear repetitive process in (10) is stable along the pass if and only if there exist two symmetric positive definite matrices P and Q such that $\Delta V_k(t) < 0$.

The main goal of this paper is to make an ISMC law for the uncertain repetitive system described in (1). Details of the design process will be formulated in the next sections.

IV. Sliding Mode Stability

In the sliding mode literature, the ISMC design is making by two stages. The first step consists on choosing the proper switching function surface for the 2D uncertain system (6). The second step is to design a suitable relay-type controller to guarantee the sliding motion asymptotically stable. Now, details of those steps will be presented.

A. Sliding Surface Choice

The first phase is to choose sliding surface which takes in consideration to eliminate the reaching phase. In this study, the switching surface is specified by the equation:

$$S_{k+1}(t) = B^{+}\eta_{k+1}(t) + \sigma_{k+1}(t) \tag{13}$$

Where $S_{k+1}(0) = 0$ for any initial conditions, $B^{+} \equiv (B^T B)^{-1} B^T$, and the function $\sigma_{k+1}(t) \in \mathbb{R}^m$ represents the solution of the following equation:

$$\begin{cases} \dot{\sigma}_{k+1}(t) = -B^{+}A\eta_{k+1}(t) - \tilde{u}_{0_{k+1}}(t) \\ \sigma_{k+1}(0) = -B^{+}\eta_{k+1}(0) \end{cases} \tag{14}$$

In addition, $\tilde{u}_{0_{k+1}}(t)$ is a nominal control law which should be designed with repetitive control feedback to achieve desired nominal performance. Then, we assume that the system (6) is forced to reach the sliding surface at the initial time t_0. The intrinsic condition of an ideal sliding motion can be achieved as follows

$$\dot{S}_{k+1}(t) = S_{k+1}(t) = 0 \ \text{ for all } t \geq t_0 \tag{15}$$

To analyze the sliding motion, consider the time derivative of (13) given by

$$\begin{aligned} \dot{S}_{k+1} &= B^{+}\dot{\eta}_{k+1} + \dot{\sigma}_{k+1} \\ &= B^{+}(\Delta A\eta_{k+1} + (B + \Delta B)\tilde{u}_{k+1} + \tilde{\varpi}_{k+1} + \tilde{f}_{k+1} - B\tilde{u}_{0_{k+1}}) \\ &= \Delta A_m \eta_{k+1} + (I_m + \Delta B_m)\tilde{u}_{k+1} + \tilde{v}_{k+1} + \tilde{g}_{k+1} - \tilde{u}_{0_{k+1}} \end{aligned} \tag{16}$$

The equivalent control $\tilde{u}_{eq_{k+1}}(t)$ with the time derivative $\dot{S}_{k+1}(t) = 0$ along the state trajectories can be written as

$$\tilde{u}_{eq_{k+1}}(t) = (I_m + \Delta B_m)^{-1}(\tilde{u}_{0_{k+1}} - \Delta A_m \eta_{k+1} - \tilde{v}_{k+1} - \tilde{g}_{k+1}) \tag{17}$$

The equivalent control is the average value such that the input control must hold the sliding motion on the sliding surface.

Remark 1. In (17), the matrix $(I_m + \Delta B_m)$ must be nonsingular which is guaranteed by Assumption 2.

B. Stability of the Sliding Motion

To obtain the sliding mode, the equivalent control (17) is not the input control law that must be applied to the system (6). In order to get the sliding motion expression, the value of $\tilde{u}_{eq_{k+1}}(t)$ can be substituted from (17) into (6), yields

$$\begin{aligned} \dot{\eta}_{k+1}(t) &= (A + \Delta A)\eta_{k+1} + \tilde{\varpi}_{k+1} + \tilde{f}_{k+1} \\ &\quad + (B + \Delta B)(I_m + \Delta B_m)^{-1} \\ &\qquad \times (\tilde{u}_{0_{k+1}} - \Delta A_m \eta_{k+1} - \tilde{v}_{k+1} - \tilde{g}_{k+1}) \\ &= A\eta_{k+1}(t) + B\tilde{u}_{0_{k+1}}(t) \end{aligned} \tag{18}$$

We remark from (18) that the effect of the uncertainties, nonlinearities and external disturbances during the sliding mode is completely rejected. As a result, uncertain system (6) is reduced to

$$\begin{cases}\dot{\eta}_{k+1}(t) = A\eta_{k+1}(t) + B\tilde{u}_{0_{k+1}}(t) \\ e_{k+1}(t) = \text{-}C\eta_{k+1}(t) + e_k(t)\end{cases} \tag{19}$$

The resulting system (19) is insensitive to matched uncertainties. Then, to achieve desired nominal performance, $\tilde{u}_{0_{k+1}}(t)$ is a nominal control input that can be designed by repetitive control.

C. ISMC design based Repetitive Control

The structure of the basic repetitive control system is shown by Fig. 1 where G is the plant model. The transfer function of a basic repetitive controller is

$$C_R(s) = \frac{1}{1-e^{-sT}} \tag{20}$$

where T is the known fundamental period of the reference periodic signal and Φ defines the output signal of the repetitive controller given by the following expression:

$$\Phi_{k+1}(t) = \begin{cases} e_k(t), & 0 \le t < T \\ \Phi_k(t) + e_k(t), & t \ge T\end{cases} \tag{21}$$

The repetitive control law proposed of the system is

$$u_{0_{k+1}}(t) = K_1 x_{k+1}(t) + K_2\Phi_{k+1}(t) \tag{22}$$

where the pair (K_1, K_2) gains matrices with appropriate dimensions will be determined. Thus, these gains matrices ensure the stability along the period of the closed-loop system.

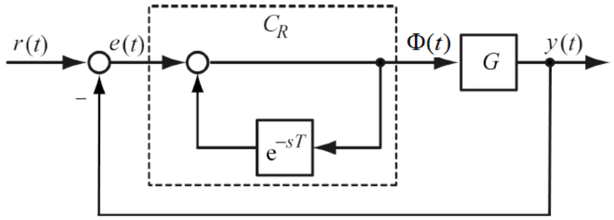


Fig. 1. Basic Repetitive Control System.

Next, the 2D nominal control input can be written as

$$\tilde{u}_{0_{k+1}}(t) = K_1\eta_{k+1}(t) + K_2 e_k(t) \tag{23}$$

By replacing the expression of $\tilde{u}_{0_{k+1}}(t)$ into (19), the new 2D state space nominal dynamics can be written as

$$\begin{cases}\dot{\eta}_{k+1}(t) = \left(A + BK_1\right)\eta_{k+1}(t) + \left(BK_2\right)e_k(t) \\ e_{k+1}(t) = \text{-}C\eta_{k+1}(t) + e_k(t)\end{cases} \tag{24}$$

The next step consists to study the stability along the pass of the sliding mode process in (24) by designing gains controller K_1 and K_2. Therefore, to achieve the design process, the following theorem gives a sufficient condition for the stability along the pass of the sliding mode process in (24) by using the LMI method and the 2D system theory. After solving it, the designed sliding function in (13) is complete.

Theorem 1. The sliding mode process is stable along the pass if and only if there exist two symmetric positive definite matrices X_1 and X_2, and two matrices W_1 and W_2 such that the following LMI

$$\begin{bmatrix} AX_1 + X_1A^T + BW_1 + W_1^TB^T & (*) & (*) \\ W_2^TB^T & -X_2 & (*) \\ CX_1 & -X_2 & -X_2\end{bmatrix} < 0 \tag{25}$$

holds, then the closed-loop system (24) is stable along the pass, that is, the stabilization gains are given by

$$\begin{cases} K_1 = W_1.X_1^{-1} \\ K_2 = W_2.X_2^{-1}\end{cases} \tag{26}$$

Proof. Consider two symmetric positive-definite matrices P and Q and choose a candidate 2D Lyapunov function $V_k(t)$ by applying lemma 1 such that

$$\begin{cases} V_{1,k}(t) \triangleq \eta_{k+1}^T(t)P\eta_{k+1}(t) \\ V_{2,k}(t) \triangleq e_k^T(t)Qe_k(t)\end{cases} \tag{27}$$

Then, the associated 2D Lyapunov function increment $\Delta V_k(t)$ given by

$$\begin{cases} \dot{V}_{1,k}(t) \triangleq \dot{\eta}_{k+1}^T(t)P\eta_{k+1}(t) + \eta_{k+1}^T(t)P\dot{\eta}_{k+1}(t) \\ \Delta V_{2,k}(t) \triangleq e_{k+1}^T(t)Qe_{k+1}(t) - e_k^T(t)Qe_k(t)\end{cases} \tag{28}$$

Therefore, the Lyapunov function increment $\Delta V_k(t)$ can be transformed into

$$\begin{aligned}\Delta V_k(t) &= \dot{V}_{1,k}(t) + \Delta V_{2,k}(t) \\ &= \dot{\eta}_{k+1}^T(t)P\eta_{k+1}(t) + \eta_{k+1}^T(t)P\dot{\eta}_{k+1}(t) \\ &\quad + e_{k+1}^T(t)Qe_{k+1}(t) - e_k^T(t)Qe_k(t) \\ &= \left[(A+BK_1)\eta_{k+1}(t) + (BK_2)e_k(t)\right]^T P\eta_{k+1}(t) \\ &\quad + \eta_{k+1}^T(t)P\left[(A+BK_1)\eta_{k+1}(t) + (BK_2)e_k(t)\right] \\ &\quad + \left[\text{-}C\eta_{k+1}(t) + e_k(t)\right]^T Q\left[\text{-}C\eta_{k+1}(t) + e_k(t)\right] \\ &\quad - e_k^T(t)Qe_k(t) \\ &= \varsigma^T(t)\psi\varsigma(t)\end{aligned} \tag{29}$$

where

$$\begin{cases} \varsigma(t) = \left[\eta_{k+1}^T(t) \quad e_k^T(t)\right]^T \\ \psi = \begin{bmatrix} (A+BK_1)^TP + P(A+BK_1) + C^TQC & (*) \\ (BK_2)^TP - QC & 0\end{bmatrix}\end{cases} \tag{30}$$

Notice that (29) implies $\Delta V_k(t) < 0$ (i.e. $\psi < 0$) for any $\varsigma(t) \neq 0$. By Lemma 1, stability along the pass of the sliding mode process in (24) is guaranteed. On the other hand, by Lemma 2 (Schur complement), LMI $\psi < 0$ is equivalent to

$$\begin{bmatrix} (A+BK_1)^T P + P(A+BK_1) & (*) & (*) \\ (BK_2)^T P & -Q & (*) \\ QC & -Q & -Q \end{bmatrix} < 0 \tag{31}$$

After that, (31) pre-multiply and post-multiply by $\Sigma = diag\left\{P^{-1}, Q^{-1}, Q^{-1}\right\}$ and its transpose respectively. Thus, LMI $\psi < 0$ is finally equivalent to

$$\begin{bmatrix} P^{-1}(A+BK_1)^T + (A+BK_1)P^{-1} & (*) & (*) \\ Q^{-1}(BK_2)^T & -Q^{-1} & (*) \\ CP^{-1} & -Q^{-1} & -Q^{-1} \end{bmatrix} < 0 \tag{32}$$

Let $X_1 = P^{-1}$, $X_2 = Q^{-1}$, $W_1 = K_1 X_1$, and $W_2 = K_2 X_2$. LMI (25) is obtained after replacing correspondent's terms in (32). Thus, the proof is finished.

V. Reachability Analysis

In the last section, a sufficient condition for the stability along the pass of the sliding mode process was derived. The next step consists to analyze the reachability of the sliding surface. Therefore, the reachability is a sufficient condition which to guarantee that the sliding mode process will converge to the sliding surface at each time instant.

A. Repetitive Integral Sliding Mode Control Law

In order to satisfy reachability condition, the following theorem proposes an RISMC law.

Theorem 2. Consider the 2D uncertain system (6) with the assumptions (1–2). Suppose that the sliding surface is given by (13) and X_1 , X_2 , W_1 and W_2 are solutions of the LMI (25). The RISMC is defined by

$$\tilde{u}_{k+1}(t) = K_1\eta_{k+1}(t) + K_2 e_k(t) - \rho\frac{S_{k+1}(t)}{\|S_{k+1}(t)\|} \tag{33}$$

where

$$K_1 = W_1.X_1^{-1},\ K_2 = W_2.X_2^{-1}$$
$$\rho = \frac{1}{1-b_m}\Big[\varepsilon + (a_m + \alpha_m\|B^+\| + b_m\|K_1\|)\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| + \beta_m\|B^+\|\Big] \tag{34}$$

with ε is a small positive scalar.

Proof. Let choose a Lyapunov function candidate to be

$$V_{k+1}(t) = \frac{1}{2}S_{k+1}^T(t).S_{k+1}(t) \tag{35}$$

In order to prove that the proposed RISMC law satisfies the reachability condition, substituting the value of (33) into (16) gives

$$\begin{aligned} \dot{S}_{k+1} &= \Delta A_m\eta_{k+1} + (I_m + \Delta B_m)\tilde{u}_{k+1} + \tilde{v}_{k+1} + \tilde{g}_{k+1} - \tilde{u}_{0_{k+1}} \\ &= (\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k + \tilde{v}_{k+1} + \tilde{g}_{k+1} \\ &\quad - \rho(I_m + \Delta B_m)\frac{S_{k+1}}{\|S_{k+1}\|} \end{aligned} \tag{36}$$

Pre-multiplying both sides of (36) by S_{k+1}^T yields

$$\begin{aligned} S_{k+1}^T\dot{S}_{k+1} &= S_{k+1}^T[(\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k \\ &\quad + \tilde{v}_{k+1} + \tilde{g}_{k+1} - \rho(I_m + \Delta B_m)\frac{S_{k+1}}{\|S_{k+1}\|}] \\ &= S_{k+1}^T[(\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k \\ &\quad + \tilde{v}_{k+1} + \tilde{g}_{k+1}] - \rho(I_m + \Delta B_m)\frac{S_{k+1}^T S_{k+1}}{\|S_{k+1}\|} \end{aligned} \tag{37}$$

and by using the property $S_{k+1}^T S_{k+1} = \|S_{k+1}\|^2$, (37) becomes

$$\begin{aligned} \dot{V}_{k+1} &= S_{k+1}^T[(\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k \\ &\quad + \tilde{v}_{k+1} + \tilde{g}_{k+1}] - \rho(I_m + \Delta B_m)\|S_{k+1}\| \\ &= S_{k+1}^T[(\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k \\ &\quad + \tilde{v}_{k+1} + \tilde{g}_{k+1}] - \rho\|S_{k+1}\| - \rho\Delta B_m\|S_{k+1}\| \\ &\le S_{k+1}^T[(\Delta A_m + \Delta B_m K_1)\eta_{k+1} + \Delta B_m K_2 e_k \\ &\quad + \tilde{v}_{k+1} + \tilde{g}_{k+1}] - \rho(1-b_m)\|S_{k+1}\| \\ &\le \|S_{k+1}\|[(\|\Delta A_m\| + \|\Delta B_m K_1\|)\|\eta_{k+1}\| + \|\Delta B_m K_2\|\|e_k\| \\ &\quad + \|\tilde{v}_{k+1}\| + \|\tilde{g}_{k+1}\|] - \rho(1-b_m)\|S_{k+1}\| \end{aligned} \tag{38}$$

Finally, by using Assumption 2, (38) can be written as

$$\begin{aligned} \dot{V}_{k+1} \le \|S_{k+1}\|[(a_m + \alpha_m\|B^+\| + b_m\|K_1\|)\|\eta_{k+1}\| \\ + b_m\|K_2\|\|e_k\| + \beta_m\|B^+\| - \rho(1-b_m)] \end{aligned} \tag{39}$$

The reachability condition is guaranteed if and only if $\dot{V}_{k+1} < 0$ is satisfied. Then, in order to enforce it, any choice of ρ must satisfy the following condition

$$\rho > \frac{1}{1-b_m}[(a_m + \alpha_m\|B^+\| + b_m\|K_1\|)\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| + \beta_m\|B^+\|] \tag{40}$$

Let ε is a small positive scalar, the inequality in (39) becomes

$$S_{k+1}^T(t)\dot{S}_{k+1}(t) \le -\varepsilon\|S_{k+1}(t)\| < 0 \tag{41}$$

Besides, the initial value of $S_{k+1}(t)$ is provided by $S_{k+1}(0) = 0$ for any initial conditions. So, the reachability of the sliding surface is guaranteed. That ended the proof.

B. Adaptative Repetitive Integral Sliding Mode Control Law

Applicability of the designed RISMC law presents two major problems. The first is the chattering phenomenon and the second is the difficulty to get the exact upper bound values of uncertainties and external disturbances [31].

In order to overcome these problems, we will present an Adaptive RISMC (ARISMC) law in this section. To get there, we start by rewriting (6) as

$$\begin{aligned}\dot{\eta}_{k+1} &= (A+\Delta A)\eta_{k+1} + (B+\Delta B)\tilde{u}_{k+1} \\ &\qquad + B\tilde{v}_{k+1} + B\tilde{g}_{k+1} \\ &= A\eta_{k+1} + (B+\Delta B)\,\tilde{u}_{k+1} \\ &\qquad + B(\Delta A_m\eta_{k+1} + \tilde{v}_{k+1} + \tilde{g}_{k+1})\end{aligned} \tag{42}$$

Let $\mu_{k+1} = \Delta A_m\eta_{k+1} + \tilde{v}_{k+1} + \tilde{g}_{k+1}$, (42) becomes

$$\begin{cases}\dot{\eta}_{k+1} = A\eta_{k+1} + (B+\Delta B)\,\tilde{u}_{k+1} + B\mu_{k+1} \\ e_{k+1} = \text{-}C\eta_{k+1} + e_k\end{cases} \tag{43}$$

To complete the description of the 2D uncertain system (43), the following assumption is used:

Assumption 3: There are unknown positive constants δ_1 and δ_2, for all $(x,t) \in \mathbb{R}^n \times \mathbb{R}$, such that

$$\|\mu_{k+1}\| \le \delta_1\|\eta_{k+1}\| + \delta_2 \tag{44}$$

To achieve design of ARISMC law, two steps are necessaries. First step, simple adaptation laws will be proposed for the upper bound of $\|\mu_{k+1}\|$. Second step consists to design a control law using this result, adaptive upper bound [31]. Let consider $\bar{\delta}_1$ and $\bar{\delta}_2$, respectively, the adaptive parameters about δ_1 and δ_2. The proposed adaptive upper bound of $\|\mu_{k+1}\|$ is defined by

$$\bar{\mu}_{k+1} = \bar{\delta}_1\|\eta_{k+1}\| + \bar{\delta}_2 \tag{45}$$

Now, define the parameter adaptation errors as $\tilde{\delta}_1 = \bar{\delta}_1 - \delta_1$ and $\tilde{\delta}_2 = \bar{\delta}_2 - \delta_2$. The simple adaptation laws proposed for the upper bound of $\|\mu_{k+1}\|$ is

$$\begin{aligned}\dot{\tilde{\delta}}_1 &\triangleq \phi_1\|\eta_{k+1}\|\|S_{k+1}\| \\ \dot{\tilde{\delta}}_2 &\triangleq \phi_2\|S_{k+1}\|\end{aligned} \tag{46}$$

where ϕ_1 and ϕ_2 are positive adaptation gains. However, δ_1 and δ_2 are assumed as constant values. Then, the result adaptation laws can be rewritten as

$$\begin{aligned}\dot{\bar{\delta}}_1 &\triangleq \phi_1\|\eta_{k+1}\|\|S_{k+1}\| \\ \dot{\bar{\delta}}_2 &\triangleq \phi_2\|S_{k+1}\|\end{aligned} \tag{47}$$

After that, by integrating (47), the adaptive parameters are described by the following expressions:

$$\begin{aligned}\bar{\delta}_1 &= \bar{\delta}_{1i} + \phi_1\int_{t_0+(k+1)T}^{t+(k+1)T}\|\eta_{k+1}\|\|S_{k+1}\|dt \\ \bar{\delta}_2 &= \bar{\delta}_{2i} + \phi_2\int_{t_0+(k+1)T}^{t+(k+1)T}\|S_{k+1}\|dt\end{aligned} \tag{48}$$

where $\bar{\delta}_{1i}$ and $\bar{\delta}_{2i}$ represents, respectively, the initial values of $\bar{\delta}_1$ and $\bar{\delta}_2$. Then, we can expose the following theorem.

Theorem 3. Consider the 2D uncertain system (6) with the assumptions (1-3). Suppose that the sliding surface is given by (13) and X_1, X_2, W_1 and W_2 are solutions of the LMI (25). The ARISMC is defined by

$$\tilde{u}_{k+1}(t) = K_1\eta_{k+1}(t) + K_2e_k(t) - \hat{\rho}\frac{S_{k+1}(t)}{\|S_{k+1}(t)\|} \tag{49}$$

where

$$\begin{aligned}&K_1 = W_1.X_1^{-1},\, K_2 = W_2.X_2^{-1} \\ &\hat{\rho}_1 = \varepsilon + (b_m\|K_1\| + \bar{\delta}_1)\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| + \bar{\delta}_2 \\ &\hat{\rho} = \frac{1}{1-b_m}\hat{\rho}_1\end{aligned} \tag{50}$$

with ε is a small positive scalar. By employing the adaptation laws (46) and the control law (49), $S_{k+1} = 0$ is stable along the pass.

Proof. Let choose an improved Lyapunov function candidate instead of $S_{k+1}^T\dot{S}_{k+1}$ to be

$$V_{k+1} = \frac{1}{2}\left(S_{k+1}^TS_{k+1} + \phi_1^{-1}\tilde{\delta}_1^2 + \phi_2^{-1}\tilde{\delta}_2^2\right) \tag{51}$$

where ϕ_1 and ϕ_2 are defined in (46). Substituting the value of (49) into (16) gives

$$\begin{aligned}\dot{S}_{k+1} &= \Delta A_m\eta_{k+1} + (I_m+\Delta B_m)\,\tilde{u}_{k+1} + \tilde{v}_{k+1} + \tilde{g}_{k+1} - \tilde{u}_{0_{k+1}} \\ &= \mu_{k+1} - K_1\eta_{k+1} - K_2e_k \\ &\quad + (I_m+\Delta B_m)\,(K_1\eta_{k+1} + K_2e_k - \frac{1}{1-b_m}\hat{\rho}_1\frac{S_{k+1}(t)}{\|S_{k+1}(t)\|}) \\ &= \mu_{k+1} + \Delta B_mK_1\eta_{k+1} + \Delta B_mK_2e_k \\ &\quad - \hat{\rho}_1\frac{1}{1-b_m}(I_m+\Delta B_m)\frac{S_{k+1}}{\|S_{k+1}\|}\end{aligned} \tag{52}$$

Pre-multiplying both sides of (52) by S_{k+1}^T yields

$$\begin{aligned}S_{k+1}^T\dot{S}_{k+1} &= S_{k+1}^T(\mu_{k+1} + \Delta B_mK_1\eta_{k+1} + \Delta B_mK_2e_k) \\ &\qquad - \hat{\rho}_1\frac{1}{1-b_m}(I_m+\Delta B_m)\|S_{k+1}\|\end{aligned} \tag{53}$$

Then, the Lyapunov function increment is given by

$$
\begin{aligned}
\dot{V}_{k+1} &= S_{k+1}^T\dot{S}_{k+1} + \phi_1^{-1}\tilde{\delta}_1\dot{\tilde{\delta}}_1 + \phi_2^{-1}\tilde{\delta}_2\dot{\tilde{\delta}}_2 \\
&= S_{k+1}^T(\mu_{k+1} + \Delta B_m K_1\eta_{k+1} + \Delta B_m K_2 e_k) \\
&\quad - \hat{\rho}_1\frac{1}{1-b_m}(I_m + \Delta B_m)\|S_{k+1}\| + \phi_1^{-1}\tilde{\delta}_1\dot{\tilde{\delta}}_1 + \phi_2^{-1}\tilde{\delta}_2\dot{\tilde{\delta}}_2 \\
&= S_{k+1}^T(\mu_{k+1} + \Delta B_m K_1\eta_{k+1} + \Delta B_m K_2 e_k) \\
&\quad - \hat{\rho}_1\frac{1}{1-b_m}(I_m + \Delta B_m)\|S_{k+1}\| \\
&\quad + \tilde{\delta}_1\|\eta_{k+1}\|\|S_{k+1}\| + \tilde{\delta}_2\|S_{k+1}\| \\
&= S_{k+1}^T(\mu_{k+1} + \Delta B_m K_1\eta_{k+1} + \Delta B_m K_2 e_k) \\
&\quad - \hat{\rho}_1\frac{1}{1-b_m}(I_m + \Delta B_m)\|S_{k+1}\| \\
&\quad + (\bar{\delta}_1 - \delta_1)\|\eta_{k+1}\|\|S_{k+1}\| + (\bar{\delta}_2 - \delta_2)\|S_{k+1}\|
\end{aligned} \tag{54}
$$

Then, by using Assumption 2, (54) can be written as

$$
\begin{aligned}
\dot{V}_{k+1} &\le (\|\mu_{k+1}\| + b_m\|K_1\|\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| \\
&\quad + \frac{b_m}{1-b_m}\hat{\rho}_1)\|S_{k+1}\| - \frac{1}{1-b_m}\hat{\rho}_1\|S_{k+1}\| \\
&\quad + \left[(\bar{\delta}_1\|\eta_{k+1}\| + \bar{\delta}_2) - (\delta_1\|\eta_{k+1}\| + \delta_2)\right]\|S_{k+1}\| \\
&\le \left[\|\mu_{k+1}\| - (\delta_1\|\eta_{k+1}\| + \delta_2)\right]\|S_{k+1}\| \\
&\quad + (b_m\|K_1\|\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| \\
&\qquad + \bar{\delta}_1\|\eta_{k+1}\| + \bar{\delta}_2 - \hat{\rho}_1)\|S_{k+1}\|
\end{aligned} \tag{55}
$$

The reachability condition is guaranteed if and only if $\dot{V}_{k+1} < 0$ is satisfied. Finally, by using Assumption 3, it is sufficient to choose the value of $\hat{\rho}_1$ as

$$
\hat{\rho}_1 > b_m\|K_1\|\|\eta_{k+1}\| + b_m\|K_2\|\|e_k\| + \bar{\delta}_1\|\eta_{k+1}\| + \bar{\delta}_2 \tag{56}
$$

Let ε is a small positive scalar, the inequality in (55) becomes

$$
\dot{V}_{k+1} \le -\varepsilon\|S_{k+1}(t)\| < 0 \tag{57}
$$

In addition, the initial value of $S_{k+1}(t)$ is provided by $S_{k+1}(0) = 0$ for any initial conditions. So, the reachability of the sliding surface is guaranteed. That ended the proof.

Remark 2. In this proposition, the rate of parameter adaptation is adjusted by an appropriate choose of $\{\bar{\delta}_{1i}, \bar{\delta}_{2i}\}$ and $\{\phi_1, \phi_2\}$. In practice, choices of adaptation gains are limited by many practical considerations such as the bound of control input and other parameters.

VI. Illustrative Example

In order to prove the validity and the effectiveness of the new proposed controlled design schemes, consider the following nominal model of a submarine from [32].

$$
\begin{cases}
\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 \\ -0.0071 & -0.111 & 0.12 \\ 0 & 0.07 & -0.3 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ -0.095 \\ 0.072 \end{bmatrix} u(t) \\
y(t) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x(t)
\end{cases} \tag{58}
$$

where the state vector is defined by $x(t) = \begin{bmatrix} \theta & \dfrac{d\theta}{dt} & \alpha \end{bmatrix}^T$ with θ is the inclination of the submarine and α is the angle of attack. It's assumed that can only measure $\dfrac{d\theta}{dt}$ and suppose that

$$
\Delta A = \begin{bmatrix} 0 & -\zeta_1 & 0 \\ 0.0071\zeta_2 & 0.111\zeta_2 & -0.12\zeta_2 \\ 0 & -0.07\zeta_3 & 0.3\zeta_3 \end{bmatrix}
$$

$$
\Delta B = \begin{bmatrix} 0 \\ 0.095\zeta_2 \\ -0.072\zeta_3 \end{bmatrix}, \ \zeta_i = \frac{\lambda_i}{1+\lambda_i} \ \text{ for } i = 1,\dots,3
$$

$$
\varpi(t) = \begin{bmatrix} 0 \\ -0.02\sin\left(\dfrac{3\pi}{4T}t + \dfrac{T}{4}\right) \\ 0.02\sin\left(\dfrac{3\pi}{4T}t + \dfrac{T}{4}\right) \end{bmatrix}, \quad f(x,t) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
$$

Consider the following periodic reference input:

$$
r(t) = 0.5\sin\left(\frac{2\pi}{T}t\right) + 0.5\sin\left(\frac{4\pi}{T}t\right) \tag{59}
$$

Therefore, a feasible solution of the LMI (25) is defined as

$$
X_1 = \begin{bmatrix} 64.9274 & -0.0000 & -1.6712 \\ -0.0000 & 26.0324 & -19.7298 \\ -1.6712 & -19.7298 & 98.3993 \end{bmatrix}, X_2 = 72.1792
$$

$$
W_1 = \begin{bmatrix} 267.0619 & 276.3066 & 112.2730 \end{bmatrix}, W_2 = -274.0254
$$

$$
\begin{cases}
K_1 = \begin{bmatrix} 4.2146 & 13.5996 & 3.9394 \end{bmatrix}, \|K_1\| = 14.7726 \\
K_2 = -3.7965, \qquad \|K_2\| = 3.7965
\end{cases}
$$

For $\lambda_i = 0.1$, $i = 1,\dots,3$, we get

$$
\Delta A_m = \begin{bmatrix} -0.0047 & -0.1097 & 0.2322 \end{bmatrix}, \ \|\Delta A_m\| = 0.2569
$$

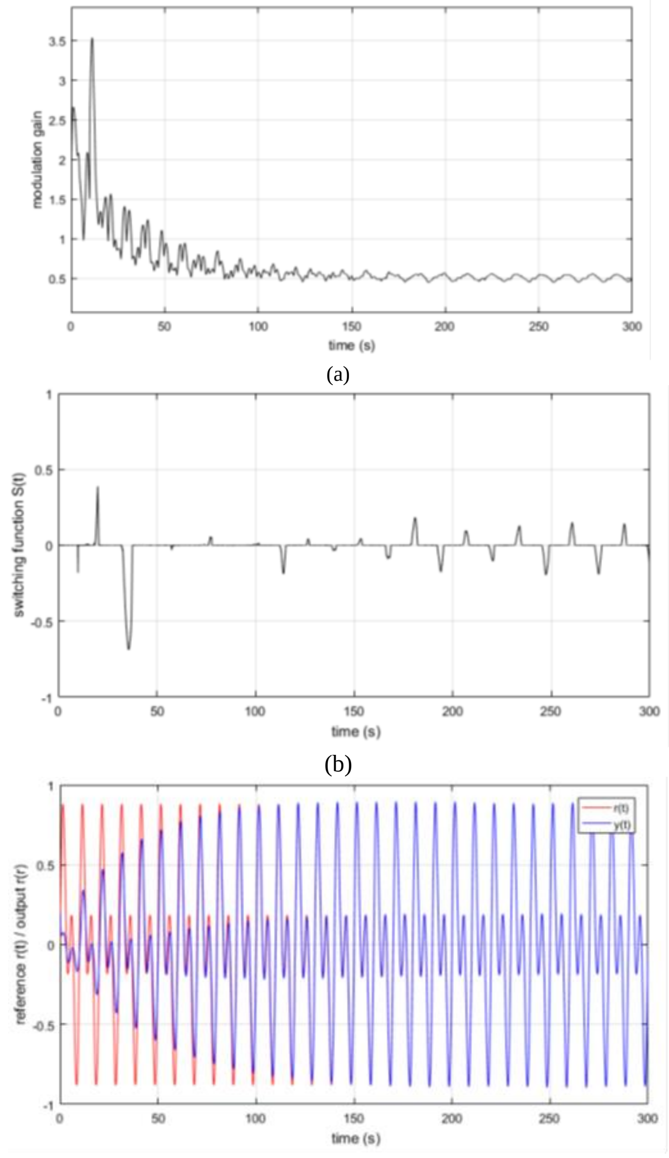
$$
\Delta B_m = -0.1000, \ \|\Delta B_m\| = 0.1000, \ \|B^+\| = 8.3892
$$

$$\|\varpi(t)\| = 0.0283\left|\sin\left(\frac{3\pi}{4T}t+\frac{T}{4}\right)\right|$$

Then, using assumption 1, we get $a_m = 0.3569$, $b_m = 0.2000$, $\alpha_m = 0.0010$, $\beta_m = 0.0283$ and $\varepsilon = 0.1000$. In addition, using (48), let $\bar{\delta}_{1i} = 0.1$, $\phi_1 = 0.1$, $\bar{\delta}_{2i} = 0.1$ and $\phi_2 = 0.1$.

The simulation results are given for the fundamental repetition period $T = 10s$ and the initial state vector $x_1(0) = [0.1 \quad 0.2 \quad 0.3]^T$.

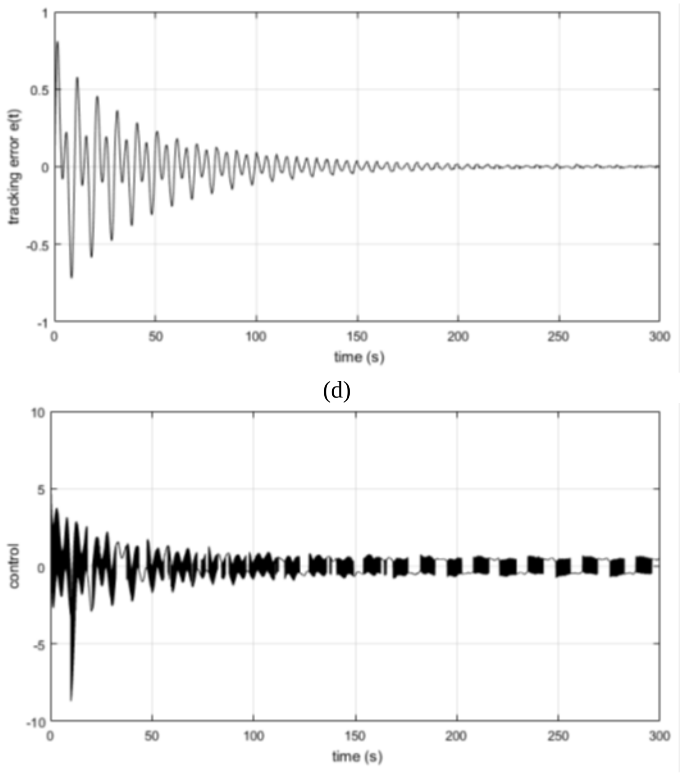
Figures 2 and 3 show the evolution of modulation gain ρ, switching function $S(t)$, reference $r(t)$ and output $y(t)$, tracking error $e(t)$, and controller $\tilde{u}(t)$ using the proposed RISMC (33-34) and ARISMC (49-50), respectively.



(a)

(b)

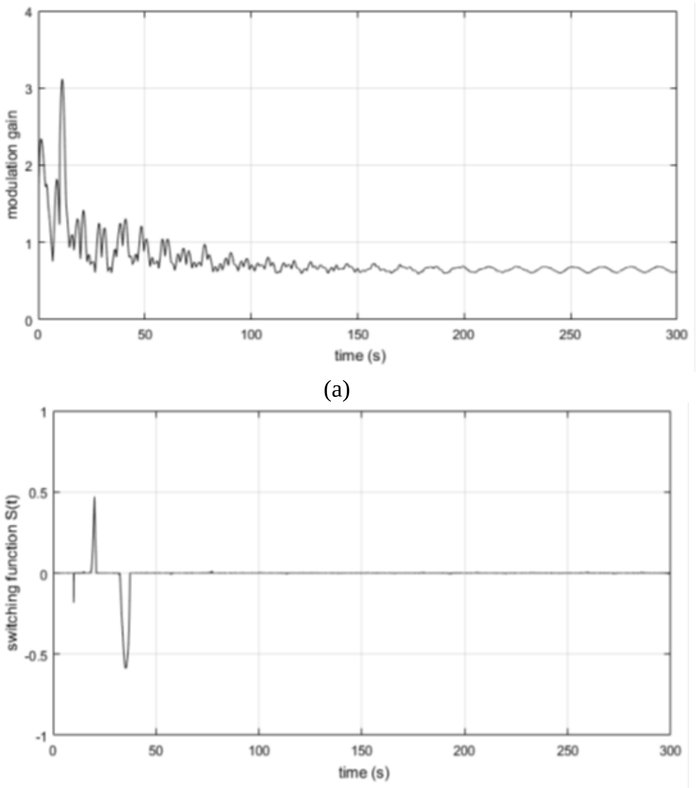
(c)



(d)

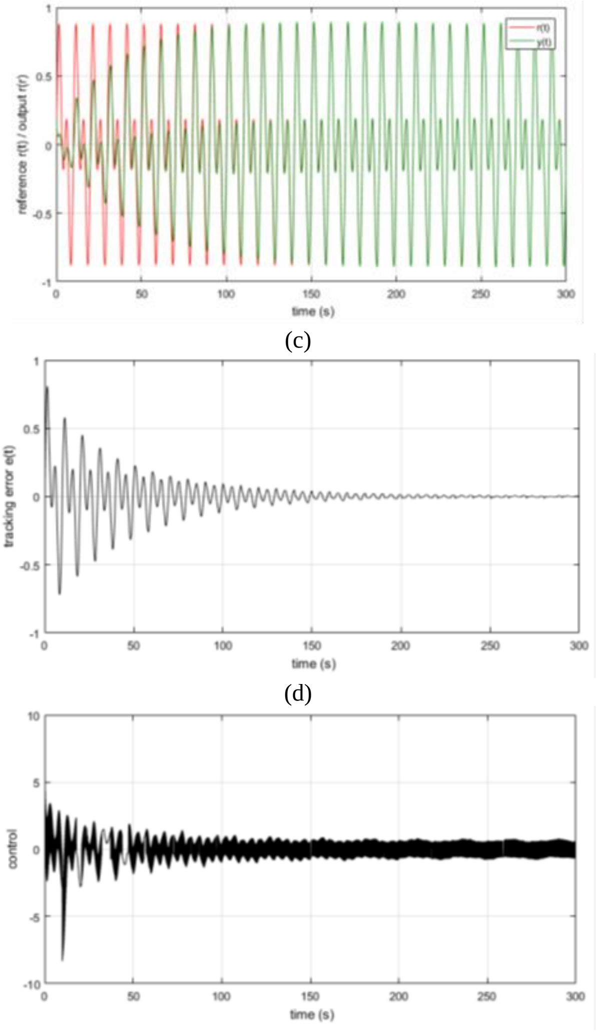
(e)

Fig. 2. Simulation Results using RISMC: Modulation Gain ρ (a), Switching Function $S(t)$ (b), Reference $r(t)$ and Output $y(t)$ (c), Tracking Error $e(t)$ (d), Controller $\tilde{u}(t)$ (e).



(a)

(b)



(c)

(d)

(e)

Fig. 3. Simulation results using ARISMC: modulation gain ρ (a), switching function $S(t)$ (b), reference $r(t)$ and output $y(t)$ (c), tracking error $e(t)$ (d), controller $\tilde{u}(t)$ (e).

Thus, it is easy to remark that the proposed control laws ensure the stability along the pass of the closed-loop system. However, the chattering phenomenon appears in the evolution of the controller for the RISMC approach. So, the ARISMC approach has overcome this problem and the controller evolution confirms the elimination of discontinuities in high frequency.

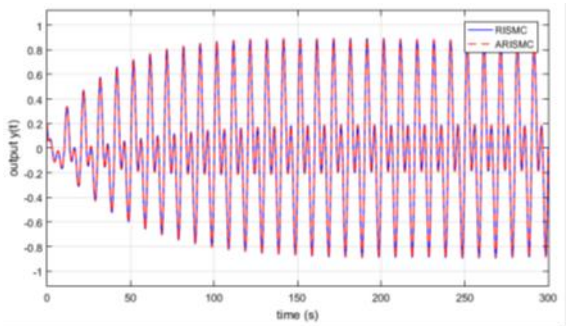


Fig. 4. Output Evolution According to RISMC and ARISMC.

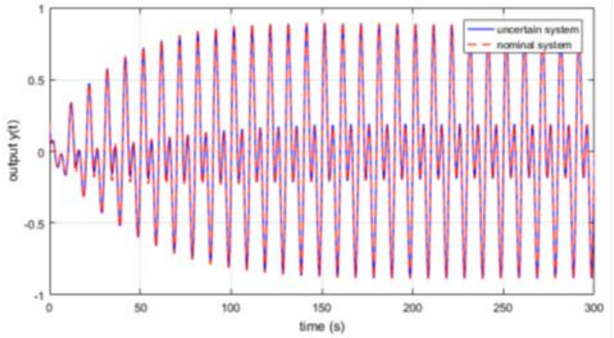


Fig. 5. Output Evolution According to ARISMC for the Uncertain and Nominal System.

Figure 4 shows the evolution of the output $y(t)$ for both RISMC and ARISMC methods, and the evolution of $y(t)$ for both uncertain and nominal systems are presented in Fig. 5. It is obvious from Fig. 4 that the ARISMC approach almost gives the same performances as the RISMC approach. In addition, evolution outputs from the uncertain and nominal systems are superposed (see Fig. 5). Then, simulation results confirm the robustness of the proposed methods.

VII. Conclusion

In this paper, the problem of integral sliding mode control of uncertain repetitive processes in the presence of matched uncertainties, external disturbances and norm-bounded nonlinearities was studied.

A combination of repetitive control and sliding mode approach was exploited to reject the effect of uncertainties, nonlinearities and disturbances and a sufficient condition of the existence of sliding mode was studied based on basic repetitive control and a sliding mode controller which was synthesized by means of linear matrix inequalities, that guarantees the stability along the periods of the controlled closed-loop process. An adaptive integral sliding mode controller based on repetitive control was also proposed to improve performances of the synthesized control scheme.

The simulation results using the new approaches have given good performances and confirm the efficiency of the new proposed controlled design schemes. In addition, the RISMC problem for linear repetitive processes with mismatched uncertainties can be examined as our future work.

References

[1] L. Zhou, J. She, C. Li, and C. Pan, “Robust aperiodic-disturbance rejection in an uncertain modified repetitive-control system,” International Journal of Applied Mathematics and Computer Science, vol. 26, no. 2, pp. 285–295, 2016.

[2] S. G. Yuan, M. Wu, B. G. Xu, and R. J. Liu, “Design of Discrete-time Repetitive Control System Based on Two-dimensional Model,” International Journal of Automation and Computing, vol. 9, no. 2, pp. 165-170, 2012.

[3] Z. Li, W. Zhang, Y. Zhang, and X. Xu, “Robust Repetitive Control Design and Its Application on Linear Servo Systems,” International Journal of Precision Engineering and Manufacturing, vol. 16, no. 1, pp. 21-29, 2015.

[4] J. D. Ratcliffe, J. J. Hatonen, P. L. Lewin, E. Rogers, and D. H. Owens, “Repetitive control of synchronized operations for process applications,”

International Journal of Adaptive Control and Signal Processing, vol. 21, pp. 300–325, 2007.

[5] Y. Song, X. Yu, G. Chen, J. X. Xu, and Y. P. Tian, "Time delayed repetitive learning control for chaotic systems," International Journal of Bifurcation and Chaos, vol. 12, no. 5, pp. 1057-1065, 2002.

[6] J. X. Xu and R. Yan, "On repetitive learning control for periodic tracking tasks," IEEE Transactions on Automatic Control, vol. 51, no. 11, pp. 1842-1848, 2006.

[7] B. A. Francis and W. M. Wonham, "The internal model principle for linear multivariable regulators," Applied Mathematics and Optimization, vol. 2, pp. 170-194, 1975.

[8] C. Mnasri and M. Gasmi, "LMI-based adaptative fuzzy integral sliding mode control of mismatched uncertain systems," International Journal of Applied Mathematics and Computer Science, vol. 21, no. 4, pp. 605-615, 2011.

[9] M. Wu, Y. H. Lan, J. She, Y. He, and L. Xu, "Optimal repetitive control based on two-dimensional model," International Journal of Innovative Computing, Information and Control, vol. 8, no. 3A, 2012.

[10] Z. H. Wang, L. Z. Yi, Y. H. Lan, and C. X. Chen, "Design of observer-based discrete repetitive-control system based on 2D model," Journal of Central South University of Technology, vol. 21, pp. 4236−4243, 2014.

[11] L. Zhou, J. She, and S. Zhou, "A 2D system approach to the design of a robust modified repetitive-control system with a dynamic output-feedback controller," International Journal of Applied Mathematics and Computer Science, vol. 24, no. 2, pp. 325–334, 2014.

[12] M. A. Emelianov, P. V. Pakshin, K. Galkowski, and E. Rogers, "Stabilization of Differential Repetitive Processes," Automation and Remote Control, vol. 76, no. 5, pp. 786–800, 2015.

[13] P. Yu, M. Wu, J. She, and Q. Lei, "Robust repetitive control and disturbance rejection based on two-dimensional model and equivalent-input-disturbance approach," Asian Journal of Control, vol. 18, no. 6, pp. 1–11, 2016.

[14] L. Zhou and J. She, "Aperiodic Disturbance Rejection in a Modified Repetitive-control System," International Journal of Control, Automation and Systems, vol. 14, no. 4, pp. 883-892, 2016.

[15] Q. Zhu, J. X. Xu, S. Yang, and G. D. Hu, "Adaptive backstepping repetitive learning control design for nonlinear discrete-time systems with periodic uncertainties," International Journal of Adaptive Control and Signal Processing, vol. 29, pp. 524–535, 2015.

[16] Y. H. Yang and C. L. Chen, "Spatial Domain Adaptive Control of Nonlinear Rotary Systems Subject to Spatially Periodic Disturbances," Journal of Applied Mathematics, 2012.

[17] L. Zhou, J. She, S. Zhou, and Q. Chen, "H∞ Controller Design for an Observer-Based Modified Repetitive-Control System," International Journal of Engineering Mathematics, 2014.

[18] Y. Wang, R. Wang, X. Xie, and H. Zhang, "Observer-Based H∞ Fuzzy Control for Modified Repetitive Control Systems," Neurocomputing, vol. 286, no. 3, pp. 141-149, 2018.

[19] Z. Shao, S. Huang, and Z. Xiang, "Robust H∞ Repetitive Control for a Class of Linear Stochastic Switched Systems with Time Delay," Circuits, Systems, and Signal Processing, vol. 34, no. 4, pp. 1363-1377, 2015.

[20] Chi-Ying Lin and Hong-Wu Jheng, "Active Vibration Suppression of a Motor-Driven Piezoelectric Smart Structure Using Adaptive Fuzzy Sliding Mode Control and Repetitive Control," Applied Sciences, vol. 7, no. 3, 2017.

[21] R. Sakthivel, K. Raajananthini, P. Selvaraj, and Y. Ren, "Design and analysis for uncertain repetitive control systems with unknown disturbances," Journal of Dynamic Systems, Measurement, and Control, vol. 140, 2018.

[22] S. Mobayen, "A novel global sliding mode control based on exponential reaching law for a class of underactuated systems with external disturbances," Journal of Computational and Nonlinear Dynamics, vol. 11, no. 2, 2016.

[23] M. Golestani, S. Mobayen, and F. Tchier, "Adaptive finite-time tracking control of uncertain non-linear n-order systems with unmatched uncertainties," IET Control Theory and Applications, vol. 10, no. 14, pp. 1675–1683, 2016.

[24] B. Vaseghi, M. A. Pourmina, and S. Mobayen, "Secure communication in wireless sensor networks based on chaos synchronization using adaptive sliding mode control," Nonlinear Dynamics, 2017.

[25] J. L. Chang, "Dynamic output feedback integral sliding mode control design for uncertain systems," International Journal of Robust and Nonlinear Control, vol. 22, pp. 841–857, 2012.

[26] R. Chuei, Z. Cao, M. Mitrevska, and Z. Man, "Sliding mode based repetitive control for improved reference tracking," Proceedings of 2014 International Conference on Modelling, Identification and Control, Melbourne, Australia, pp. 166-177, 2014.

[27] W. Sun, H. Cai, F. Zhao, Z. Zhong, "Repetitive Control Design of Simulation Turntable Based on Integral," Proceedings of 2012 International Conference on Modelling, Identification and Control, , Wuhan, China, pp. 849-854, 2012.

[28] L. Wu, H. Gao, and C. Wang, "Quasi sliding mode control of differential linear repetitive processes with unknown input disturbance," IEEE Transactions on Industrial Electronics, vol. 58, no. 7, pp. 3059-3068, 2011.

[29] E. Rogers, K. Galkowski, and D. H. Owens, "Control systems theory and applications for linear repetitive processes," Lecture Notes in Control and Information Sciences, vol. 349, Springer, Germany, 2007.

[30] W. Paszke, K. Galkowski, E. Rogers, and D. H. Owens, "H∞ control of differential linear repetitive processes," IEEE Transactions on Circuits and Systems II, vol. 53, no. 1, pp. 39–44, 2006.

[31] D. S. Yoo and M. J. Chung, "A variable structure control with simple adaptation laws for upper bounds on the norm of the uncertainties," IEEE Transactions on Automatic Control, vol. 37, no. 6, pp. 860-864, 1992.

[32] R. C. Dorf and R. H. Bishop, "Modern control systems," Pearson Prentice Hall, 12th edition, 2011.

Towards end-to-end Continuous Monitoring of Compliance Status Across Multiple Requirements

Danny C. Cheng[1], Jod B. Villamarin[2], Gregory Cu[3], Nathalie Rose Lim-Cheng[4]
College of Computer Studies, De La Salle University
Manila, Philippines

***Abstract*—Monitoring compliance status by an organization has been historically difficult due to the growing number of compliance requirements being imposed by various standards, frameworks, and regulatory requirements. Existing practices by organizations even with the assistance of security tools and appliances is mostly manual in nature as there is still a need for a human expert to interpret and map the reports generated by various solutions to actual requirements as stated in various compliance documents. As the number of requirements increases, this process is becoming either too costly or impractical to manage by the organization. Aside from the numerous requirements, multiple of these documents actually overlap in terms of domains and actual requirements. However, since current tools do not directly map and highlight overlaps as well as generate detailed gap reports, an organization would perform compliance activities redundantly across multiple requirements thereby increasing cost as well. In this paper, we present an approach that attempts to provide an end-to-end solution from compliance document requirements to actual verification and validation of implementation for audit purposes with the intention of automating compliance status monitoring as well as providing the ability to have continuous compliance monitoring as well as reducing the redundant efforts that an organization embarks on for multiple compliance requirements. This research thru enhancing existing security ontologies to model compliance documents and applying information extraction practices would allow for overlapping requirements to be identified and gaps to be clearly explained to the organization. Thru the use of secure systems development lifecycle, and heuristics the research also provide a mechanism to automate the technical validation of compliance statuses thereby allowing for continuous monitoring as well as mapping to the enhanced ontology to allow reusability via conceptual mapping of multiple standards and requirements. Practices such as unit testing and continuous integration from secure systems development life cycle are incorporated to allow for flexibility of the automation process while at the same time using it to support the mapping between compliance requirements.**

***Keywords*—Compliance management, continuous compliance monitoring; ontology mapping; natural language processing; secure systems development lifecycle**

I. Introduction

The need to conduct compliance activities within an organization has never been more apparent that it is in recent times. Regulations such as the General Data Protection Regulation or GDPR which aims to protect the privacy rights of individuals make it a requirement for an organization to look into and implement compliance efforts. However, even by just considering the compliance requirements for data privacy alone, an organization would have to consider all the various versions in different countries where the law has a counterpart and the organization is dealing with data subjects from those countries. The number of regulations, standards, frameworks, architectures, and practices that an organization is required to or would benefit from by complying is overwhelming and is continuously increasing in number and complexity as technology and the environment changes over time.

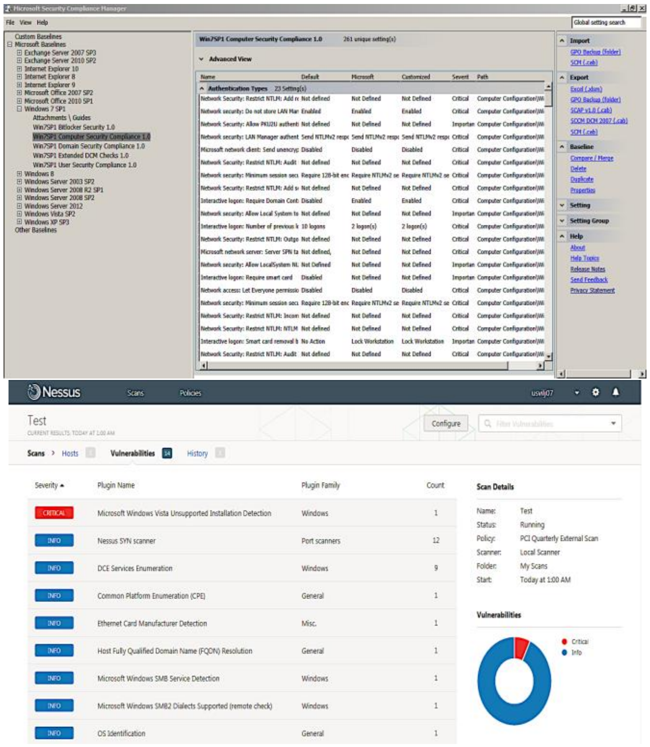
Although they are increasing, it can also be noticed that multiple requirements can also have a large number of overlaps or commonalities that are shared among various regulations, standards, and frameworks. Due to the increase, organizations are attempting to improve on their compliance practices by minimizing redundant new organizational units within the organization [1]. Governance Risk and Compliance (GRC) systems such as those from IBM, SAP, Oracle, and even open source versions such as Eramba have been developed to manage compliance monitoring for enterprises. Systems that generate compliance reports based on predefined templates are also being deployed in an effort to manage and improve on compliance activities. However, compliance monitoring or management systems commonly still rely heavily on human or expert intervention in order to generate reports that answer simple management questions like "If we are already compliant with Standard A what else are we missing to comply with Standard B?". Mapping across requirements and determining overlaps are not commonly found in such systems and thus the process of determining gaps or level of compliance across multiple compliance requirements are repeated for each new regulation as well as for every organizational unit that is to be affected or is required to comply.

Several ontology models for information security, compliance, as well as policy concepts already exist. However, these models have been developed by different researchers whereby information security and compliance are viewed as two separate activities. Although these models are conceptually correct and accurate, the concepts in these models do not capture the concepts that can be seen in the actual compliance requirements statements written in the documents in order to identify, determine, and explain the status of compliance efforts and providing understandable responses to inquiries on level of compliance and areas of non-compliance based on compliance requirements statements to senior management. [2][3][4]

This paper presents a framework for continuous compliance monitoring for multiple requirements documents by enhancing and harmonizing existing ontology definitions for security and compliance focusing on the concepts present in compliance documents. Test cases and unit testing frameworks as practiced in secure software development lifecycle are incorporated in order to audit and validate controls implementation in a flexible and customizable manner. A prototype for mapping test cases and scripts to exact words and phrases in compliance documents which serves as "use cases" allow for the ability to generate reports that show specific deficiencies in compliance based on the document requirements. The mapping can also be used as an input for a lexicon specific to the domain needed in the domain ontology defining concepts such as controls and assets as well as be used to improve the ontology mapping and alignment between different standards and compliance requirements. The research takes a different path by using testing frameworks and scripting based on software quality assurance and secure software development lifecycle methodologies and practices as opposed to the use of existing notations such as Business Process Management Notation (BPMN) as a means to model processes and map to compliance documents [5] as there still exists a great majority of organizations that are not using BPMN within the organization.

II. Ease of Use

Software tools or appliances currently available and deployed to do compliance monitoring, management, and reporting can be grouped together into categories like compliance managers, vulnerability scanners, penetration testers, security events managers, and even governance risk and compliance tools as shown in Fig. 1.



Win7SP1 Computer Security Compliance 1.0 261 unique setting(s)

Advanced View

Name	Default	Microsoft	Customized	Severit	Path
Authentication Types 23 Setting(s)					
Network Security: Restrict NTLM: Add re	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Network security: Do not store LAN Man	Enabled	Enabled	Enabled	Critical	Computer Configuration\Wi
Network Security: Allow PKU2U authent	Not defined	Not Defined	Not Defined	Importan	Computer Configuration\Wi
Network security: LAN Manager authent	Send NTLMv2 respc	Send NTLMv2 resp	Send NTLMv2 resp	Critical	Computer Configuration\Wi
Microsoft network client: Send unencryp	Disabled	Disabled	Disabled	Critical	Computer Configuration\Wi
Network Security: Restrict NTLM: Audit	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Network security: Minimum session secu	Require 128-bit enc	Require NTLMv2 se	Require NTLMv2 se	Critical	Computer Configuration\Wi
Network Security: Restrict NTLM: Add s	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Interactive logon: Require Domain Cont	Disabled	Enabled	Enabled	Critical	Computer Configuration\Wi
Network security: Allow Local System to	Not defined	Not Defined	Not Defined	Importan	Computer Configuration\Wi
Interactive logon: Number of previous k	10 logons	2 logon(s)	2 logon(s)	Critical	Computer Configuration\Wi
Network Security: Restrict NTLM: Outgo	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Microsoft network server: Server SPN ta	Not defined,	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Network security: Allow LocalSystem NL	Not Defined	Not Defined	Not Defined	Importan	Computer Configuration\Wi
Interactive logon: Require smart card	Disabled	Not Defined	Not Defined	Importan	Computer Configuration\Wi
Network access: Let Everyone permissio	Disabled	Disabled	Disabled	Critical	Computer Configuration\Wi
Network security: Minimum session secu	Require 128-bit enc	Require NTLMv2 se	Require NTLMv2 se	Critical	Computer Configuration\Wi
Network Security: Restrict NTLM: Incom	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Network Security: Restrict NTLM: NTLM	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi
Interactive logon: Smart card removal b	No Action	Lock Workstation	Lock Workstation	Importan	Computer Configuration\Wi
Network Security: Restrict NTLM: Audit	Not defined	Not Defined	Not Defined	Critical	Computer Configuration\Wi

Test

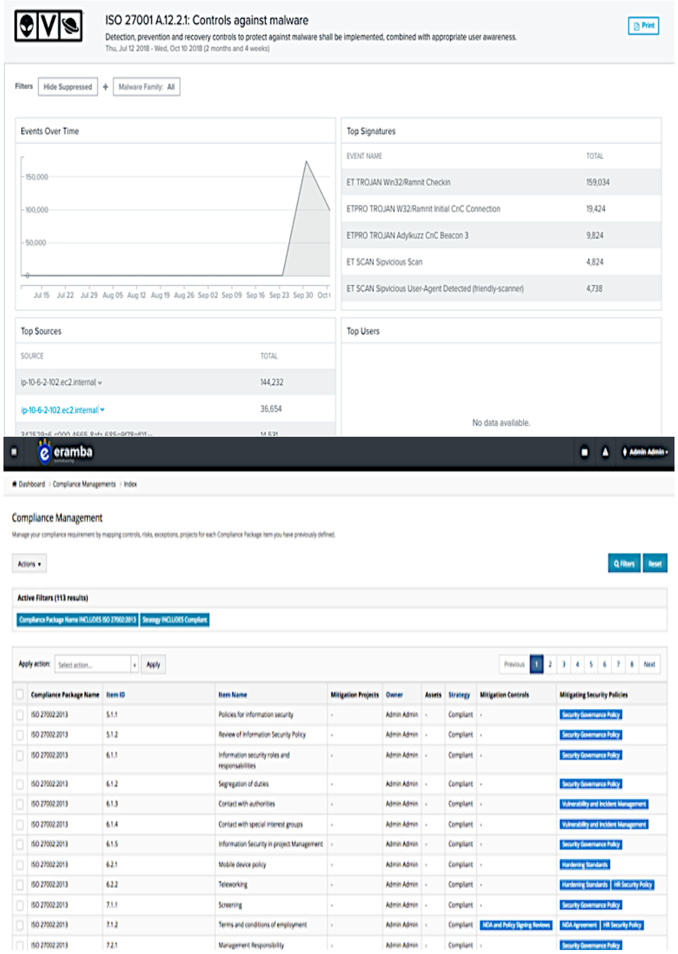
CURRENT RESULTS: TODAY AT 1:00 AM

Severity	Plugin Name	Plugin Family	Count
CRITICAL	Microsoft Windows Vista Unsupported Installation Detection	Windows	1
INFO	Nessus SYN scanner	Port scanners	12
INFO	DCE Services Enumeration	Windows	9
INFO	Common Platform Enumeration (CPE)	General	1
INFO	Ethernet Card Manufacturer Detection	Misc.	1
INFO	Host Fully Qualified Domain Name (FQDN) Resolution	General	1
INFO	Microsoft Windows SMB Service Detection	Windows	1
INFO	Microsoft Windows SMB2 Dialects Supported (remote check)	Windows	1
INFO	OS Identification	General	1

Scan Details

Name:	Test
Status:	Running
Policy:	PCI Quarterly External Scan
Scanner:	Local Scanner
Folder:	My Scans
Start:	Today at 1:00 AM

Vulnerabilities



ISO 27001 A.12.2.1: Controls against malware

Detection, prevention and recovery controls to protect against malware shall be implemented, combined with appropriate user awareness.

Thu, Jul 12 2018 - Wed, Oct 10 2018 (2 months and 4 weeks)

Top Signatures

EVENT NAME	TOTAL
ET TROJAN Win32/Ramnit Checkin	159,034
ETPRO TROJAN W32/Ramnit Initial CnC Connection	19,424
ETPRO TROJAN Adylkuzz CnC Beacon 3	9,824
ET SCAN Sipvicious Scan	4,824
ET SCAN Sipvicious User-Agent Detected (friendly-scanner)	4,738

Top Sources

SOURCE	TOTAL
ip-10-6-2-102.ec2.internal	144,232
ip-10-6-2-102.ec2.internal	36,654

Compliance Management

Manage your compliance requirement by mapping controls, risks, exceptions, projects for each Compliance Package item you have previously defined.

Active Filters (113 results)

Compliance Package Name INCLUDES ISO 27002:2013 | Strategy INCLUDES Compliant

Compliance Package Name	Item ID	Item Name	Mitigation Projects	Owner	Assets	Strategy	Mitigation Controls	Mitigating Security Policies
ISO 27002:2013	5.1.1	Policies for information security	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	5.1.2	Review of Information Security Policy	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	6.1.1	Information security roles and responsabilities	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	6.1.2	Segregation of duties	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	6.1.3	Contact with authorities	-	Admin Admin	-	Compliant	-	Vulnerability and Incident Management
ISO 27002:2013	6.1.4	Contact with special interest groups	-	Admin Admin	-	Compliant	-	Vulnerability and Incident Management
ISO 27002:2013	6.1.5	Information Security in project Management	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	6.2.1	Mobile device policy	-	Admin Admin	-	Compliant	-	Hardening Standards
ISO 27002:2013	6.2.2	Teleworking	-	Admin Admin	-	Compliant	-	Hardening Standards HR Security Policy
ISO 27002:2013	7.1.1	Screening	-	Admin Admin	-	Compliant	-	Security Governance Policy
ISO 27002:2013	7.1.2	Terms and conditions of employment	-	Admin Admin	-	Compliant	NDA and Policy Signing Reviews	NDA Agreement HR Security Policy
ISO 27002:2013	7.2.1	Management Responsibility	-	Admin Admin	-	Compliant	-	Security Governance Policy

Fig. 1. Microsoft Security Compliance Manager [6], Nessus Vulnerability Assessment Tool [7], Alien Vault Unified Security Management[8], Eramba Opensource GRC [9].

Thru a survey of some common and popular tools, this research observed that most of these tools focus heavily on the technical compliance requirements and thus generate compliance reports based on templates that still require human interpretation and translation to actual compliance requirements in order to provide an actual compliance status report. As an example, one tool can generate a report that contains the top sources of malware infections, top signatures detected, and malware events over time as a report for to comply with the statement from ISO27001 stating "Detection, prevention and recovery controls to protect against malware shall be implemented, combined with appropriate user awareness." However, details like prevention and recovery are not clearly seen in the report and user awareness is not considered as well [8]. That being said, assuming that the organization is already compliant with the statement of ISO 27001, when ask the question, is the organization also compliant with Payment Card Industry Data Security Standard (PCIDSS) where the statement is "Ensure that anti-virus programs are capable of detecting, removing, and protecting against all known types of malicious software.", it is also not readily determined by the system if the two requirements are equivalent, or just overlap, or which part of the requirement is still non-compliant.

III. Our Approach

Existing security and compliance ontologies [2][3][4] model domain concepts such as controls in its ideal end state and concepts usually refer to tangible objects that should already exist. However, in considering the compliance documents statements, it can be seen that these documents do not only state the end state for specific requirements, they also state concepts like actions (e.g. "Ensure", "should be implemented") that needs to be performed and conditions (e.g. "actively running") that have to be met or maintained by the specific requirements. As such, enhancements to the existing ontologies were introduced [10] in order to model and capture the statements from the perspective of a "requirement" rather than just a "control". This perspective was taken in order to also facilitate mapping and translation between compliance documents.

TABLE I. COMPLIANCE DOCUMENT STRUCTURE

Document	Structure
CoBIT 5	Stakeholder Drivers -> Stakeholder Needs -> Enterprise Goals -> IT Related Goals -> Enabler Goals -> Process Goal -> Process Practices -> Process Activities
ISO 27002	Security Control Clause/ Domains -> Security Categories -> Controls -> Objective, Implementation Guidance
PCIDSS	Requirements -> Testing Procedure -> Guidance

A. Compliance Documents Conceptual Overlaps

One of the goals of this research is to allow for a more granular understanding and assessment of the compliance level of an organization. In considering the documents to be used in the research, Table I illustrates the inherent document structure of each of the documents that are to be used in this research. To improve on the granularity aspect, the research focuses on the leaf nodes of the structure of the document as the actual requirements are stipulated in this section of the document.

High level alignments and mappings are already available both for older versions of the documents as well as the current versions. However, having the mapping stop at a higher section level of the document loses the needed details to actually determine how similar or different the requirements are to each other as well as the potential gaps among these requirements. Table IV shows the level with which current alignments and mappings are being developed and published. Although such guides provide a good starting point for mapping and determining related requirements, it lacks the level of detail to clearly describe the differences and redundancies. One advantage in looking at standards documents is that each requirement is stated in an enumerated structure such that it is possible to perform comparison without having to locate relevant sections as well as remove unneeded and unwanted text. The same cannot be said on other forms of documents such as the traditional book or news articles as well as corporate governance documents that are more descriptive based rather than being itemized [11].

To model compliance as per the compliance documents, this research took the perspective of a compliance auditor that is checking based on the statements or requirements of the compliance documents. Table II shows the definitions of the concepts from the actual document while Table III shows sample excerpts from these documents. It can be seen from both tables that there are conceptual overlaps both in definition and in the actual requirements as stated in the documents. Given the overlaps, several attempts and efforts have been performed to map and align these documents in order to enforce compliance. However, current efforts are done manually and independent of any systems that can monitor compliance. Compliance and audit practices are also currently performed at an individual standard or requirement basis as there is no definitive granular mapping that can illustrate exact overlaps and gaps among different compliance requirements.

TABLE II. DEFINITION OF THE BASIC CONCEPTS USED AS BASIS FOR EXTRACTION

Document and Concept	Definition
CoBIT 5 - Activity	The main action taken to operate a process. Describe a set of necessary and sufficient action-oriented implementation steps to achieve a Governance ***Practice*** or Management Practice (ISACA 2012)
ISO 27002 - Control	The means of managing risk, including policies, procedures, guidelines, ***practices*** or organizational structures, which can be administrative, technical, management or legal nature. (ISACA 2012)
PCIDSS - Requirement	Compliance validation basis, considered in-place if ***controls*** are implemented or scheduled to be implemented. (Payment Card Industry 2016)

TABLE III. EXCERPT ON THE STANDARDS DOCUMENTS REFERRING TO MALWARE PROTECTION

Document	Excerpt
CoBIT 5 - Activity	Implement and maintain preventive, detective and corrective measures in place (especially up-to-date security patches and virus control) across the enterprise to protect information systems and technology from malware (e.g., viruses, worms, spyware, spam).
ISO 27002 - Control	Detection, prevention and recovery controls to protect against malware should be implemented, combined with appropriate user awareness.
PCIDSS - Requirement	Deploy anti-virus software on all systems commonly affected by malicious software (particularly personal computers and servers).

Existing works such as [11] map compliance document concepts at a section level (see Table IV) which leads to loss of detail as well as the inability to develop a system that can automate compliance status reporting with respect to actual compliance requirement statements. Our research looks at the lower level (see Table V) in modeling the concepts for the compliance ontology in an attempt to be implementable by a system.

TABLE IV. RECENTLY PUBLISHED MAPPING DOCUMENTS THAT SHOW LEVELS OF MAPPING

Document	Structure
CoBIT 5	Stakeholder Drivers -> Stakeholder Needs -> Enterprise Goals -> IT Related Goals -> Enabler Goals -> Process Goal -> ***Process Practices*** -> Process Activities
ISO 27002	Security Control Clause/ Domains -> Security Categories -> ***Controls*** -> Objective, Implementation Guidance
PCIDSS	***Requirements*** -> Testing Procedure -> Guidance

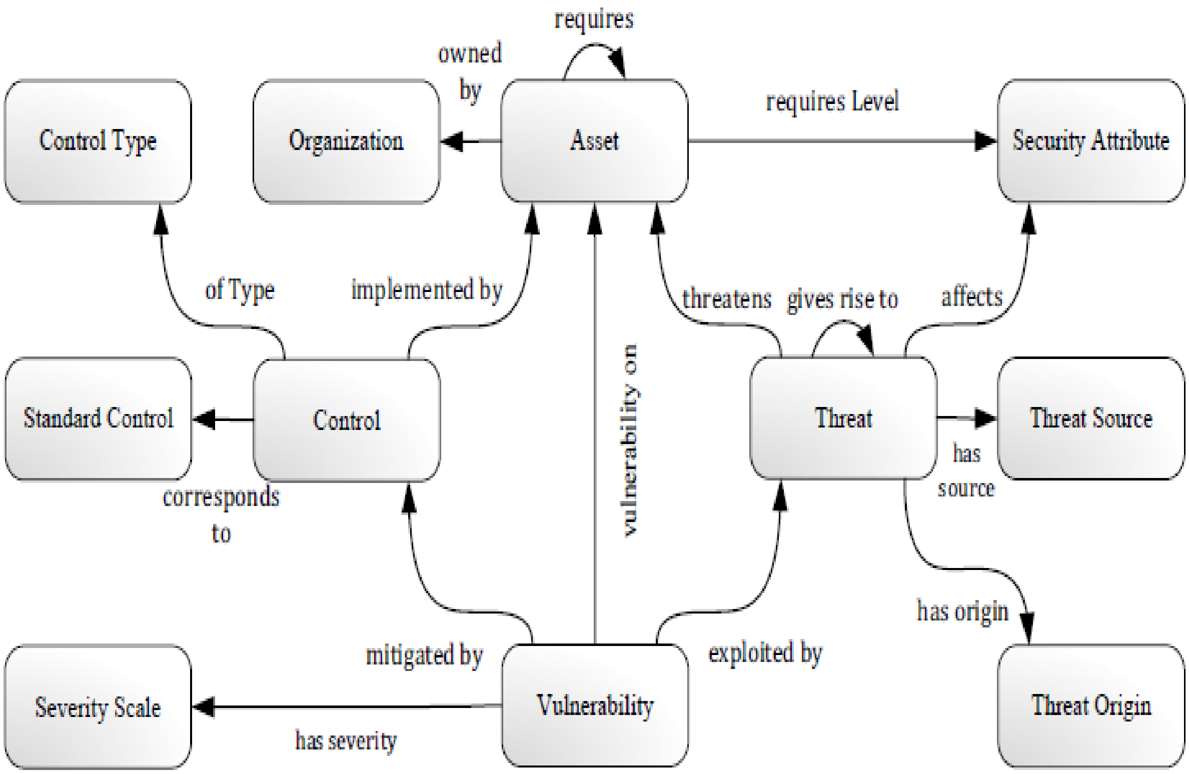
TABLE V. COMPLIANCE DOCUMENT LEVEL USED FOR MODELING

Document	Structure
CoBIT 5	Stakeholder Drivers -> Stakeholder Needs -> Enterprise Goals -> IT Related Goals -> Enabler Goals -> Process Goal -> Process Practices -> ***Process Activities***
ISO 27002	Security Control Clause/ Domains -> Security Categories -> Controls -> ***Objective, Implementation Guidance***
PCIDSS	***Requirements*** -> Testing Procedure -> Guidance

B. Compliance Ontology Enhancement

Multiple efforts in defining ontologies that can be used in security and compliance activities have been conducted and defined. As can be seen in Fig. 2, high-level conceptual modelling of information security and compliance have been developed and defined [12] [13] that shows the corresponding relationships of the different concepts involved. However, these models are not linked to the compliance documents and model mostly based on technical aspects or general concepts of information security or compliance. Several concepts defined in these researches are not readily visible or available within the statements of the compliance documents. In order to link such models to the actual statements in the documents, this research focused on the concept of Assets and Controls and put aside concepts such as Threats and Vulnerabilities as although these are valid information security concepts, such concepts would normally not be found in the statements of the compliance documents. Fig. 3 shows the enhancements performed on the information security ontology.

Upon evaluating the structure of the requirements statements in the compliance documents, a Control does not stand alone in the document as the statement included Actions to be applied to Controls. Also, although a Control is implemented by an Asset, the same Control can also be applied to other objects which are also part of the Assets of the organization (e.g. Install and regularly update anti-virus software on machines containing sensitive information). Conditions (or qualifiers) are also introduced into the ontology for the concepts of Controls and Assets as the documents contain phrases such as "regularly update", "periodic evaluation", "commonly affected", and "strong encryption". A State is also added to an Asset as the documents require checking for phrases like "actively running assets". The Action can be composed of multiple actions as some documents would provide detailed guidelines on things to do, while others would be more general on their statements. The scope of an action would cover statements that contain specific ranges or domains of applicability such as perimeter or date ranges.



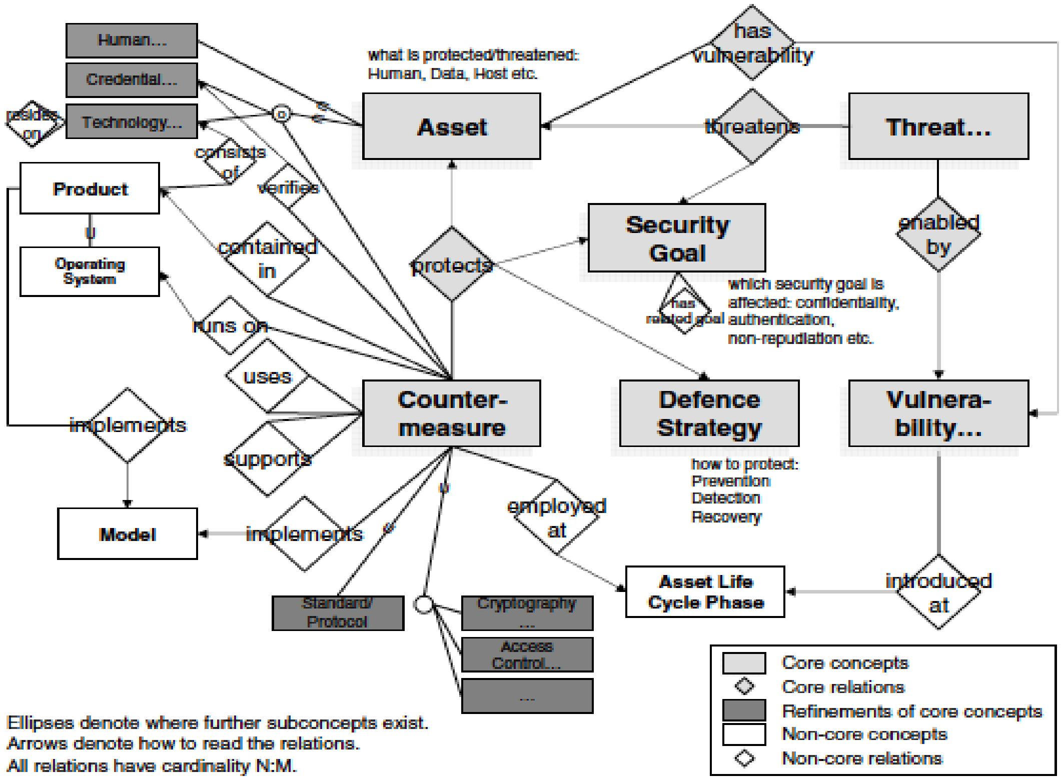


Fig. 2. Existing Informaiton Security Ontology [12][13].

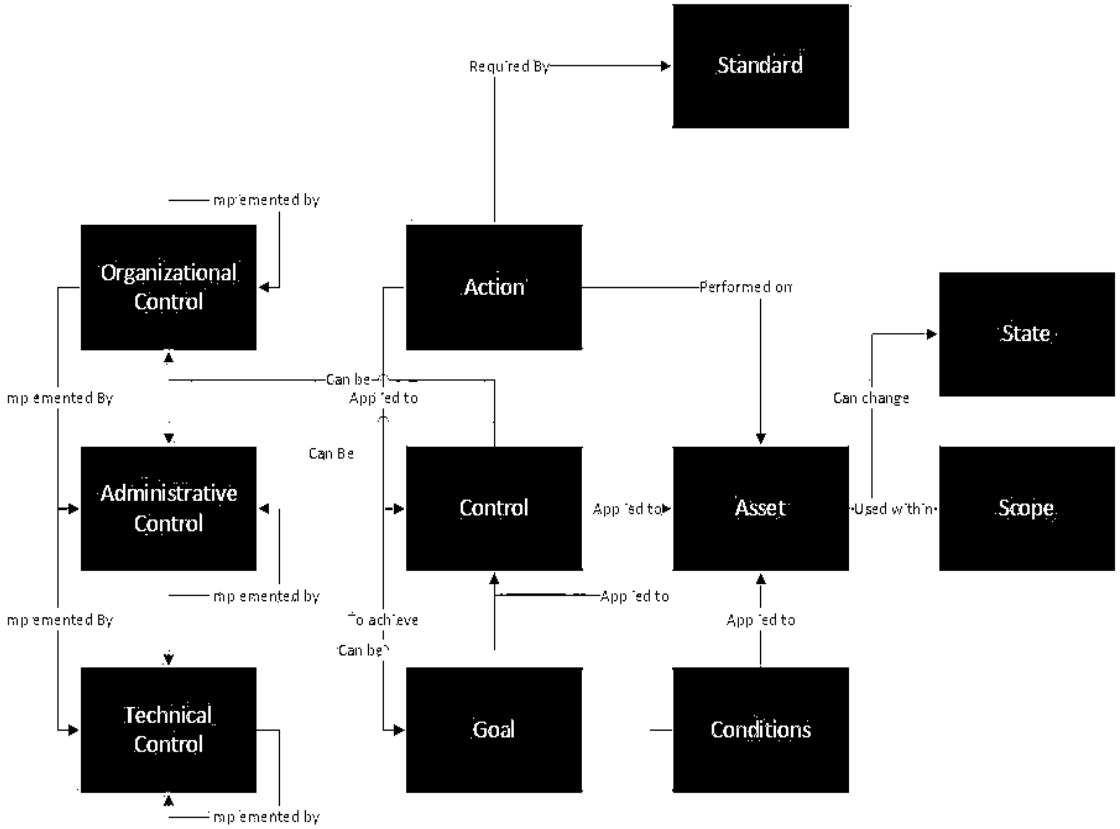


Fig. 3. Enhancements to Compliance Ontology.

In Fig. 2, the concept of the hierarchy of controls in introduced to be part of the enhanced ontology in order to tackle the complete compliance requirements rather than focusing solely on the technical aspects of controls and compliance. An example non-technical requirement would be the development of an Information Security Management Program as required by ISO 27001:2013. This requirement cannot be modelled or represented as a technical control but is an organizational requirement that is supported or implemented by administrative controls such as review process and procedures which in turn is implemented through the various technical controls deployed within an organization. The motivation behind modeling organizational and administrative controls is that although they are non-technical or manual in nature, scripts can be developed through the use of rules to automatically verify or validate if the controls are being implemented and enforced. An example case would be the administrative control requirement of an annual review process for the documents. A script can be developed with a rule to check the last date of review or revision to the document that is stored in a repository which can be used to automatically verify and validate if the requirement for annual review is being complied with or not.

C. Information Extraction in Populating the Ontology

As overlaps clearly exist among and between compliance documents, the need to reduce redundant or repetitive efforts in compliance enforcement and audit becomes paramount as the number of compliance requirements increase at an alarming rate over time [14]. In order to achieve this, automation of mapping of the compliance document to the enhanced compliance ontology is needed. To populate the enhanced ontology, Natural Language Processing tools and techniques such as information extraction approaches have been applied to construct structured information from unstructured data sources such as the compliance documents. Standards or Compliance documents state requirements as a series of imperative statements. A semantic relationship between two concepts is expressed by a verb in natural language texts [15], hence it is first necessary to identify where the verb is in the requirement statement. From there, the noun phrases and other clauses (e.g., prepositional phrases) can be processed to determine the relationships of the verb with the other entities in the sentence. For this process, Stanford CoreNLP toolkit [16] was employed. Each statement in the compliance document were pre-processed to form complete sentences and the output was given to the CoreNLP toolkit to parse and extract based on Parts of Speech (POS) generating a dependency tree [10]. The dependency tree serves as the basis for identifying what items will be used to populate the ontology (e.g. actions are populated based on verbs identified).

Upon studying the resulting annotations from Stanford CoreNLP, it was apparent that there are patterns to the type of concept to be extracted in relation to its part of speech and/or with its semantic dependency. The following lists the general patterns that can be applied to extract data and populate the ontology.

1) The ROOT is extracted as ACTION. Depending on the POS of the ROOT, the extracted data may need to be lemmatized.

a) If there is an immediate child node of this ROOT that has POS of CC and SD of CC, we then look for child nodes in the same level that has SD of CONJ to also serve as ACTION. Each ACTION extracted is stored as a separate entry in the populated ontology. Later, once the asset is determined (as stated in the following patterns), the associated ASSET and its constituents (e.g., QUANTITY and ASSET_TYPE, if applicable) are copied.

2) The first child node, where the POS is NNS and the SD is DOBJ or NSUBJPASS, is extracted as ASSET. This child node usually appears as the immediate child node or as a second-degree child node.

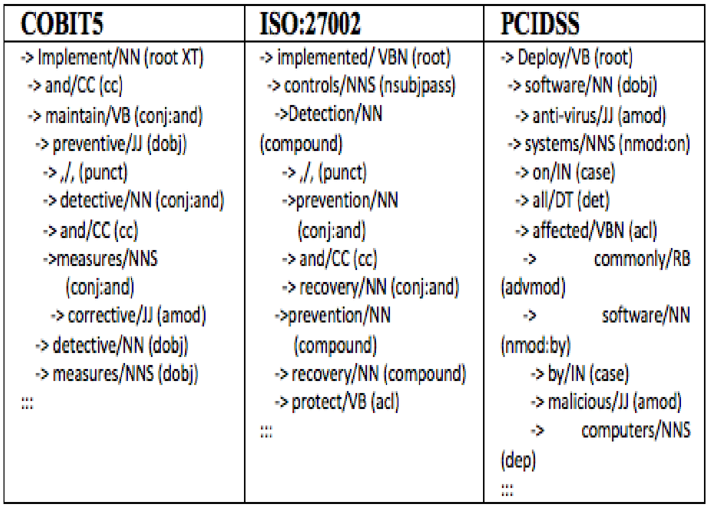
a) If there is an immediate child node of this ASSET that has POS of CC and SD of CC, we then identify other child nodes in the same level, with POS NNS and SD as CONJ, also as ASSET.

3) The immediate child node of the ASSET, where the POS is JJ and SD is AMOD, is the ASSET_TYPE. Similarly, the immediate child node of the ASSET, where the POS is NN and SD is COMPOUND, is extracted as ASSET_TYPE.

4) The immediate child node of the ASSET, where the POS is DT and SD is DET, is the QUANTITY.

5) The first child node of the ROOT that is a modifier of the DOBJ, e.g., NMOD that is associated with a marker like on or across, this subtree is extracted as SCOPE.

6) The first child node where the SD is ACL and has immediate child node marker to, the entire subtree is extracted as GOAL.



COBIT5	ISO:27002	PCIDSS
-> Implement/NN (root XT) -> and/CC (cc) -> maintain/VB (conj:and) -> preventive/JJ (dobj) -> ,/, (punct) -> detective/NN (conj:and) -> and/CC (cc) ->measures/NNS (conj:and) -> corrective/JJ (amod) -> detective/NN (dobj) -> measures/NNS (dobj) :::	-> implemented/ VBN (root) -> controls/NNS (nsubjpass) ->Detection/NN (compound) -> ,/, (punct) ->prevention/NN (conj:and) -> and/CC (cc) -> recovery/NN (conj:and) ->prevention/NN (compound) -> recovery/NN (compound) -> protect/VB (acl) :::	-> Deploy/VB (root) -> software/NN (dobj) -> anti-virus/JJ (amod) -> systems/NNS (nmod:on) -> on/IN (case) -> all/DT (det) -> affected/VBN (acl) -> commonly/RB (advmod) -> software/NN (nmod:by) -> by/IN (case) -> malicious/JJ (amod) -> computers/NNS (dep) :::

Fig. 4. Sample Results on using Stanford Corenlp with POS Extraction on Compliance Documents.

It should be noted that the Stanford CoreNLP tool sometimes produces erroneous tags. This phenomenon where the root identified is correct but was given a wrong POS tag appeared for quite a few samples, so far all from Control Objectives for Information and Related Technologies (COBIT)5. Similar to that in Fig. 4, the tag for "implement" was that of noun, instead of verb, even when the text is followed by a conjunction to another verb. As the assumption is that statements are written in correct English grammar, we can resolve this by comparing (and aligning) the parts of speech of both conjuncts. One possible cause for this error can be attributed to the actual structure of the document and a way to resolve this is to perform a pre-processing step to first split statements to individual goals and individual assets to, not only provide a more accurate result, but also to granularly store and perform better inferences on data later. That is, a pre-processor may first split the said COBIT5 activity into the statements in Table VI.

TABLE VI. RESTRUCTURED COBIT5 DOCUMENT FORUSE IN STANFORD CORENLP POS EXTRACTION

Implement preventive measures in place … across the enterprise to protect information systems and technology from malware.
Implement detective measures in place … across the enterprise to protect information systems and technology from malware.
Implement corrective measures in place … across the enterprise to protect information systems and technology from malware.
Maintain preventive measures in place … across the enterprise to protect information systems and technology from malware.
Maintain detective measures in place … across the enterprise to protect information systems and technology from malware.
Maintain corrective measures in place … across the enterprise to protect information systems and technology from malware.

```
-> data/NNS (root)
  -> Record/NNP (compound)
  -> events/NNS (nmod:on)
    -> on/IN (case)
    -> risk/NN (compound)
    -> caused/VBN (acl:relcl)
      -> that/WDT (nsubj)
      -> have/VBP (aux)
      -> or/CC (cc)
      -> cause/VB (conj:or)
        -> that/WDT (nsubj)
        -> may/MD (aux)
        -> impacts/NNS (dobj)
        -> IT/PRP (nmod:to)
          -> to/TO (case)
      -> enablement/NN (dobj)
        -> benefit/value/NN (compound)
```

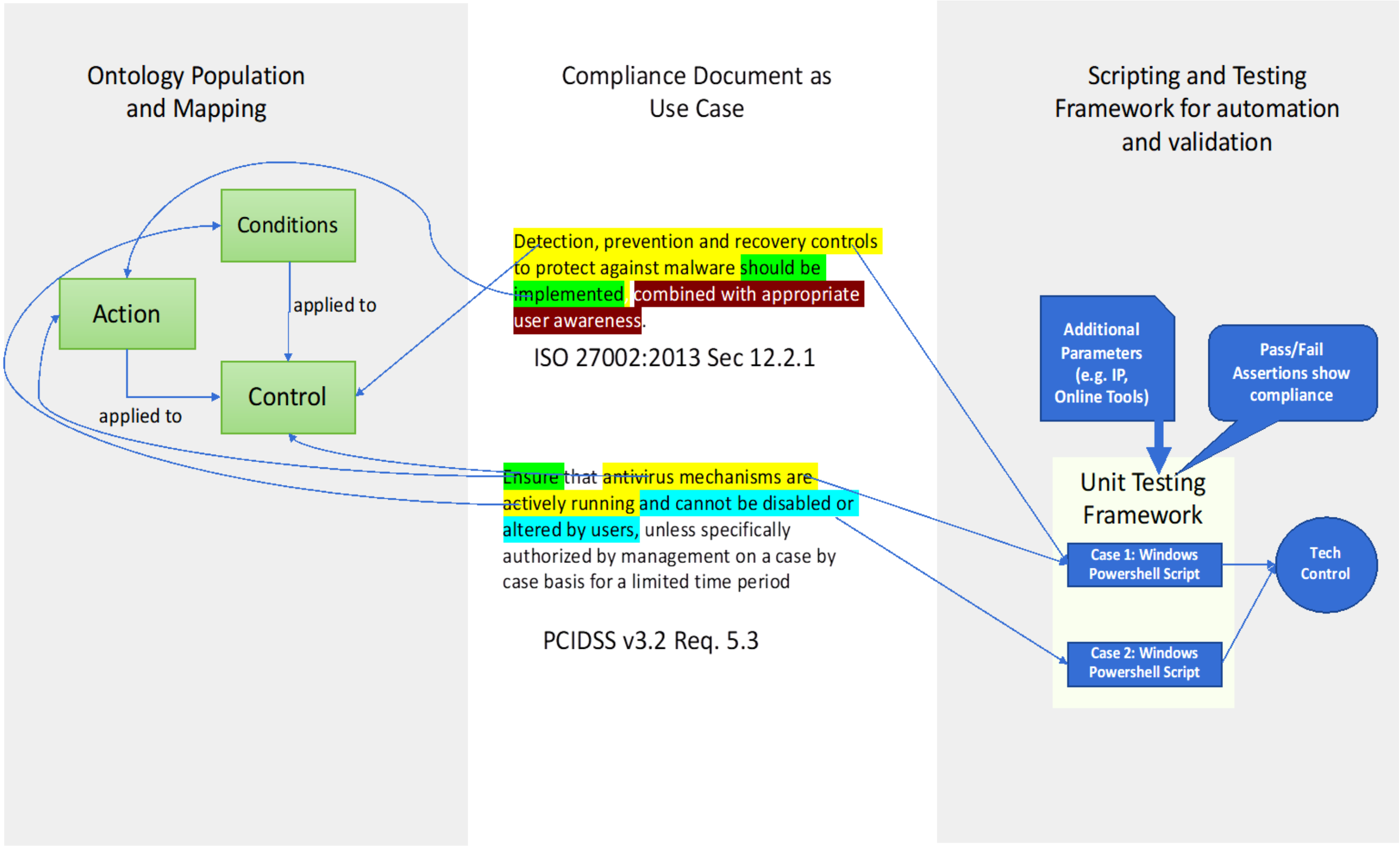
Fig. 5. Sample Result Showing Incorrect ROOT Node.

However, incorrect tags can be more of a problem should the root identified be incorrect, as such affecting the dependency tree as well. One such example is another activity in COBIT stating: "Record data on risk events that have caused or may cause impacts to IT benefit/value enablement, IT programme and project delivery, and/or IT operations and service delivery." Fig. 5 shows an excerpt of the resulting annotation by Stanford CoreNLP. A possible resolution to this phenomenon is currently still under research.

D. *Linking Compliance Documents to Verification Automation Scripts*

In order to complete the link between compliance requirements and implementation verification as well as provide the ability to support continuous compliance monitoring and process audit [17], there is a need to link the populated ontology and its related compliance document to actual technical verification tools such as audit scripts in order to provide real-time compliance status feedback (see Fig. 6). There is a need to map the audit scripts to the compliance documents in order automate compliance monitoring. Existing tools that monitor compliance map to industry practices which makes them unable to directly show areas of compliance and deficiencies with respect to compliance requirements (e.g. "unsupported installation: Nessus Report" is not linked or mapped to "establishing a formal policy prohibiting the use of unauthorized software: ISO 27002:2013 section 12.2.1a" and even if its mapped, the requirement of "establishing a formal policy" cannot be seen in the initial report as the policy itself if defined is located in a different system (document management system) that is not usually part of the compliance monitoring tools.

Scripting through the use of Windows Powershell for the proof-of-concept (see Fig. 7 and Fig. 8) was implemented rather than the use of solutions such as BPMN [5] as there is a varied set of tools and controls that need to be monitored for compliance and not all tools would support BPMN. The requirement is for the script to be as atomic as possible in order for it to be reusable (e.g. 1 script for checking antivirus deployments with input parameters such as list of IPs that needs have the antivirus deployed as stated in the compliance document). In doing so, having atomic scripts and mapping it to compliance documents can also aid in the mapping and translation of requirements through the enhanced ontology by providing a common vocabulary such as common controls that was previously unavailable. Heuristics can also be incorporated to improve the mapping and translation between different compliance requirements through the enhanced ontology. Scripts mapped to similar sections or phrases within a compliance document with similar customization parameters can be used to identify overlapping requirements in multiple compliance documents as well as validation of the mapping of the ontology between documents. Relationships of controls and compliance documents such as subsumption of requirements can also be inferred by analyzing the similarity in scripts and customization parameters.



Ontology Population and Mapping

Compliance Document as Use Case

Scripting and Testing Framework for automation and validation

Detection, prevention and recovery controls to protect against malware should be implemented combined with appropriate user awareness.

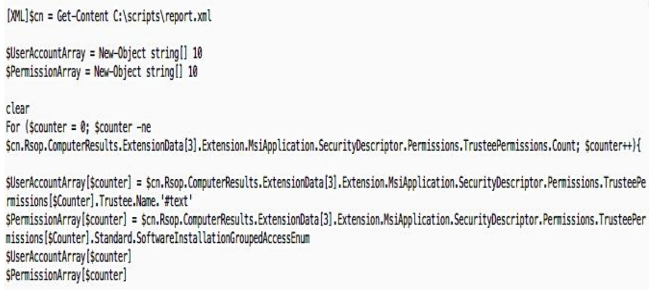
ISO 27002:2013 Sec 12.2.1

Ensure that antivirus mechanisms are actively running and cannot be disabled or altered by users, unless specifically authorized by management on a case by case basis for a limited time period

PCIDSS v3.2 Req. 5.3

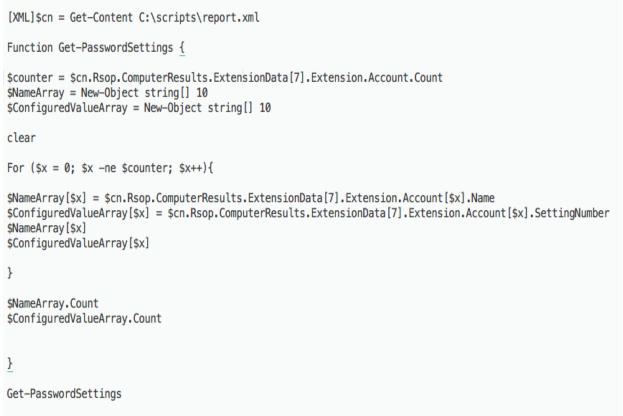
Additional Parameters (e.g. IP, Online Tools)
Pass/Fail Assertions show compliance
Unit Testing Framework
Case 1: Windows Powershell Script
Case 2: Windows Powershell Script
Tech Control

Fig. 6. Mapping and Population of Compliance Ontology based on Compliance Document and Linking it to a Testing Framework for Automation and Validation.



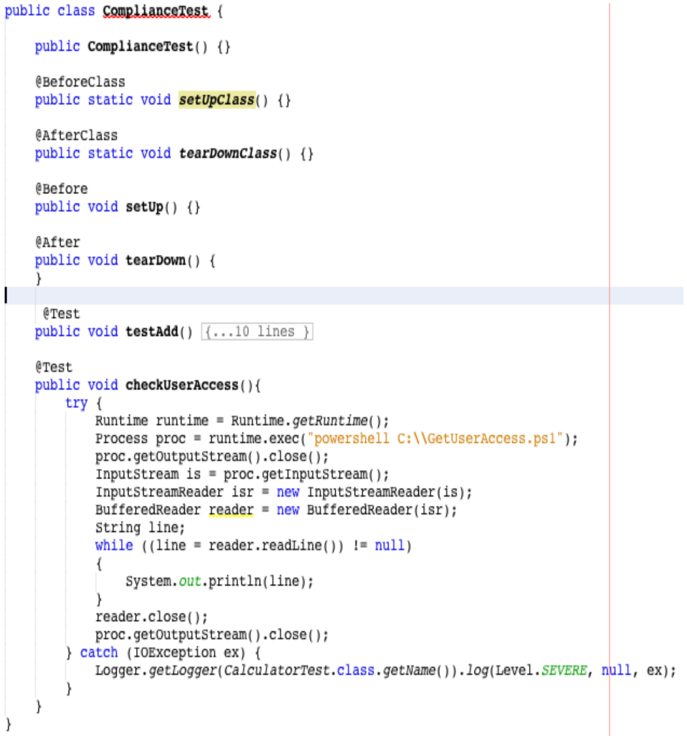
```
[XML]$cn = Get-Content C:\scripts\report.xml

$UserAccountArray = New-Object string[] 10
$PermissionArray = New-Object string[] 10

clear
For ($counter = 0; $counter -ne
$cn.Rsop.ComputerResults.ExtensionData[3].Extension.MsiApplication.SecurityDescriptor.Permissions.TrusteePermissions.Count; $counter++){

$UserAccountArray[$counter] = $cn.Rsop.ComputerResults.ExtensionData[3].Extension.MsiApplication.SecurityDescriptor.Permissions.TrusteePe
rmissions[$Counter].Trustee.Name.'#text'
$PermissionArray[$counter] = $cn.Rsop.ComputerResults.ExtensionData[3].Extension.MsiApplication.SecurityDescriptor.Permissions.TrusteePer
missions[$Counter].Standard.SoftwareInstallationGroupedAccessEnum
$UserAccountArray[$counter]
$PermissionArray[$counter]
```

Fig. 7. Sample Windows Powershell Script for Getting the Access Rights of the user.



```
[XML]$cn = Get-Content C:\scripts\report.xml

Function Get-PasswordSettings {

$counter = $cn.Rsop.ComputerResults.ExtensionData[7].Extension.Account.Count
$NameArray = New-Object string[] 10
$ConfiguredValueArray = New-Object string[] 10

clear

For ($x = 0; $x -ne $counter; $x++){

$NameArray[$x] = $cn.Rsop.ComputerResults.ExtensionData[7].Extension.Account[$x].Name
$ConfiguredValueArray[$x] = $cn.Rsop.ComputerResults.ExtensionData[7].Extension.Account[$x].SettingNumber
$NameArray[$x]
$ConfiguredValueArray[$x]

}

$NameArray.Count
$ConfiguredValueArray.Count


}

Get-PasswordSettings
```

Fig. 8. Sample Windows Powershell Script for Checking Password Settings.



```
public class ComplianceTest {

    public ComplianceTest() {}

    @BeforeClass
    public static void setUpClass() {}

    @AfterClass
    public static void tearDownClass() {}

    @Before
    public void setUp() {}

    @After
    public void tearDown() {
    }

     @Test
    public void testAdd() {...10 lines }

    @Test
    public void checkUserAccess(){
        try {
            Runtime runtime = Runtime.getRuntime();
            Process proc = runtime.exec("powershell C:\\GetUserAccess.ps1");
            proc.getOutputStream().close();
            InputStream is = proc.getInputStream();
            InputStreamReader isr = new InputStreamReader(is);
            BufferedReader reader = new BufferedReader(isr);
            String line;
            while ((line = reader.readLine()) != null)
            {
                System.out.println(line);
            }
            reader.close();
            proc.getOutputStream().close();
        } catch (IOException ex) {
            Logger.getLogger(CalculatorTest.class.getName()).log(Level.SEVERE, null, ex);
        }
    }
}
```

Fig. 9. Sample Integration of Windows Powershell Compliance Test Scripts to a unit Testing Framework to Model Compliance Requirements as Test Cases[18].

In order to fully support compliance requirements statements, the audit scripts are formalized through the use of test cases in testing frameworks used in software quality assurance to allow for customization and configuration based on audit practices (see Fig. 9) [11]. The introduction of the concept of test cases allows for a formal mapping as the compliance requirements can now be modelled also as use cases similar to what is being used in secure software development lifecycles (see Fig. 6). Complexities such as scope of control and non-technical controls can also be supported by audit scripts with the assumption that additional input may be given, or artifacts can be found in digital storage. In the case of PCIDSS v3.2 the statement "Deploy antivirus software on all systems commonly affected by malicious software (particularly personal computers and servers)" contains a scope of "all systems commonly affected". In a manual audit, the audit team would ask for network diagrams and segments to determine scope, for the audit scripts, the scope can be given as a set of IP addresses that meet the criteria. For the case of document requirements such as a formal policy, audit scripts can also be developed to check document repositories for the existence and updates to policies of the organization.

Aside from checking the existence of completed compliance implementations, the research also aims to support partial compliance and evidence mapping by linking compliance requirements to tickets within a project management tool so as to show the status of an activity or project for compliance for monitoring purposes. If the activity or project has been completed, the link can also serve as a means to derive documentary evidences for compliance report generation as needed by compliance audits.

IV. Results and Conclusions

Validation and testing for the research is currently done by parts, namely, the ontology population through information extraction, the integration of the enhanced ontology to an existing GRC system, and the validation of audit scripts with respect to compliance statements. In ontology population, the rules were manually evaluated against the resulting Parts of Speech (POS) tagging of CoreNLP and the number of terms matching the defined rules to determine the verb or action to be performed. For the integration with an existing GRC system, the results were evaluated based on the ability of the resulting extraction to conform with the process of data population or importing of data to the GRC system. Finally, in terms of the validation of audit scripts with respect to compliance statements, configurable scripts were developed that allow parameters to be given to describe the actual compliance statement requirements. These scripts and their corresponding parameters are then mapped to the parts or phrases within the compliance statement in order to determine which requirement is actually being verified by the script.

Current results of the information extraction and ontology population when tested on 356 compliance requirements statements using CoreNLP showed a 69.38% (247 statements) accuracy in action/verb identification. After additional pre-processing and testing on other similar compliance documents, the result of proper verb/action identification is currently at an average of 79% with a range of 70% to 91% as detailed in Table VII and VIII. The average result of each compliance document is used to get the average of the accuracy level across different compliance documents that were used in this research. However, the test only refers to the ability to determine the proper verb or action needed and does not yet consider the identification, extraction, and population of the other concepts as needed in the ontology. It also showed that there is potential misidentification of the action as the word can have multiple meanings [10]. Challenges currently exist in identifying the remaining concepts of the enhanced ontology from compliance documents for ontology population due to the lack of existing taxonomy or vocabularies in the domain. The ontology population is also currently mapped to an opensource GRC system (Eramba) [10] and integration to a project management system has been conceptualized through the use of hyperlinks to project tickets in order to support monitoring of partial compliance implementation.

Audit scripts have also been implemented in an atomic and customizable form in order to support the varied requirements of compliance. Scripts developed include checking antivirus deployment status, checking installed software, checking hardware inventory, and network scanning. These are initially developed as PCIDSS was the compliance document used for the basis of determining what scripts to develop. For example, the network scanning script is to be used to determine compliance to PCIDSS requirement 11.1.1 "Maintain an inventory of authorized wireless access points including a documented business justification." The same script can also be used to determine compliance for ISO 27002:2013 Section 13.1.1 Network Controls (f)(g) stating "systems on the network should be authenticated" and "systems connection to the network should be restricted" Current implementation is limited to controls that can be validated through scripting in a Windows Powershell environment.

TABLE VII. Definition of the basic Concepts used as basis for Extraction

Documents	**Requirements or Domains or Sections**	**No. of Sentences**	**No. of Sentences w/ acceptable processing by CoreNLP**	**%**
CSC CIS Requirements	20	272	202	74%
ISO 27001	14	111	87	78%
ISO 9001 2015	7	126	115	91%
PCI V3	13	297	229	77%
PCI DSS 3.2	12	357	249	70%
NIST 80053	24	553	453	82%
TOTAL	**90**	**1716**	**1335**	**79%**

TABLE VIII. DEFINITION OF THE BASIC CONCEPTS USED AS BASIS FOR EXTRACTION

Documents	**Requirements or Domains or Sections**	**No. of Sentences**	**No. of Sentences w/ acceptable processing by CoreNLP**	**%**
CSC CIS Requirements	1	10	8	80%
	2	9	9	89%
	3	19	15	84%
	4	19	14	81%
	5	13	8	77%
	6	11	9	78%
	7	17	15	80%
	8	11	7	78%
	9	8	7	79%
	10	9	7	79%
	11	12	11	80%
	12	20	14	78%
	13	16	11	78%
	14	10	7	77%
	15	16	9	76%
	16	23	16	75%
	17	10	5	74%
	18	16	13	74%
	19	10	9	75%
	20	13	8	74%
TOTAL	20	272	202	74%
ISO 27001	5	2	2	100%
	6	7	6	89%
	7	6	3	73%
	8	10	10	84%
	9	13	9	79%
	10	2	2	80%
	11	15	13	82%
	12	14	10	80%
	13	7	4	78%
	14	13	9	76%
	15	4	3	76%
	16	7	6	77%
	17	4	3	77%
	18	7	7	78%
TOTAL	14	111	87	78%
ISO 9001 2015	4	13	11	85%
	5	6	6	89%
	6	8	8	93%
	7	28	24	89%
	8	49	46	91%
	9	15	13	91%
	10	7	7	91%
TOTAL	7	126	115	91%
PCI V3	1	25	20	80%
	2	18	15	81%
	3	28	24	83%
	4	7	3	79%
	5	6	4	79%
	6	36	22	73%
	7	11	8	73%
	8	31	25	75%
	9	41	35	77%
	10	26	17	76%
	11	21	13	74%
	12	42	39	77%
	A	5	4	77%
TOTAL	13	297	229	77%
PCI DSS 3.2	1	24	15	63%
	2	19	16	72%
	3	37	28	74%
	4	6	5	74%
	5	8	5	73%
	6	37	17	66%
	7	11	6	65%
	8	37	26	66%
	9	41	24	65%
	10	49	47	70%
	11	30	21	70%
	12	58	39	70%
TOTAL	12	357	249	70%
NIST 80053	1	43	35	81%
	2	22	19	83%
	3	14	13	85%
	4	22	15	81%
	5	28	24	82%
	6	25	19	81%
	7	29	27	83%
	8	9	6	82%
	9	11	9	82%
	10	37	33	83%
	11	13	13	84%
	12	28	24	84%
	13	19	15	84%
	14	9	6	83%
	15	29	22	83%
	16	23	21	83%
	17	40	37	84%
	18	31	21	83%
	19	21	15	83%
	20	34	27	82%
	21	18	13	82%
	22	4	4	82%
	23	29	23	82%
	24	15	12	82%
TOTAL	24	553	453	82%

Although the continuous compliance monitoring framework has been identified and defined, the research still needs to validate if the use of heuristics from the perspective of audit scripts can help build the taxonomy or vocabulary needed to improve the population of the enhanced compliance ontology which in turn will improve the mapping and translation of compliance requirements with the eventual goal of reducing compliance efforts and activities in an ever growing complexity of compliance requirements.

ACKNOWLEDGMENT

This research is partly funded by the University Research Coordination Office (URCO) of De La Salle University Manila and was conducted with the assistance of research assistant Julie Ann Salido.

REFERENCES

[1] Falcione A., McKillop, J. 2016. “PwC State of Compliance Study 2016 Laying a strategic foundation for strong compliance risk management” in https://www.pwc.com/us/en/risk-assurance/state-of-compliance-study/assets/state-of-compliance-study-2016.pdf

[2] N. S. Abdullah, M. Indulska, and S. Sadiq. 2016. “Compliance management ontology --- a shared conceptualization for research and practice in compliance management.” Information Systems Frontiers 18, 5 (October 2016), 995-1020. DOI: https://doi.org/10.1007/s10796-016-9631-4)

[3] Fenz, S., Ekelhart, A.: 2009. “Formalizing information security knowledge.” in ASIACCS 2009: Proceedings of the 2009 ACM symposium on Information, computer and communications security. ACM, New York

[4] Schmidt, R., Bartsch, C., Oberhauser, R., 2007 “Ontology-based representation of compliance requirements for service processes” SBPM 2007 Semantic Business Process and Product Lifecycle Management. http://ceur-ws.org/Vol-251/paper4.pdf

[5] Sunkle S., Kholkar D., and Kulkarni V. 2016 “Toward Better Mapping between Regulations and Operational Details of Enterprises Using Vocabularies and Semantic Similarity” Complex Systems Informatics and Modeling Quarterly CSIMQ, Issue 5, December 2015 / January 2016, Pages 39-60

[6] Microsoft Security Compliance Management (retrieved October 2018) https://www.microsoft.com/en-us/download/details.aspx?id=53353

[7] Nessus Vulnerability Assessment Tool (retrieved October 2018) https://www.tenable.com/products/nessus/nessus-professional

[8] Alienvault Unified Security Management (retrieved October 2018) https://www.alienvault.com/products

[9] Eramba Opensource Governance Risk and Compliance System (retrieved October 2018) http://www.eramba.org/

[10] D. C. Cheng and N. R. Lim-Cheng, "An ontology based framework to support multi-standard compliance for an enterprise," 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, 2017, pp. 1-6.

[11] D. Lacey, “A Practical Guide To The Payment Card Industry Data Security Standard (PCI DSS)” ISACA 2015 ISBN 978-1-60420-586-2

[12] F., S., Ekelhart, A.: 2009. “Formalizing information security knowledge.” in ASIACCS 2009: Proceedings of the 2009 ACM symposium on Information, computer and communications security. ACM, New York

[13] H., Almut & S., Nahid & D., Claudiu. (2007). An Ontology of Information Security. IJISP. 1. 1-23. 10.4018/jisp.2007100101.

[14] J. Verver (2017). Top 8 Better Practices In Compliance Management. Retrieved from http://www.acl.com/pdfs/ebook-top-8-better-practices-in-compliance-management.pdf

[15] Mercier-Laurent, E. and D. Leake, D. 2008. “Intelligent Information Processing IV” In Proceedings of the 5th IFIP International Conference on Intelligent Information Processing. Volume 288. Springer Science & Business Media.

[16] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. 2014. “The Stanford CoreNLP Natural Language Processing Toolkit.” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[17] N. Subhani and R. D. Kent, "Continuous process auditing (CPA): An audit rule ontology based approach to audit-as-a-service," 2015 Annual IEEE Systems Conference (SysCon) Proceedings, Vancouver, BC, 2015, pp. 832-838.

[18] Y. H. Tung, S. C. Lo, J. F. Shih and H. F. Lin, "An integrated security testing framework for Secure Software Development Life Cycle," 2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS), Kanazawa, 2016, p.

Cloud Computing Auditing

Roadmap and Process

Mohammad Moghadasi, Dr. Seyed Majid Mousavi, Dr. Gábor Fazekas
Department of Informatics, University of Debrecen, Hungary

***Abstract*—Cloud Computing is a new form of IT system and infrastructure outsourcing as an alternative to traditional IT Outsourcing (ITO). Hence, migration to cloud computing is rapidly growing among organizations. Adopting this technology brings numerous positive aspects, although imposing different risks and concerns to organization. An organization which officially deputes its cloud computing services to outside (offshore or inshore) providers and implies that it outsources its functions and process of its IT to external BPO services providers. Therefore, customers of cloud must evaluate and manage the IT infrastructure construction and the organization's IT control environment of BPO vendors [25]. Since cloud is an internet-based technology, cloud auditing would be very critical and challengeable in such an environment. This paper focuses on practices related to auditing processes, methods, techniques, standards and frameworks in cloud computing environments.**

Keywords—Cloud computing; cloud auditing; IT outsourcing

I. Introduction

Outsourcing IT operations is not a new concept. Recently, Cloud computing is a new concept in the outsourcing IT operations as an adopted paradigm for delivering IT services over the Internet by organizations. Maximum utilization of hardware and software by sharing resources through virtualization, elastically, flexibility and decreasing capital and operational expenditures (CAPEX and OPEX) has made popular this IT paradigm. Supporting thousands of business needs, Simplify and streamline enterprise collaboration, cost management, availability, and scalability are only a few of countless motivations for organizations to adopt cloud computing. This new technology also brings risks and concerns to organization. The number of IT outsource providers in cloud recently has increased and this increment has brought large number of risks to the scene. As well as the providers, IT outsourcing risks are considerably increased, these risks are applied and enforced all over the life cycle of cloud computing and its services, either an organization is already implemented cloud services and solutions within its environments or planning on becoming a cloud-based company or an affiliated organization [1]. This paper contributes to provide a comprehensive perspective in auditing processes, different approaches and frameworks, and key concepts in cloud computing environments.

According to SOX section 302 [2], Chief Financial Officer (CFO) and Chief Executive Officer (CEO) support the credibility of their corporation and are responsible for the accuracy of financial reports of their company annually and quarterly. Even if these business reports and relevant data exist in different locations, units, teams, departments, business sites, data centers and or in different cities or countries [35]. Thus, for organizations, it is important that the IT operations in the cloud comply with applicable legislation and SLAs (Service Level Agreement).

As cloud computing is a new orientation in IT and business processing outsourcing, organizations would make good use of this technology in their business procedures [26]. The importance of IT auditing and especially cloud computing auditing is an essential effort to ensure the proper functioning processes of an organization's IT systems, management, operations and related processes, to avoid fraudulent, in order to have comprehensive and accurate financial view of their business. Internal auditing is a crucial component of any organizational processes; thus, being a strategic collaborator to an organization is not the only essential element but performing ordinary quality assurance is also crucial in cloud-based organizations. As well as enhancing the organizations' productivity and efficacy in the improvement of their IT processes throughout these activities. [3,4 and 34]. Hence, this paper aims to provide a contribution to the understanding of different aspects in cloud auditing [33], its risks and benefits in cloud environments, in order to shed light on the cloud computing audit practices. In this paper, we address different cloud auditing practices related to processes, techniques, test steps, standards and frameworks with the purpose of answering the following questions: 1) How to maximize the value of the IT audit function? 2) What are specific components and key controls which might be necessary for cloud environment auditing? 3) How to determine appropriate cloud auditing process? 4) Which frameworks and standards are recommended to do a cloud audit?

The present paper is structured according to the followings: The forthcoming section differentiates between IT outsourcing and cloud computing [25]. The implication and importance of cloud auditing are explained in section three. After that, in section four, cloud auditing approaches and techniques are discussed. Test steps and key controls come in section five. Sixth section points out cloud auditing standards and frameworks [25]. And at last as a final result of this paper a conclusion is presented in section eight.

II. IT Outsourcing and Cloud Computing

A. IT Outsourcing

Most of the times, impossibility of conducting all aspect of affairs, business process or being temporarily some processes justify hiring external required resources and professionals to perform operations in organization [5]. Utilization of external required resources to conduct a specific business processes, is usually a strategic decision based on a desire to reduce costs and to allow a company to focus on its core competencies. IT outsourcing is a subset of business process outsourcing (BPO) [6]. The main implication of IT outsourcing is, moving all or parochial of IT functions to an external company. Two main objectives of IT outsourcing for most of organizations are: lack of adequate resources and cost reduction strategies. IT outsourcing has different models for organization which is chosen based on organization needs and strategy, but adopting a model is not always easy. Types of IT outsourcing models are: Application Service Provider (ASP), Application Develop & Maintenance (ADM), IT Infrastructure Outsourcing (ITO), and IT Services [25].

B. Cloud Computing

Cloud Computing facilitates and empowers the IT process outsourcing by which the approachability of IT-related services and resources such as delivered-platforms, hardware, software on the internet as service for a basis charge monthly, quarterly or annually [7]. National Institute of Standards and Technology also known as the NIST has the following definition for Cloud Computing [8]:

‘‘Model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” [8]. The opportunity to purchase and centralize IT support/services by using external infrastructure and another organization’s resources to provide the mentioned services is an outsourcing decision in cloud computing [6,7]. The Cloud Computing technology is processed through three layers (IaaS, PaaS, SaaS), relying on such computing concepts as virtualization, grid and distributed computing and Service Oriented Architectures (SOA) [9,25]. The main concepts which are listed by the cloud community and the NIST organization include hosting, roles, deployment and service models and necessary and important specification [10].

There are different perspectives on how IT process outsourcing differs from Cloud Computing. In case of Cloud Computing, an application is originated in the cloud itself by default, whereas with IT outsourcing, a specific function is relocated from its original geographic jurisdiction, enterprise, site or department [11]. Network-based applications, length of contract, saleable services and requests, flexible self-service interface and shared resources, in order to achieve maximum density, are the other major differences between IT outsourcing and cloud computing.

Cloud computing adoption and its popularity among companies are rapidly growing because of both economic and technical aspect. With respect to its positive aspects, Cloud Computing risks should be considered and mitigated as well. Cloud computing risks affect quality of service (QoS) to customers. The major cloud environment risks have been reported in control access management, vendor management, regulatory compliance, data privacy, and operational process. In order to overcome such risks continuous audit is needed in cloud environment.

III. Cloud Auditing

IT outsourcing life cycle consists of several phases, and each phase may encounter uncertainties and risks. Therefore, in order to risk quality control and identification of IT outsourcing, a systematic auditing is required to be applied to the entire life cycle of IT process outsourcing.

Cloud computing appearance is a novel form of IT system and infrastructure outsourcing as an alternative to traditional IT outsourcing. According to technical and economic perspectives, Cloud Computing has numerous advantages pushing for its development and quick assumption [13]. An organization which officially deputes its cloud computing services to external providers and points out that it outsources the process of its IT and functions to external ITO services providers [25]. According to SOX regulations, management’s responsibility in sustaining efficient internal control over financial reporting will not be affected or reduced by using a service organization [2]. Thus, Customers of cloud computing execute a consecutive and successive evaluation over the IT control and provider environment [25].

Regardless numerous advantages, cloud computing has been associated with numbers of risks and concerns, which attracted IT auditor’s attention. Auditing’s main and primary implication is; “an independent and autonomous experiment of an organization’s management assertions declared by an external sanctioning body that must follow a set of guidelines and standards” [14, 25].

Thus, IT auditors must have complete comprehension of cloud computing and auditing methods to assess, evaluate and assurance of regulatory compliance and SLAs (Service Level Agreements). The auditing work is much different and more complicated than regular IT auditing, and as a result cloud computing involves external vendor’s help or partner’s support to control [12,15,16,19 and 25]. The cloud audit can be internal or external as regular IT audit. Internal audit is performed by inside auditors to analyze and assess the data and processes for improving organization’s effectiveness and efficiency. External audit is conducted by auditing firms or expert auditors. Organizations are obliged to comply with mandatory audit to demonstrate regulatory compliance and voluntary audits include processes, practices, internal controls and the independent validation or quality assurance and those associated with certification [17]. A major asset in corporate governance and financial reporting and its public confidence improvement is efficacious auditing function if this effective action includes following elements; approved audit-charter and audit-committee existence, unlimited scope, stakeholder support, un-restricted access, sufficient staff, professional-audit standards, competent leadership, adequate funding, formal mandate and organizational independence [18,20].

Providing an assurance engagement between consumers and cloud services provider for cloud auditing is a major motivation to raise the measurement of criteria against cloud services and the confidence of cloud consumers. Cloud auditing is a major tool to help organization’s board and management, as an evaluation function to identify risks [26,29]. The auditing in cloud environment may be applied on specific sector such as: entity-levels, application systems, security, application systems, data center, virtualized environments and web application (IT governance and customer relationship management (CRM) and enterprise resources planning (ERP)) [11,26].

IV. Cloud Auditing Techniques and Process

The auditing process is a consolidation of different ISA (International Standards on Auditing) and audit methodology which can be performed by internal or external auditors. It is important for auditor to consider that auditing is a continuous process and performed at various stages. It is also noteworthy for auditor to note that some stages can be combined with other stages or may need to return and rethink on previous completed stages.

The main purpose of IT auditing is to provide an independent opinion to ensure whether IT operations and governance comply with standards and SLAs. Substantive testing phase alongside Tests of controls as well as planning of an audit are the three main phases that can be performed by IT-audit process [12,21 and 25].

- *Audit planning phase*: In the first phase, planning phase, auditor must gain deep comprehension of the nature of business. Collecting and analyzing important information (such as IT operations, internal controls and risks) must be performed by auditor. During this phase auditor comprehend organization's policies, practices and structure. Practical approaches to gain evidence are: reviewing documentation and application, interview (management, employees), questionnaires. The three major stages of audit planning phase are introduced as followings: (1) procedures of substantive examination and scheme examination of controls, (2) organization's structure, course, terms and conditions review, (3) application and comprehensive control review [12,25].
- *Control examination phase*: In second phase auditor performs different tests to ensure internal control compliance over IT operational activities. During this phase, the auditor assesses quality of controls. The stages of this phase are: (1) control stage of specifying the reliance degree, (2) executing examinations of controls, (3) test outcome evaluation [4,5,21 and 25].
- *Main substantial testing phase:* The third phase focuses on investigation of financial data. During this phase substantive tests are performed in data files by using appropriate audit tools and techniques. The three stages of this phase include: (1) executing main substantial tests, (2) report of result assessment and issuing report of auditor, (3) creating audit report [21].

Checking the quality of processes of an IT operational are the purposes an IT audit's, whether the objectives and targets of a company or establishment are met by their IT processes or not [5,25]. An independent and systematic test, by ISO describes a quality audit as; "to achieve an organization's targets, planned regulation and adjustments must be complied with results of relevant quality operations whether or not, effective implementation of these regulations is suitable" [30,36]. It is explicit that IT auditing and all related areas on cloud computing operations can be developed by the quality concepts (Merhout. J.W., Havelka. D) [14,36], amongst the 108 identified unique factors, and based on existence control, efficacy, factor determination and what or who a propounded framework of IT audit quality consisting of eight categories. These eight categories are described in Table 1.

TABLE I. Descriptions of IT Audit Quality Framework Categories (Merhout and Havelka)

Categories	*Description*
Factors of Audit team	Teamwork experience, communications and teamwork quality,
Audit methodology and processes factors	IT audit team follows particular practice and procedure,
Client-controlled organizational factors	Critical client partnerships during an audit's course, management's support and adequacy of documentation,
IT Audit-Controlled Organizational Factors	Business unit comprehension, client relationship with organization, allocating sufficient time for all of audit, leadership and IT organizational assessment and change ability,
Technical qualification factors of IT audit personnel	Personnel experience, risks understanding and weakness control project management,
Interpersonal and social factors of IT audit personnel	Enthusiasm, capability and willingness to change, communication proficiency, motivation and independence,
Organizational environment and enterprise environmental factors	Internal audit's reporting structure, recent audit numbers, corporate culture, financial resources and audit's value perception,
System and target process factors	System type and complexity, processing of manual versus automatic amount, system or process documentation level, clearly defined project scope

Additional value over the primary assurance objectives can be provided by IT audit activities [14,25]. Work of regular IT audit can be similar to the cloud computing audit work as long as effective auditing framework and risk assessment method are chosen and followed by cloud computing's IT auditors. [25]. There are two major IT audit processes: Risk-based IT audit and Value-added IT audit [14]. Thus, during the cloud audit, the audit team establishes auditing process on value-added audit or risk-based audit. Each process has a specific auditing domain.

C. Value-Added IT Audit

Value-added and quality are consumer-focused concepts in organizations, and a new trend to conduct IT auditing. Value-added IT focuses on the organization's IT operations and capabilities. Value-added audit is a proven method to assess effectiveness of an organization's operations (such as quality, business process, IT, etc.) that verifies compliance with policies and procedures. According to the IIA (Institute of Internal Auditors) [9,31], following scopes are covered by a quality audit: Business process efficiencies, Trade process and business control, Commerce risks, Quality and utilizable efficiency and effectiveness, Cost diminution situations, Corporate governance effectiveness and Waste deletion opportunities [4,22,23], and also value-added auditing is defined as follows: "Internal auditing is designed to improve companies' operations and enhance relevant assurance values, since this activity is an independent, objective and consulting, and using this would facilitate organizations through the accomplishment of their objectives by bringing disciplined and systematic techniques to appraise and progress control, risk effectiveness management and governance methods" [14,25,36]. Some of value-added IT audit benefits are in Table 2.

TABLE II. VALUE-ADDED IT AUDIT SERVICES (MERHOUT AND HAVELKA)

1. Improved information technology governance by using proven return on investment in IT
2. Improved business process management or operational expediency and productivity through IT process and business progress reengineering by using audit documentation,
3. Improved risk mitigation through enhanced enterprise risk management (ERM) awareness by using audit observations,
4. Improved business continuousness and associated systems disaster recovery planning,
5. Improved systems development quality approach,
6. Increased trust development and organizational communication through facilitation among various stakeholders,

D. Risk-based IT Audit Process

An audit includes risk-based audit to focus effectively and expeditiously on the timing, nature and extension process in the mentioned scopes and to assure of having misstatement cause and its potential material in financial reports [23,25]. Internal audit can be accredited and empowered by risk-based audit to assure the board that whether or not risk management processes are complying effectively.

A risk-based IT audit identifies substantial IT threats and risks in IT operations area such as: risk assessment, security, data safety, IT governance, and systems development. Risk-based IT audit defines appropriate strategy for assessing IT operations and present proper solutions for risks mitigation. Even though IT function's quality and modality maintenance as well as value development is targeted by value-added IT audit but maximizing IT quality is the goal. [25].

V. CLOUD AUDITING: TEST STEPS AND KEY CONTROLS

Following objectives must be covered by the cloud audit as defined by ISACA [28]:

- Providing stakeholders with internal security policy and successful control process of the cloud computing service provider and productiveness evolution
- Providing an interface between the service provider and organization's client for identifying insufficiencies and inadequacies of internal control
- Providing an assessment criteria and report of capability and quality to audit stakeholders to be confident of the certification and accreditation of service provider and its internal controls

In addition to above objectives, the auditor must consider control access, authorization and trusted control frameworks, communications latency, data breach notification and international laws. A transferred system to the cloud or/and IT services support and reinforce business functions which must be contemplated and considered by the cloud auditor [3,32].

Cloud auditor must understand the associated risks, dealing with, and ability to develop an audit strategy and plan. Since cloud computing architectures consist of different models, services and components from other form of IT outsourcing, cloud auditor also must consider following points:

1) During cloud migration one or some parts of an application may not be compatible with cloud environment. Because the most of applications and related functions rely on internal corporate's network, not over the internet.

2) Web applications must be assessed to assure access controls, authentication, and monitoring.

3) Regardless any complexity, Identity and access management must be assessed, to ensure appropriate control access over resources.

4) Assessing endpoint systems ensure auditors that systems have sufficient security to gain legitimate access to cloud resources.

5) All communications and correspondence between vendors and corporate should be inspected based on SLA.

During the cloud auditing, audit models, standards and frameworks would be determined by audit team. As indicated in section four, cloud auditing can be performed as value-added or risk-based. The main differences between risk-based audit and value-added audit are that risk-based audit brings data security, data protection and risk assessment into focus however the concentration of value-added audit is on risk migration and cloud investment and their improvement [15,25 and 37]. Quality of services and risks assessment of cloud environment also are two important issues in firm by seeking an audit request through internal or external cloud auditors.

In a cloud computing environment, cloud computing audit can be conducted in an alternative way, in which the auditors intelligibly should comprehend the available technology of cloud computing thought a value-added method and the related value would be created once the organization adopts the approach. [25, 37]. And focusing on targeted attractive features to clients by the auditing work would provide followings available benefits, values and possibilities if cloud computing is adopted:

Solutions for every financial plan and necessity, more appropriate use of resources, raised flexibility, bigger agility and supported efficacy, ameliorated collaboration, cost avoidance, reformed cost model, access to novel technology, and developed security [25,29].

In general, the first step for a firm to adopt cloud computing is to select the right cloud vendor. Since cloud vendors have direct impact on cost, quality, and operational processes in cloud environments, thus, selection, and continuous evaluation of cloud vendor should be considered by cloud auditors. Auditing work in cloud computing requires more effort than ordinary IT auditing processes, as it requires the support and supervision of wanted information technologies of an external vendor [12,15,16 and 19]. Important factors in vendor's selection and evaluation measurements are: financial heath, operational performance metrics, expertise, risk factors, etc.

The next crucial step is providing a service contract with strong service level agreements (SLAs). The best time for cloud auditing is before the contract is finalized and signed. In this step all contractual obligations should be clearly stated. Some of these obligations are: SLAs (Availability, Performance, support coverage), SLA security (Encryption, Data privacy, data retention, data destruction, security training and background check, control frameworks), compliance assessments (SSAE 16, ISEA 3402), penalties for non-performance, condition for terminating, subcontracting relationships (right of denial, access to subcontractor's ISAE 3402), etc.

Eventually, providing a comprehensive report is required. The report is written in a standard format and included all audited cloud sections. Preparing a complete report is a major factor by which the reputation of internal audit department is established such as: data storage, cost savings, security issues, cloud governance, risks and so on [39]. Cloud auditor's report contains five sections: objectives, procedures, findings, recommendations and limitations based on the policies, standards, risks and business process.

After the cloud auditing process, an organization gain comprehensive view of pros and cons of cloud environment in its business. Moreover, strengths, weaknesses, associated risks and security breaches are clearly realized. Logical solutions can be recommended by auditors and organization executives and board need to work out strategies for solving occurring IT-related imperfections and later emphasize their complete business process management [25].

VI. Frameworks and Standards

IT audit standards provide a set of criteria, guidance, frameworks, procedures and methodologies that help to determine the extent of audit steps in order to how an audit should be conducted and what audit reports should be issued for IT engagements.

The fast evolution of cloud computing services and lack of sufficient standardization for these services caused utilizing many traditional IT audit standards for cloud auditing. Security, privacy and SLA's are potential challenges and concerns by cloud computing. There are several active organizations which have a number of guidance, standards, frameworks and metrics to assess cloud computing environment such as: ENSIA (European Network and Information Security Agency), ISACA (Information Systems Audit and Control Association), CSA (Cloud Security Alliance), and NIST. The publications of these organizations can be robust references for cloud audits. In the following we explain some of these organizations in short:

A summary of the possible negative outcomes in information security was provided by ENISA to stakeholders. It is a completely consultative organization, at the same time, it has accredited research related to security issues, such as "Cloud Computing Risk Assessment" [24], published in 2009. This paper strongly recommended several key points such as: continuous trust between cloud vendors and clients, Data protection in large-scale environments, large-scale systems' interoperability, resiliency, and monitoring. ENISA is following up different cloud activities and has robust frameworks which can be utilized as a useful reference for cloud auditors such as: Managing security through SLAs, Critical cloud services, Cloud Security and Resilience Expert Group, Good practice guide for Governmental clouds, Incident reporting for Cloud Computing, Certification in the EU Cloud strategy, and Cloud Certification Schemes List (CCSL).

A non-commercial formation as CSA with the purpose of supporting the application of most suitable practices, aims to assure security in cloud environments. Access Management Guidance and Identity has been introduced by CSA [26], including the most optimal solutions to ensure secure access management and identities. CSA research areas include cloud standards, frameworks, certification, guidance and tools. Some of important CSA's published documents and frameworks which can be very useful for cloud auditors during auditing process are: Cloud Computing's Critical Areas of Focus and their Security Guidance, Top Threats to Cloud Computing [32], GRC (Governance, Risk and Compliance) Stack, Cloud Controls Matrix (CCM), Cloud-Trust Protocol, and Consensus Assessments Initiative Research.

ISACA is a leading global organization in the development, adoption, and practices for information systems [27,28]. The popular ISACA's framework for IT management and governance are Control Objectives for Information and Related Technology also known as COBIT [15, 19 and 38]. "Controls and Assurance in the Cloud" is one of ISACA publications consist of practical guidance to provide cloud governance and control frameworks through an audit program by using COBIT 5. "Cloud Computing Management Audit/Assurance Program" is another useful published resource providing guidelines for the finalization of a particular and especial assurance procedure by ISACA [12,27 and 28].

The Unite States Department of Commerce's non-regulatory delegation known as NIST, supports innovation through research, measurements, standards, business services and other programs [8, 26]. NIST has three Special Publication subseries: SP800, SP1800, and SP500. The NIST 800 series is a set of publications as result of exhaustive research work for optimizing the computer security and describe policies, procedures and guidelines [10,11,26]. These publications cover all NIST-recommended procedures and criteria for assessing, threats, vulnerabilities and risk mitigation. The publications can be utilized as directions for the implementation of safety norms and auditing procedures as juristic references [37].

VII. Conclusion

Cloud computing is the latest evolution in the IT outsourcing world. Cloud adoption brings benefit for enterprises such as: business agility, data availability, ease of use, cost savings, and sustainability, but these benefits must be weighed against potential risks. In this paper we have provided a comprehensive perspective in the field of cloud computing audit. We discussed different approaches and techniques for auditing in Cloud environment that have strong benefit for cloud adopters and auditors. To contextualize and study Cloud auditing, we have investigated the implications of IT auditing, cloud computing and definitions for key concepts. We have then determined test steps, key controls, and additional factors which have to be considered in cloud environment. Finally, we have introduced active organizations and their appropriate standards and frameworks for cloud computing assessment and audit.

ACKNOWLEDGMENT

The present paper was supported by the University of Debrecen to development and research program for the project “Auditing and quality assurance strategies in Cloud Computing services”. I am immensely honored and gratefully thank those whom I have had the pleasure to work with on this paper and other related manuscript and projects.

REFERENCES

[1] Kalaiprasath, R., R. Elankavi, and R. Udayakumar. "Cloud security and compliance-a semantic approach in end to end security." International Journal on Smart Sensing and Intelligent Systems 10 (2017): 482-495.

[2] David Balovich: "Sarbanes-Oxley Document Retention And Best Practices" by 3JM Company Inc., Lake Dallas, Tx, May 09, 2007, Creditworthy News.

[3] Al-Twaijry, A. A. M, Brierley, J. A, & Gwilliam, D. R. (2003). The development of internal audit in Saudi Arabia: An Institutional Theory perspective. Critical Perspective on Accounting, 14, 507-531. doi:10.1016/S1045-2354(02)00158-2.

[4] Savouk, O. (2007). Internal audit efficiency evaluation principles. Journal of Business Economics and Management, 8(4), 275–284.

[5] D.C. Chou, An investigation into IS outsourcing success: the role of quality and change management, Int. J. Inf. Syst. Chang. Manag. 2 (2) (2007) 190–204.

[6] Mousavi. SM.,et al.: “Increasing QoS in SaaS for low Internet speed connections in cloud”, The 9th International Conference on Applied Informatics, Eger, Hungary Feb 1 2014, pp. 195-200.

[7] Yigitbasioglu, Ogan, Kim Mackenzie, and Rouhshi Low. "Cloud Computing: How does it differ from IT outsourcing and what are the implications for practice and research?." The International Journal of Digital Accounting Research 13 (2013): 99-121.

[8] P. P. Mell, T. Grance, “The NIST Definition of Cloud Computing”, NIST Special Publication 800-145, National Institute of Standards and Technology Gaithersburg, MD 20899-8930, September 2011.

[9] YOUSEFF, L., BUTRICO, M. & DA SILVA, D. (2008): "Toward a Unified Ontology of Cloud Computing", Gce: 2008 Grid Computing Environments Workshop: 42-51.

[10] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, D. Leaf, NIST Cloud Computing Reference Architecture NIST Special Publication 500-292, 2011.

[11] Mousavi. SM., Fazekas.G.: “A Novel Algorithm for Load Balancing using HBA and ACO in Cloud Computing Environment”, International Journal of Computer Science and Information Security, June 2016, 14(6), pp.48-52.

[12] S. Gadia, “Cloud computing: an auditor’s perspective”, ISACA (Information Systems Audit and Control Associatio) Volume 6, J. 6 (2009).

[13] Aceto, G., Botta, A., De Donato, W., & Pescapè, A. (2013). Cloud monitoring: A survey. Computer Networks, 57(9), 2093-2115.

[14] Mousavi. SM. et al: “A load balancing algorithm for resource allocation in cloud computing”, Advances in Intelligent Systems and Computing, Recent Global Research and Education: Technological Challenges, Springer International Publishing 2017, vol 66(16), pp. 289-296.

[15] V. Raval, Risk landscape of cloud computing, ISACA J. 1 (2010).

[16] Mousavi. SM. et al: “Dynamic Resource Allocation in Cloud Computing”, Journal Acta Polytechnica Hungarica, March 2017, 14(3), pp. 80-101.

[17] Gantz, Stephen D. The Basics of IT Audit: Purposes, Processes, and Practical Information. Elsevier, 2013.

[18] Belay, Z. (2007). A Study on effective implementation of internal audit function to promote good governance in the public sector. Presented to the “The Achievements, Challenges, and Prospects of the Civil Service Reform program implementation in Ethiopia” Conference Ethiopian Civil Service College Research, Publication & Consultancy Coordination Office.

[19] T.W. Singleton, IT audits of cloud and SaaS, ISACA J. 3 (2010) 1–3.

[20] De Smet, D, & Mention, A. (2011). Improving auditor effectiveness in assessing KYC/AMLpractices: Case study in a Luxembourgish context. Managerial Auditing Journal, 26(2), 182–203.

[21] J.A. Hall, Information Technology Auditing and Assurance, third edition South-Western Cengage Learning, Mason, OH, 2011.

[22] Mousavi. SM., Fazekas. G.: “Dynamic resource allocation using combinatorial methods in Cloud: A case study”, 16th international conference CogInfoCom 2017, pp. 221-232.

[23] Yazdankhah, F, Honarvar, AR "An Intelligent Security Approach using Game Theory to Detect DoS Attacks in IoT." International Journal Of Advanced Computer Science And Applications 8.9 (2017): 313-318.

[24] Sookhak, M, et al. "Remote data auditing in cloud computing environments: a survey, taxonomy, and open issues." *ACM Computing Surveys (CSUR)* 47.4 (2015): 65.

[25] Chou, David C. "Cloud computing risk and audit issues." Computer Standards & Interfaces 42 (2015): 137-142.

[26] Mousavi. SM., Fazekas. G. (2017): “Dynamic resource allocation in Cloud Computing using a new hybrid Metaheuristic algorithm”. PhD thesis 2017.

[27] Identity Management Audit/Assurance Program, ISACA, by ISACA,Identity Management, 2013.

[28] ISACA Issues Four New Audit Programs on Cloud Computing, Crisis Management, Security and Active Directory, ISACA, 2010.

[29] Mousavi. SM.,Fazekas. G.: “Increasing QoS in SaaS for low Internet speed connections in cloud”, The 9th International Conference on Applied Informatics, Eger, Hungary Feb 1 2014, pp. 195-200.

[30] D.C. Chou, A.Y. Chou, Analyses of software quality and auditing, in: C.V. Brown, H. Topi (Eds.), IS Management Handbook, seventh editionCRC Press, Boca Raton, FL,2000.

[31] Internal Auditors, Global Institute of Internal Auditors (IIA), Available at http://www.theiia.org2012 (accessed November 1, 2012).

[32] Halpert, B. Auditing Cloud Computing: A Security and Privacy Guide, Wiley Corporate, 2011.

[33] Richard Bradford-Knox. "Approaches to and the Management of the Audit Process in the Food Industry", British Food Journal, 2017.

[34] Cloud Computing, European International Journal of Science and Technology (EIJST).

[35] NACM National Trade Credit Reports issued by trade credit report team, published on credit worthy and tradecreditreport[dot]com websites.

[36] Merhout. J.W., Havelka. D.,"Development of an Information Technology Audit Process Quality Framework". Conference: Reaching New Heights. 13th Americas Conference on Information Systems, AMCIS 2007.

[37] Kenneth G Crowther, Yacov Y. Haimes, M. Eric Johnson. "Principles for Better Information Security through More Accurate, Transparent Risk Scoring", Journal of Homeland Security and Emergency Management, 2010.

[38] David C. Chou. "Cloud computing: A value creation model", Computer Standards & Interfaces, 2015.

[39] Jaydip S., "Security and Privacy Issues in Cloud Computing", in book: Architectures and Protocols for Secure Information Technology Infrastructures, Edition: First Edition., Chapter: 1, Publisher: IGI-Global, USA, September, 2013.

Recommender System based on Empirical Study of Geolocated Clustering and Prediction Services for Botnets Cyber-Intelligence in Malaysia

Nazri Ahmad Zamani[1], Aswami Fadillah Mohd Ariffin[2], Siti Norul Huda Sheikh Abdullah[3]
[1,2]CyberSecurity Malaysia, Level 5, Sapura Mines Seri Kembangan, Malaysia
[3]Cyber Security Faculty of Information Science and Technology Universiti Kebangsaan Bangi, Malaysia

Abstract—**A recommender system is becoming a popular platform that predicts the ratings or preferences in studying human behaviors and habits. The predictive system is widely used especially in marketing, retailing and product development. The system responds to users preferences in goods and services and gives recommendations via Machine Learning algorithms deployed catered specifically for such services. The same recommender system can be built for predicting botnets attack. Via our Integrated Cyber-Evidence (ICE) Big Data system, we build a recommender system based on collected data on telemetric Botnets networks traffics. The recommender system is trained periodically on cyber-threats enriched data from Coordinated Malware Eradication & Remedial Platform system (CMERP), specifically the geolocations and the timestamp of the attacks. The machine learning is based on K-Means and DBSCAN clustering. The result is a recommendation of top potential attacks based on ranks from a given geolocations coordinates. The recommendation also includes alerts on locations with high density of certain botnets types.**

Keywords— Botnets; recommender system; predictive analytics; Big Data; cyber-threat intelligence; K-Means; DBSCAN

I. Introduction

Botnets are growing threats at global scale in recent years. This cyber phenomenon is growing exponentially with the increased of broadband penetration onto the global population. Furthermore, this trend is thought be directly related with mobile devices and computers are getting cheaper and more powerful over time. Botnets are malicious software (or malware) that ceded control of users devices and computers. These malware are responsible for being mediums for DDOS attacks, ransomware, and data mining [8]. According to SpamHaus Project site [9], the company malware division identified that there are 9500 botnets C&C servers detected on 1122 different networks worldwide. Malaysia is no stranger to the threat- the country is listed in rank 21 in the world botnet threat list with 96049 listed incidents detected in just 2017 alone [10, 11]. Adopting cyber security strategy framework provides an insight into the government's approach for protection of cyber space in the country [17].

Curbing the botnets incidents is quite an impossible feat as botnets infect machines via strength-in-numbers strategy. The more machine the botnets are able to infect, the better. Looking on the bright side, botnets are always in communication with their bot herders (or Command and Control) via their respective communication protocols. The exhibit communication IP addresses of these botnets can be traced of their geolocations. This information can be analyzed via Machine Learning to predict their next attacks patterns via geolocation vs. time information. Through this predictive analysis, IT experts are able to take precautions steps in mitigating the risks before the predicted botnets attack are taking place. Whether the predicted attacks are precise or the otherwise, considering the volume and the velocity of botnets attacks any good preparation could save an organization from any damage that they might incurred.

A recommender system is one of the popular applications that is built on top of a Machine Learning predictive analytics algorithms. It is widely used by companies such as Amazon, Target Corporation, and Netflix to drive sales. A recommender system predicts consumers' interests and provides recommendation to them through Machine Learning. The same concept can be applied for botnets patterns. This paper demonstrates our implementation of such recommendation system that is developed from acquired telemetric botnets sensors data. The predictive analytic is used to learn and analyze botnets source IPs and the target IPs. The results from such learning can be used to provide either a warning on the top-10 botnets attack or a warning on certain botnets attack based on user input geolocations.

The feasibility of such system opens up a novel defense mechanism that is scalable to the volume, velocity and the veracity of botnets activities nationwide versus timeline. Such scalability is an enabling feature to analyze more data and to come up with more accurate results in this era of ubiquitous computing. An accurate recommendation system is essential in not just as monitor-alert system but as to provide better consequential planning and actions for remedial or triage.

II. Background

Botnet signals the bot herders from internet-connected devices IP addresses. From the signals the location of the source IPs and the Command & Control (C&C) IPs can be located and logged via sensors. The number of recorded signals can be in huge volume and in high velocity in daily basis. Analyzing the logs can cost a lot of computing resources and therefore only a Big Data set up can handles such magnitude.

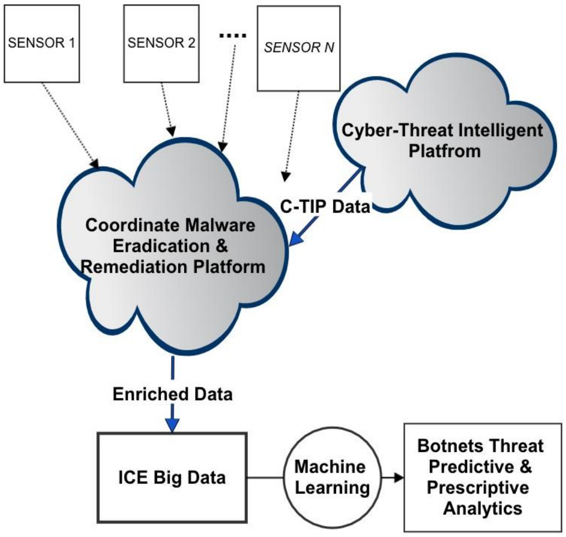

SENSOR 1
SENSOR 2
....
SENSOR N
Cyber-Threat Intelligent Platfrom
C-TIP Data
Coordinate Malware Eradication & Remediation Platform
Enriched Data
ICE Big Data
Machine Learning
Botnets Threat Predictive & Prescriptive Analytics

Fig. 1. ICE Recommendation System for Botnets Attack Prediction Services Schematic.

CyberSecurity Malaysia has developed Integrated Cyber-Evidence (ICE), a Big Data platform to analyze huge volume of digital evidence. The system comprises of nine processing nodes in which are all managed via Apache Metron [14].

For the recommendation system, ICE gets the input from malware cyber-intelligence enriched data from Coordinated Malware & Remediation Platform (CMERP). CMERP is another Big Data platform that CyberSecurity Malaysia developed to analyze in real-time data-in-motion from sensors deployed and from the Microsoft Cyber-intelligent Platform (or C-TIP). C-TIP here is a cloud service platform that Microsoft developed to monitor and fight botnets and the threat actors that run the malicious codes [13]. The enriched data that the CMERP tabulated is a precursor to the ICE system, in which the predictive analytics that leads to recommendation system can take place while the data is at rest. The following Fig.1 shows the schematic of how ICE botnets recommendation system works.

In building a recommender system, it is essential to firstly determine the type of application that is required and how it can be useful from the data ingested. The first step is to do the Exploratory Data Analysis (or EDA). EDA is performed in order to understand the data through statistical and visualization means in order to knife-out strategy to build a recommendation system. The botnets cardinality and IP addresses geolocations are drawn to determine the right algorithms that could cluster the geo-patterns. K-Means is one of the common used clustering algorithms in partitioning observations on the data-space into Voronoi cells. Another clustering algorithm under consideration is Density-based spatial clustering of applications with noise (or DBSCAN). DBSCAN is known for its density-based clustering- it groups together points that are closely packed together. With similar motivation of [18,19] study, projecting these Botnets into these data-space clusters may fit the predictive requirements for the geo-patterns of the signals.

III. Related Works

There are several ways used in various cyber security researches [16] to detect malware either in anti-virus software or end point protection such as signature based and behavioral based. Unlike signature-based, behavioral based able to detect malware that uses obfuscation technique even though it is time consuming with considerable false positive. This paper is inspired by the works of Coreia [12] and Casey [15]. Casey proposed a recommendation-verification system for predicting the Zeus malware infections via the Signaling Game methodology. Coreia on the other hand is using statistical characterization of botnets and their respective Command & Control traffics. His studies are focusing on the characterization of Denial of Service (DoS) attacks, spamming or phishing activities. Wei Xu [3] build a system on top various data feeds that predicts malware via malicious DNS domain. The system is leveraging on the knowledge of the life cycle of malicious domains, as well as the observation of resource re-use across different attacks. Another similar work by Truong & Cheng, where they proposed a method to detect Zeus and Conficker that utilizes domain fluxing by analyzing the extracted the DNS traffic length and expected value that can distinguish between a domain name, by a human or botnets [5]. Nguyen & Tran on the other hand, modeled user behaviors and applying heuristic analysis approach to mobile logs generated during device operation process [7]. For the task, they proposed a lightweight semantic formalization in the form of physical and logical taxonomy for classifying collected raw log data. There is also work done on honeypot dataset presented by Dowling & Seamus in [4]. In this paper, Seamus presented their analysis on honeypot dataset to establish attack types and corresponding temporal patterns. Their analysis shows the calculation of the probability of each attack type occurring at a particular time of day. Then they test these probabilities with a random sample from the honeypot dataset to see the geo-distributed patterns of the attacks. These attacks can take many forms and can come from different geographical sources. Another work that involves the applications of honeypot and sandboxing is by Mariconti, where he uses classification technique to learn about the different network behavior patterns demonstrated by target malware and generic malware [1]. They set up a sandbox and infected virtual machines with malware, recording all resulting malware activities on the network and then extracted meaningful features for classification. Gupta in [6] used Probabilistics Data Structure, specifically Bloom filter for setting membership on data the number of hits on suspicious nodes per unit time in network traffic data for their IDS (Intrusion Detection System). Lastly, Mezzour used empirical test alternative hypotheses on factors variations in the number of malware encounters [2]. Their analysis is focusing on regression analysis to test for the effect of computing and monetary resources, web behavior, computer piracy, cyber-security expertise, apart from the number of malware encounters. Another improvement of anomaly intrusion detection system has been proposed based on hybrid approach namely Fuzzy-ART and k-means clustering algorithm [20]. They swap the usage of both of the methods for getting the initial stated value for K-means, using Fuzzy-ART whereas

Fuzzy-ART uses K-means to reassign the data instances within clusters.

IV. Proposed System

The proposed system comprises of two sub-systems. The first sub-system is the training on the geolocation and the latter is the recommendation system. The training is the process of learning the enriched data, in which consists of preprocessing, clustering, in-hulling and merging geolocation data from cyber-threats enriched intelligence data. A prerequisite for the training phase to initiate is the EDA in getting the data overview in term of statistics, data visualization.

The recommender system is the process that responds to the geolocation input by user and return location alerts on types of botnets threat in the enquired vicinity. The system would also check at which cluster hulls that the requested geolocations inputs are closely correlate. It will return a warning for a user if the location is indeed a hotbed for botnets infections.

A. Exploratory Data Analysis

The main task of EDA is the hypothesis testing via statistics and data visualization. The first step is to seek the magnitude by cardinality of botnets activities in Malaysia. The following Fig.2 shows the word clouds visualization of the botnets infection in Malaysia in the first quarter of 2016 alone.

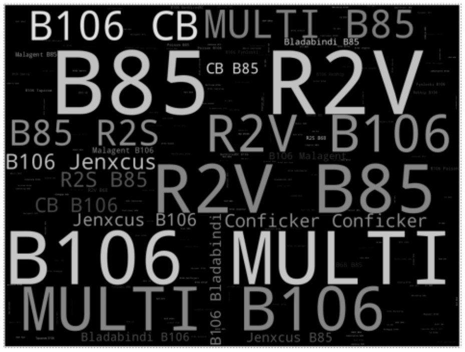


Fig. 2. The Word Clouds that Depict the Magnitude of Botnets Infection in Term of Cardinality.

From the word cloud, we can assume the top-3 botnets infections (with their own descriptions and threat risk) as tabulated in Table 1.

TABLE I. The Rank of Botnets Infections Analyzed from the Data

Rank	Threat code	Description	Threat Risk
1	B85-R2V	Family of worms that can steal usernames and passwords. Some can use PC for DoS attacks.	**Severe**
2	B106-MULTI	PC infected with multiple types of Botnets malware.	**Severe**
3	B106-CB	PC that is infected with both B106-Bladabindi & B106-Jenxcus	**Severe**

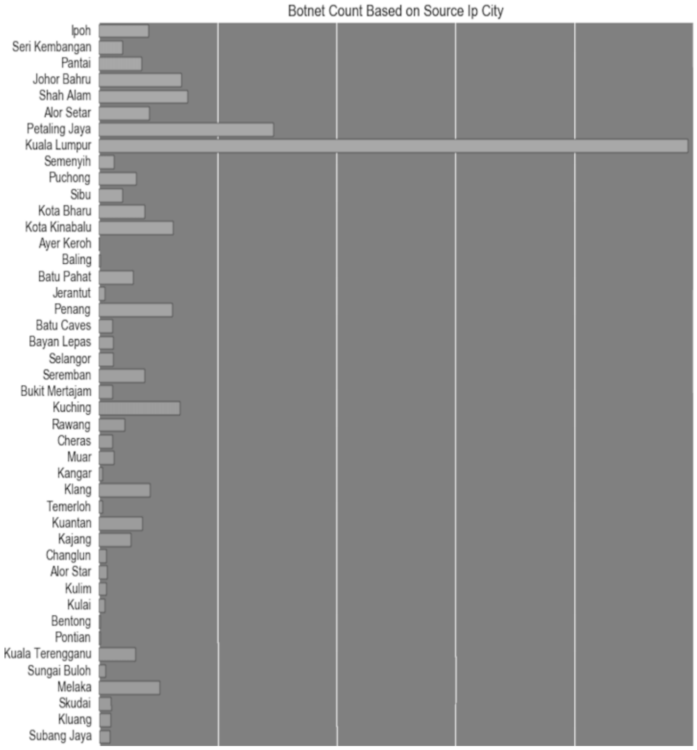


Fig. 3. The Histogram of Botnets Infections Versus Source IP Cities.

The next analysis is to find the distribution of infections in each of the city observed. Fig.3 shows the histogram of the number of infections over cities source IPs. The two cities which have the highest number of botnets infections depicted from the histogram are Petaling Jaya and Kuala Lumpur. Kuala Lumpur has the highest record of infection: almost reached a ten millions of records at just in the first quarter of 2016. Petaling Jaya on the other hand is lagged behind in number by less than a million. The numbers of infections of these two adjacent cities can easily balance out the number of other cities combined. In fact 70% of the infections are focused on regions adjacent to these two cities, in which are hubs for government offices, financials, manufacturing and commercials. Therefore, the predictive analytics for the botnet infections can be given top priority on these cities. Furthermore, the threats presented in Table 1 are commonly found in these areas.

B. Training Process

The Training process is a phase of selected machine learning algorithm learning from the ingested enriched data. The data is further pre-processed to discriminate columns required for the analysis. The data is then trained in two-tiers Machine Learning – K-Means and DBSCAN. K-Means is the first-tier clustering that clustered and partitioned the botnets IP address geolocation based on the number of botnets observations. The resulting Voronoi partitions are then taken on to the second-tier clustering, in which DBSCAN is to mark outlier points that lie alone in low-density regions (whose nearest neighbors are wide apart). The final cluster results are then merged for the next Recommendation process. Fig.4 shows the explained the proposed Training process.

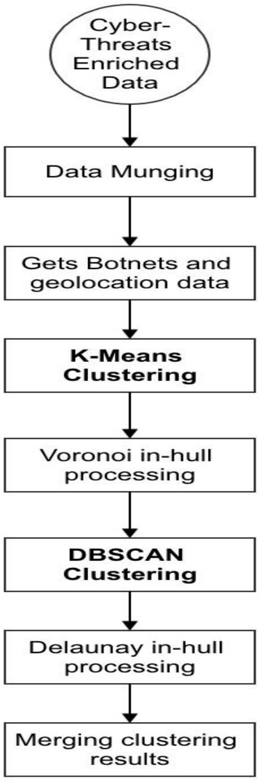

Cyber-Threats Enriched Data
Data Munging
Gets Botnets and geolocation data
K-Means Clustering
Voronoi in-hull processing
DBSCAN Clustering
Delaunay in-hull processing
Merging clustering results

Fig. 4. The Proposed Training Process Flow.

One of the challenges in the training phase is to determine the optimal value of *k* in calculating the K-Means for such huge dataset. We solved the problem by estimating the value via the Elbow method. The idea of the elbow method is to run K-Means clustering on the data for a range of values of *k* (for this setup 1 to 15), and for each value of *k* we calculate the sum of squared errors (SSE). The goal is to choose a small value of *k* that is still has a low SSE, and the elbow of the curve is usually represents of where the SSE value have diminishing returns by the increasing *k*. Fig. 5 shows the line plot of SSE versus the increasing *k* values. Here, the optimal value *k* at the elbow of the curve is *k*=5. This value is the one applied to the K-means clustering for the dataset.

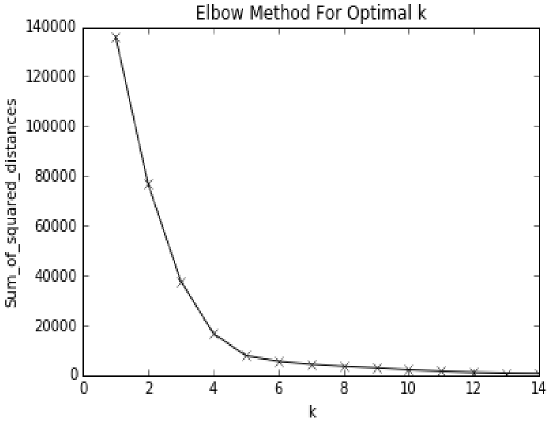


Fig. 5. The Elbow Method for Determining the Optimal *K* Value for K-Means.

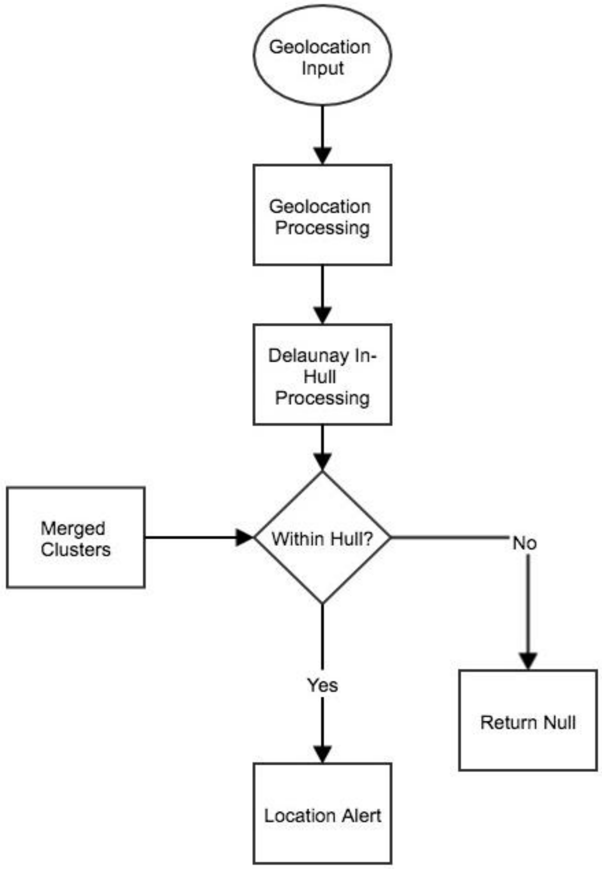


Fig. 6. The Recommendation Process Flow.

C. *Recommendation Process*

The recommendation process is where the system gives possible predicted botnets attacks based on a geolocation input. The input is processed with in-hull Voronoi and Delaunay processing as to find whether the distance correlation between the enquired geolocation points to the clusters of trained botnets threats is within hull. If the points are within hull, the location alert is issued with suggestion of related botnets types that may exist within the vicinity of the location. Fig.6 shows the Recommendation process flow.

V. METHOD OF EXPERIMENTS

The experiment is conducted on the enriched Cyber-Threats Intelligence data from the mentioned first quarter of year 2016. From the EDA performed on the data, we had decided to focus the predictive analysis on Kuala Lumpur, Petaling Jaya and other connected cities around Klang Valley. The recommendation system will show the top-3 botnets within the enclosure of the Klang Valley. In the second part of the recommendation system, an alert is given by the system on the threats that may exist from a given geolocation.

VI. ANALYSIS AND RESULTS

Our analysis shows that from the millions of log on the botnets infections, the K-Means clusters formed the Voronoi as depicted in the Fig.7. Each of the Voronoi cell represents malware cluster label.

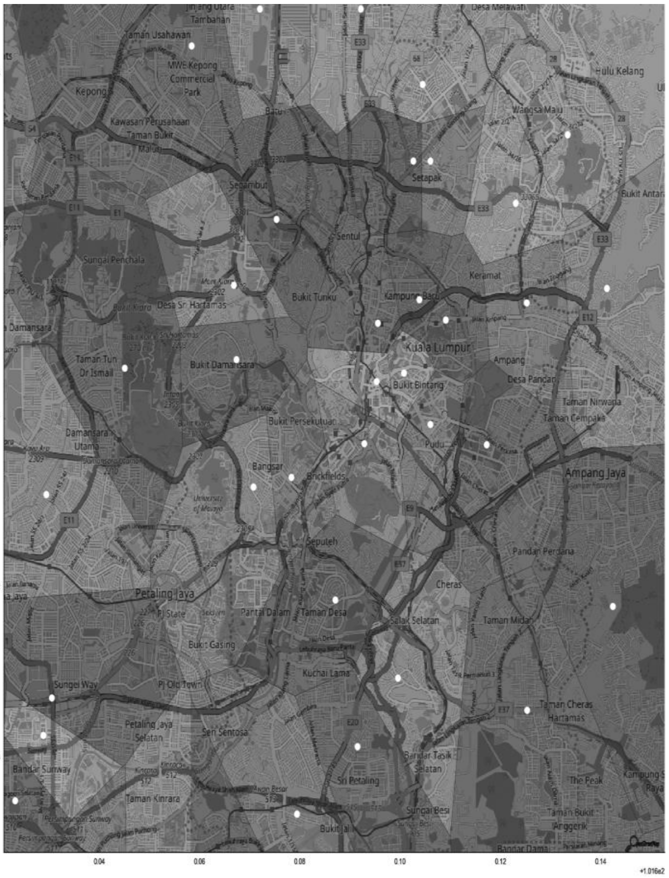


Fig. 7. The Voronoi Cells of K-means Clustering of Botnet Threats Around the Klang Valley.

From a selected geolocation point around the city center of Klang Valley, we can see the top 3 botnets threats, and the suggested locations of each suggested botnet within radius from the requested point. The Fig.8 shows the example of top three predicted botnets and the suggested dots are areas of which botnets infections are found to be active within the point parameters.

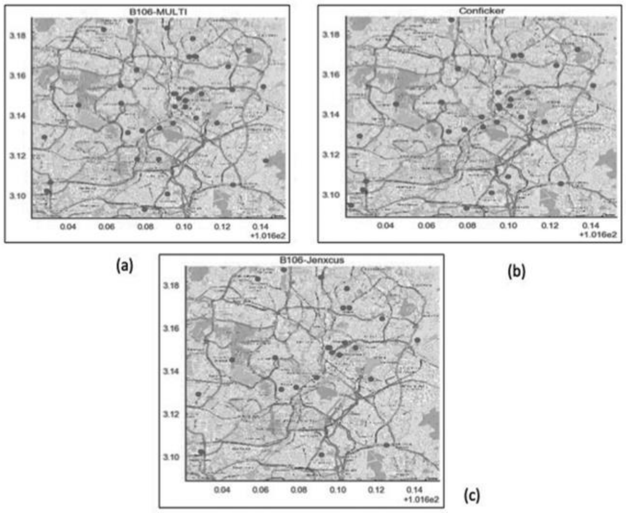


Fig. 8. The Example of Suggested top-3 Botnet Threats from a Requested Geolocations Point–(a) B106-MULTI, (b) Conficker, and (d) B106-Jenxcus.

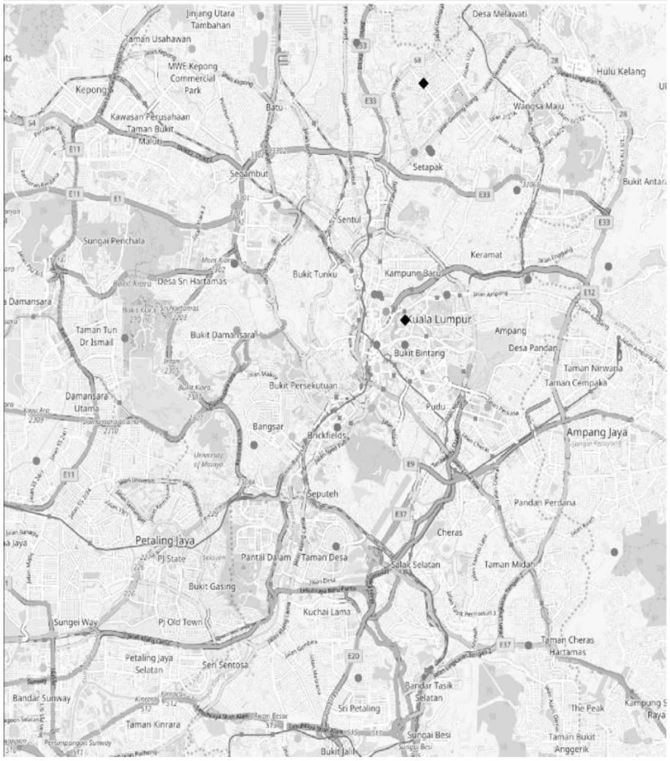


Fig. 9. The Alert of Botnet Threats around the given Geolocations Point, (*101.67, 3.139003*).

In the second part of the recommendation system, a geolocation is chosen randomly at (*101.67, 3.139003*). As shown in the Fig. 9, the system gives a recommendation of types of botnets threats existed. There are 73 warnings issued on the threats existed within the parameter of the requested geolocations, and is dot-labeled to differentiate each of the threat.

VII. CONCLUSIONS

Recommendation system is crucial in informing the public on the predicted botnet threat landscape based on their local area. The system is capable of providing predicted botnets threats that users have to be aware of in an area. The recommendation system works via Machine Learning algorithms inside ICE Big Data environment. ICE system learns the botnets threats IP geolocations through K-Means and DBSCAN clustering and partitioning the threats geographically. As a result the system will provide the top-three botnets in the alert feeds along with other suggested targeted areas on the map for awareness.

ACKNOWLEDGMENT

This Botnet Recommendation system is a project that is a part of the DSTIN program on Integrated Cyber-Evidence platform development, granted by Ministry of Energy, Science, Technology, Environment and Climate Change (MESTECC) to CyberSecurity Malaysia. A research grant UKM-AP2017-005/2 under Universiti Kebangsaan Malaysia has been a collaborator and partner to CyberSecurity Malaysia in research, academic and outreach programs especially in the field of digital forensics.

References

[1] Mariconti, Enrico, et al. "What's your major threat? On the differences between the network behavior of targeted and commodity malware." Availability, Reliability and Security (ARES), 2016 11th International Conference on. IEEE, 2016.

[2] Mezzour, Ghita, Kathleen M. Carley, and L. Richard Carley. "An empirical study of global malware encounters." *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*. ACM, 2015.

[3] Xu, Wei, Kyle Sanders, and Yanxin Zhang. "We know it before you do: predicting malicious domains." *Virus Bulletin Conference September*. 2014.

[4] Dowling, Seamus, Michael Schukat, and Hugh Melvin. "Using analysis of temporal variances within a honeypot dataset to better predict attack type probability." Internet Technology and Secured Transactions (ICITST), 2017 12th International Conference for. IEEE, 2017.

[5] Truong, D.-T. Cheng, G. “Detecting domain-flux botnet based on DNS traffic features in managed network”. Security and Communication Networks. Issues 14 Volume. 2016.

[6] Gupta, D, Garg, S Singh, A Batra, S Kumar, N Obaidat, M S. "ProIDS: Probabilistic Data Structures Based Intrusion Detection System for Network Traffic Monitoring". GLOBECOM 2017 - 2017 IEEE Global Communications Conference. 2017.

[7] Nguyen, G, B M Nguyen, D Tran, and L Hluchy. “A Heuristics Approach to Mine Behavioural Data Logs in Mobile Malware Detection System.” Data and Knowledge Engineering 115: 129–51, 2018.

[8] Tara Seals. “Bad Botnet Growth Skyrockets in 2017”. Infosecurity Magazine. https://www.infosecurity-magazine.com/news/bad-botnet-growth-skyrockets-in. 2018.

[9] SpamHaus. The SpamHaus Project. https://www.spamhaus.org/statistics/botnet-cc. 2018.

[10] Composite Block Listing. “CBL Breakdown by Country, Highest by Count”. Composite Block Listing, A Division of SPAMHAUS. https://www.abuseat.org/public/country.html. 2018.

[11] SeyedAliReza Vaziri. “Botnets: As We See Them in 2017”. Ripe NCC. https://labs.ripe.net/Members/alireza_vaziri/botnet. 2018.

[12] Correia, Pedro, Eduardo Rocha, António Nogueira, and Paulo Salvador. “Statistical Characterization of the Botnets C&C Traffic.” Procedia Technology 1: 158–66. 2012.

[13] Microsoft. Microsoft Secure. https://www.microsoft.com/en-us/security/intelligence. 2018.

[14] Apache Metron. http://metron.apache.org. 2018.

[15] William Casey, Evan Wright, Jose Andre Morales, Michael Appel, Jeff Gennari, Bud Mishra. "Agent-based trace learning in a recommendation-verification system for cybersecurity". 2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE). IEEE. 2015.

[16] Rami Sihwail, Khairuddin Omar, Khairul Akram Zainol Ariffin. A Survey on Malware Analysis Techniques: Static, Dynamic, Hybrid and Memory Analysis, 8 (4-2) : 1662-1671, 2018

[17] Salamzada, Khosraw and Zarina Shukur, and Marini Abu Bakar. “A framework for cybersecurity strategy for developing countries: case study of Afghanistan”. Asia-Pacific Journal of Information Technology and Multimedia, 4 (1). pp. 1-10. ISSN 2289-2192. 2015.

[18] Mohammed Ariff Abdullah, S.N.H.S. Abdullah, Md Jan Nordin. “Additional Feet-on-the-Street Deployment Method for Indexed Crime Prevention Initiative”, Jurnal Pengurusan 53(2018) 20 pages ,2018.

[19] Kamarul Ismail, and Nasir Nayan, and Siti Naielah Ibrahim. “Improving the tool for analyzing Malaysia’s demographic change: data standardization analysis to form geo-demographics classification profiles using k-means algorithms”. Geografia : Malaysian Journal of Society and Space, 12 (6). pp. 34-42. ISSN 2180-2491. 2016.

[20] Zulaiha Ali Othman, and Afaf Muftah Adabashi, and Suhaila Zainudin, and Saadat M. Al Hashmi. “Improvement anomaly intrusion detection using Fuzzy-ART based on K-means based on SNC Labeling”. Jurnal Teknologi Maklumat dan Multimedia, 10 . pp. 1-11. ISSN 1823-0113 Item availablity restricted. 2011.

Cyber Romance Scam Victimization Analysis using Routine Activity Theory Versus Apriori Algorithm

Mohd Ezri Saad[1]
Commercial Crime Investigation Department
Royal Malaysia Police, Bukit Aman
Kuala Lumpur, Malaysia

Siti Norul Huda Sheikh Abdullah[2], Mohd Zamri Murah[3]
Center for Cyber Security Faculty of Information
Science and Technology Bangi,
Selangor, Malaysia

***Abstract*—The advance new digital era nowadays has led to the increasing cases of cyber romance scam in Malaysia. These technologies have offered both opportunities and challenge, depending on the purpose of the user. To face this challenge, the key factors that influence the susceptibility to cyber romance scam need to be identified. Therefore, this study proposed cyber romance scam models using statistical method and Apriori techniques to explore the key factors of cyber romance scam victimization based on the real police report lodged by the victims. The relationship between demographic variables such as age, education level, marital status, monthly income and independent variables such as level of computer skills and the level of cyber-fraud awareness has been investigated. Then, the result of this study was compared with Routine Activity Theory (RAT). This study found that those between the ages of 25 and 45 years were likely to be the victims of cyber romance scams in Malaysia. The majority of the victims are educated and having a Diploma. In addition, this research shows that married people are more likely to be the victims of cyber romance scams. Study shows that non-income individuals are also vulnerable to being the victims because the study shows that 17 percent of respondents who are the victims are from this group. As expected, those who work and have monthly income between RM2001 and above are more likely to be targeted and become a victim of cyber romance scams. The study also shows that those who lack computer skills and less levels of cyber-fraud awareness are more likely to be victims of cyber romance scams.**

***Keywords*—Cybercrime; love-scam; routine activity theory**

I. Introduction

In the era of globalization information, users should be wiser in sharing their information especially personal information. Users negligence will give advantage to the cyber scammers [1], [2]. These scammers usually targeting people who looking for romantic partners as a victim, often via dating websites, apps or social media by pretending to be prospective companions. Scammers typically create fake online profiles designed to lure victims in [3]. They will start by profess strong feeling for the victim and ask to chat with them privately. As soon as the scammer has gained the trust from the victim, they begin requesting money by pretending to need money for some sort of personal emergency. Among other possibilities, the scammer may request photos or personal information that could eventually be used to blackmail the victim and extort more money [4].

These cyber romance scams has fraud Australians out of millions every years [1], [5]. No exception to Malaysia that has been the top 6 countries that recorded highest number of cybercrime cases, with total loses reaching RM1 billion (Source: Bernama Report in 2014). Other countries like Taiwan, China, Thailand, Indonesia and Hong Kong also were listed together. While, Sophos Threat Report in 2014 stated that cybercrime attacks have been increased in the first quarter of 2014 and 81% of them happened in Malaysia. This phenomenon was in line with the rapid growth of Malaysia's digital economy [6].

Furthermore, Business Insider reports that Malaysia has become a center of cybercrime mastermind by Nigerian. They have been successfully tricked hundreds American woman into cyber criminals with the average loses is USD250 thousand [5]. Their operation is same by start hacking the internet infrastructure, then broke into Malaysian banking system. This repeating cases have been shows that Malaysia is lacked of resources and expertise to handle these cybercrimes [6], [7]. This situation has been drop Malaysian image and credibleness in combating cybercrime. Despite the fact that cybersecurity in Malaysia has been improved in term of the national policy formulation and the management mechanism of national cyber crisis. However, it can be seen that this strategy are still unsuccessful and need to be enhanced [7], [8]. If this situation continued, it will give a serious impact on political, economic and social sectors in Malaysia.

This cyber romance scams happen on a global scale and there is no international statistical center that stores the victim's data and the exact amount of loss [9]. In contrast, Malaysian have been recorded those important information by the Commercial Crime Investigation Department (CCID), Royal Malaysian Police (PDRM). Based on the statistics recorded by CCID, the increasing in cybercrime cases in Malaysia have gone up at an alarming rate and this situation has inadvertently urge agencies and authorities in Malaysia to analyze cyber romance scam in Malaysia as a whole to seek a preventive measure in reducing cybercrime rate.

However, [9] emphasized that the authorities only received a portion of the report because there are some victims is shame to appear after recognizing themselves being deceived by this syndicated, or some of the victims still do not realize they have been deceived. Nevertheless, cybercrime statistics obtained from the CCID still can be used. Users who like to find a romantic partner or soul mate online is the main target

Sponsored by Ministry of Higher Education using grant AP2017-005/2

of this syndicate. Therefore, in-depth studies need to be conducted to find out more details and pattern about these cybercriminal [8].

On the other hand, [10] have been introduced a theory to understand the crime victimization that called Routine Activity Theory (RAT). This theory suggests three key factors to identify the crime victimization: (i) the presence of a motivated criminal, at the same time and place, (ii) the existence of an opportunity to meet the appropriate target, (iii) the target or the victim has no adequate care. This theory has been successfully applied for a tendency crime prediction such as robbery, theft, vandalism, rape, assault and fraud [11]. In recent years, RAT theory has been tested in cyber criminology and they found very significant correlation between individual tendencies to receive virus in cyberspace and the tendency to become cybercrime victim [11]. While [12] found a correlation between online shopping behavior and the tendency to become a cybercrime victim of financial fraud. Hence, these findings clearly show that RAT theory also can be used as a framework of study involving fraud or crime in cyberspace. Therefore, this study will compared the findings of cyber romance scam victimization analysis using RAT theory and Apriori algorithm. Then proposed a cyber romance scam model.

II. Method

Quantitative method such as questionnaire survey was used in this study to investigate cyber romance scam victimization tendency factors. Furthermore, to identify the reliability of instruments used, a pioneer test was carried out before the actual studies were conducted. There are three methods used in this study which is (i) participants, (ii) materials, and (iii) pioneer test. The highlighted method is explained as below:

A. Participants

The participant of this study is cyber romance scam victims based on the sample survey in Selangor, Malaysia. The total samples are 280 surveys that represent 2508 of target populations. The participants included 42 men and 238 women. Aged below 25 years old is 17, aged between 25-35 is 92, aged between 36-45 is 87, and aged above 46 is 84. The education level of participants included 92 bachelor degree holder, diplomas holder is 104, and STPM holder is 17. For marital status, 84 is single, 28 divorced, and 168 married. Working status included 224 is working, 8 retired, and 48 unemployed. While, monthly salary included 48 participants has no income, 22 participants got salary between RM1-RM2000, 92 participants got salary between RM2001-RM4000, 73 participants got salary between RM4001-RM6000, and 42 participants got salary from RM6001 and above.

B. Materials

The materials used in this study is based on questionnaire. The questionnaire consists of three section and all section is required to be answered by the participant. The first section explains the objective of research and the background information regarding to this study to ensure that the respondent understands the purpose of the study. The contact information such as phone numbers and e-mail address also stated in this section to allow the respondents contacting the researchers if they need any clarification regarding the survey. Then, this section is followed by the respondents' demographics related-questions such as gender, age, education level, marital status, employment status and monthly income. The dichotomous question types in used to able respondents selecting only one answer based on several given answers.

Whereas for second section is consist of question that is used as an instrument in measuring relationship between levels of cybercrime awareness and the tendency to become a victim of cyber romance crime. There are six items used to measure the level of cyber-fraud awareness as shown in table 1. The questions in this section are adapted from [1], which the study is investigating a tendency to become a victim of online scam.

The last section is consisting of question that is used as an instrument in measuring relationship between the level of computer-based skills and the tendency to be a victim of cyber romance crime. There are three items used to measure the level of computer skills such as: (i) Period of using computer (years), (ii) Period of using computer (hours) and (iii) Taking IT courses. The questions in this section are also adapted from the study conducted by [1]. To gather the information, both section which is section two and three is using dichotomous and Likert scale of questionnaires types in range between 1 to 6. The dichotomous method is a straight question and answer type which allows respondents to select only one simple answer. While Likert scare are used to select the degree to which respondents agree to a specific statement.

C. Pioneer Test

Reliability is the level of suitability and the accuracy of the instrument to measure the variable studies. Therefore, reliability analysis is carried out using pioneer test before the actual survey was distributed to ensure that the research findings are consistently based on the selected data collection method.

These tests were conducted in early December 2017 and distributed randomly to the victims who has been lodge a police report regarding cyber romance crime. This pioneer test has been used to ensure the usability and validity of the instrument's content used. Then the reliability of this instrument is measured by the Cronbach Alpha Coefficient.

For the purpose of this pioneer test, survey questions are built online using Google Forms services and links to the questionnaire are sent using WhatsApp application. Within two weeks, 350 surveys were distributed to the victims of cyber romance scams. However, only 300 surveys were answered. 20 of the 300 surveys should not be used. This is due to some of unanswered survey questions by the respondents, so it cannot be applied to this study. Therefore, the number of surveys used in this study is 280 and respondents' response rate is 80 percent. All respondents find that the question in the instrument are easy to understand and the contents were clearly explained. The average time to answer the questionnaire is about 10 minutes.

Then, the Cronbach Alpha Coefficient is computed using SPSS software. All used variables shows the result between 0.83 - 0.92. Where the level of cyber-fraud awareness got a higher Cronbach Alpha values (0.92) against computer skill level (0.83). This result has been confirmed that the instruments used in assessing the level of computer-based skills and the level of cyber-fraud awareness is reliable and consistent.

III. Results and Discussion

A. Preliminary Analysis

In preliminary analysis, normality test was conducted to ensure that the normality of the data recorded in this study is approaching a normal distribution. Normality is the assumption used that involving two or more variables. There are two types of variables in this study namely demographic variables and manipulated variables including the level of computer skills and the level of cyber-fraud awareness. While the responding variable is a tendency to become a victim of cyber romance scams. As shown in table 1, there are six questions used to study the level of cyber-fraud awareness. The mean values for these questions are between 1.157 - 1.891. While three questions are used to study the level of computer skills and the mean or average values for these questions are between 1.231 - 2.752. It can be conclude that the recorded survey data in this study is approaching a normal distribution.

TABLE I. Descriptive Analysis of Variables

Variables: Levels of cyber-fraud awareness	**Min (Average)**
Knowledge level about cyber romance scam	1.157
Awareness levels about cyber romance scam	1.591
Knowing about 419 scam / Nigeria scam 419 / parcel scam	1.825
Awareness levels about 419 scam / Nigeria scam 419 / parcel scam	1.386
Knowing about phishing scams	1.891
Awareness levels about phishing scam	1.758
Variables: Levels of computer skills	**Min (Average)**
Period of using computer (years)	2.752
Period of using computer (hours)	2.517
Taking IT courses	1.231

TABLE II. Pearson Correlation Analysis

Variables	**Tendency to become a victim of cyber romance crime**	**Levels of cyber-fraud awareness**	**Levels of computer skills**
Tendency to become a victim of cyber romance crime	1	-	-
Levels of cyber-fraud awareness	0.626	1	-
Levels of computer skills	0.306	0.403	1

Then, Pearson correlation coefficients is used to evaluate the degree of linear relationship between all variables in this studies. Based on the result in table 2, the correlation between the level of cyber-fraud awareness and the tendency to become a victim of cyber romance scam is 0.626 (r = 0.626, p <0.01). This value indicates a positive relationship between these two variables. On the other hand, a weak positive relationship can be seen among computer skills and the tendency to be a victim of cyber romance scam with coefficient value of 0.306 (r = 0.306, p <0.01).

After that, in order to identify cases (or respondents) that are above or below standard deviation units, the Casewise Diagnostics analysis need to be implemented. Three unusual cases have been identified as shown in Table 3. However, this unusual cases are not affect the analysis result since the number of cases identified is 1% of the total cases and Cook's coefficient value is less than 1. Therefore, this unusual cases will not be removed.

TABLE III. Casewise Diagnostics and Analysis of Cook Distance

Total cases	**Levels of cyber-fraud awareness**	**Expected value**	**Cook coefficient**
91	4.23	1.3001	
110	3.50	1.8972	
			0.446

Lastly, regression analysis was applied. Unlike correlation analysis, regression analysis can assess the strength of causal relationships between manipulated variables and responding variables. Multiple regression coefficients are used to determine the strength of relationships between respondents' variables (the tendency to become a victim of cyber romance crime and manipulated variables (levels of cyber-fraud awareness and computer skills). Multiple regression coefficients, R2 measures the variation of respondents' variables that are likely to become a victims of cyber romance scams that can be statistically explained by manipulated variables i.e. level of cyber-fraud awareness and computer skill level.

TABLE IV. Multiple Regression Coefficients Analysis

Variables	**R2**	**F-values**	**Beta coefficient**	**t**	**P-values**
Tendency to become a victim of cyber romance crime	0.669	31.275			0.000
Levels of cybercrime awareness			0.626	8.730	0.000
Levels of computer skills			0.306	3.493	0.001

Based on table 4, the R2 value which is 0.669 indicates that 66.9 percent of the variance in “Tendency to become a victim of cyber romance crime” can be explained by a regression model. The value of F is 31.276 with a value of p 0.000. This is means that the probability of this decision happens by chance is less than 0.0005. Hence, significant relationships have been shown among the level of cyber- fraud awareness is 8.730 and the t value for computer skill level is 3.493. Again, the probable probability of the result is less than 0.05.

Therefore, based on the results obtained and the interpretations made from the statistical analysis, it can be concluded that four hypotheses have been supported (H1, H2, H5 and H6). The hypothesis summary of their relevant research questions is shown in table 5. Next, the studies were continued to examine how far the demographic variables influencing the tendency to become a victim of cyber romance crime using Apriori based association rules technique.

TABLE V. HYPOTHESIS SUMMARY BASED ON INTERPRETATION OF STATISTICAL ANALYSIS

No	Research Question	Hypothesis	Results
H1	Is age influencing the tendency to become a victim of cyber romance crime?	Older individuals are more likely to be a victims of cyber romance scams.	Supported
H2	Is education level influencing the tendency to become a victim of cyber romance crime?	Uneducated individuals are more likely to be a victims of cyber romance scams.	Supported
H3	Is marital status influencing the tendency to become a victim of cyber romance crime?	Single individuals are more likely to be a victims of cyber romance scams.	Unsupported
H4	Is monthly income influencing the tendency to become a victim of cyber romance crime?	High-income individuals are more likely to be a victims of cyber romance scams.	Unsupported
H5	Is levels of computer skills influencing the tendency to become a victim of cyber romance crime?	Individuals with low computer skills are more likely to be a victims of cyber romance scams.	Supported
H6	Is levels of cybercrime awareness influencing the tendency to become a victim of cyber romance crime?	Individuals who lack of cyber-crime awareness are more likely to be a victims of cyber romance scams.	Supported

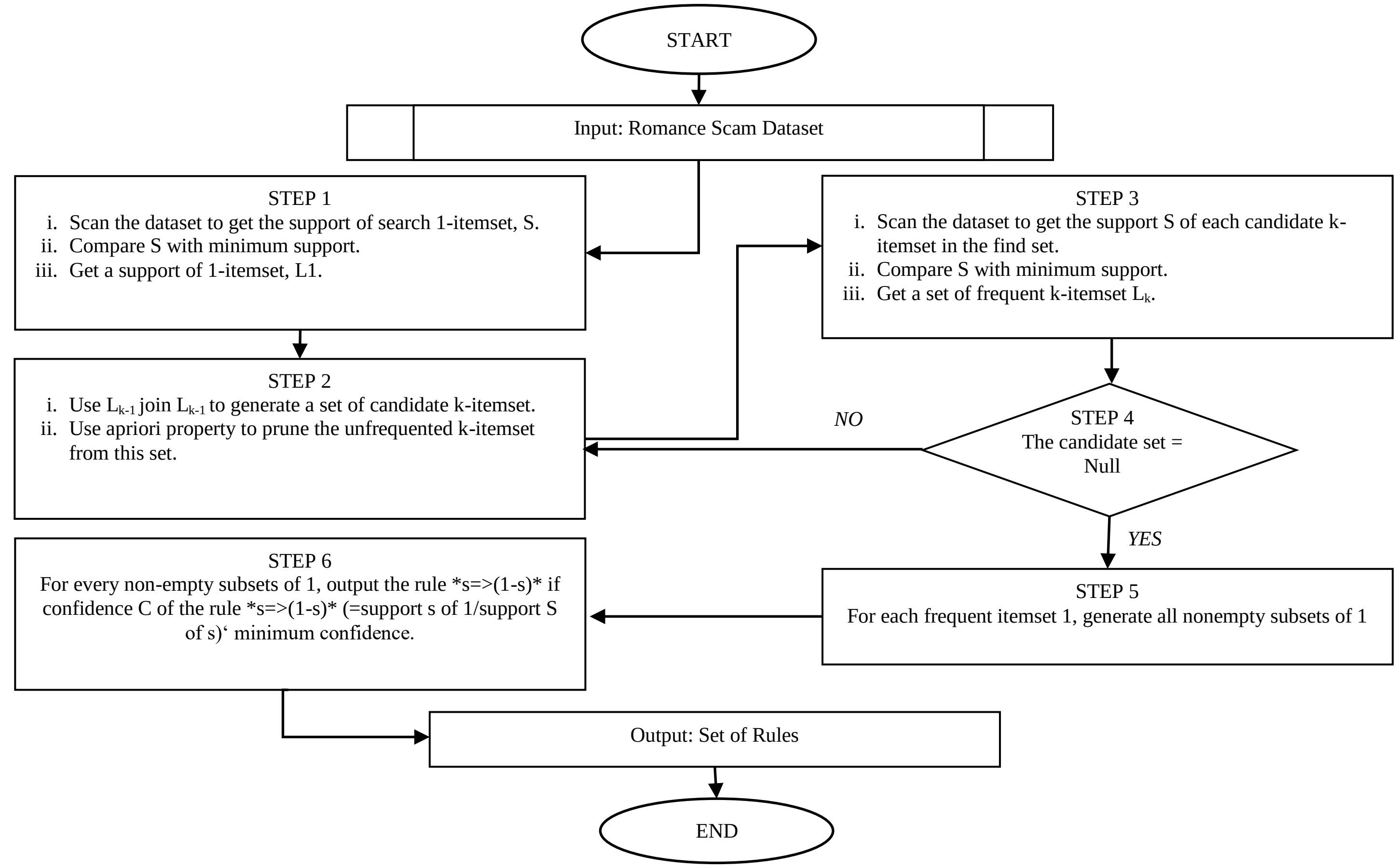

START
Input: Romance Scam Dataset
STEP 1
i. Scan the dataset to get the support of search 1-itemset, S.
ii. Compare S with minimum support.
iii. Get a support of 1-itemset, L1.
STEP 3
i. Scan the dataset to get the support S of each candidate k-itemset in the find set.
ii. Compare S with minimum support.
iii. Get a set of frequent k-itemset L_k.
STEP 2
i. Use L_{k-1} join L_{k-1} to generate a set of candidate k-itemset.
ii. Use apriori property to prune the unfrequented k-itemset from this set.
NO
STEP 4
The candidate set = Null
YES
STEP 6
For every non-empty subsets of 1, output the rule *s=>(1-s)* if confidence C of the rule *s=>(1-s)* (=support s of 1/support S of s)‘ minimum confidence.
STEP 5
For each frequent itemset 1, generate all nonempty subsets of 1
Output: Set of Rules
END

Fig. 1. The Steps of Generating Association Rules using Apriori Algorithms on Romance Scam Dataset.

B. Cyber Love Fraud Pattern Recognition in Malaysia using Apriori based Association Rules Technique

This study will use the Apriori based Association Rules algorithm to get a meaningful information from the data. This algorithm works by generating frequent item set and then generating a set of rules. Association rule learning is a prominent and a well-explored method for determining relations among variables in large databases compared to other methods. In addition, this method is still popular and been used recently to solve the problem in various domain [13]–[15]. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time [16].

Association rule generation is usually split up into two separate steps which is a minimum support threshold is applied to find all frequent item sets in a database; and a minimum confidence constraint is applied to these frequent item sets in order to form rules. The first step needs more attention, while the second step is straightforward [17]. Figure 1 shows the steps of generating association rules using Apriori algorithms that have been used in this study

These experiments were conducted by involving the relationship between seven main attributes/variables namely Races, Gender, Marital Status, Age, Occupation, Type of fraud and Total losses. The data set used is an original record of cyber romance fraud victims in Selangor that representing 2274 cases in 2017. The data set is recorded by the CCID in Bukit Aman. Strong, interesting and authentic association between attributes is generated as a set of rules for the pattern recognition of cyber romance fraud in Malaysia. This experiment was conducted using the Waikato Environment for Knowledge Analysis (WEKA) software.

TABLE VI. RULES SUMMARY FOR ALL SELECTED ATTRIBUTES

Min. Support 0.1			
No.	**Rules**	**Min. Confidence**	**Frequency**
1	**IF** TypeOfFraud=ParcelScam **AND** Occupation=UnEmployed **AND** MaritalStatus=Married **THEN** Gender=Female	0.95	257
2	**IF** TypeOfFraud=ParcelScam **AND** Age=36-45 **AND** MaritalStatus=Married **THEN** Gender=Female	0.86	320
3	**IF** Races=Chinese **AND** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **THEN** Gender=Female	0.86	382
4	**IF** TypeOfFraud=ParcelScam **AND** Age=25-30 **AND** MaritalStatus=Single **THEN** Gender=Female	0.85	249
5	**IF** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **AND** Age=25-30 **THEN** Gender=Female	0.85	337
6	**IF** TypeOfFraud=ParcelScam **AND** Age=25-30 **AND** MaritalStatus=Married **THEN** Gender=Female	0.85	230
7	**IF** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **AND** MaritalStatus=Single **THEN** Gender=Female	0.83	288
8	**IF** Races=Chinese **AND** Occupation=PrivateSector **AND** MaritalStatus=Married **THEN** Gender=Female	0.82	256
9	**IF** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **AND** Age=36-45 **THEN** Gender=Female	0.81	228
10	**IF** Races=Malay **AND** TypeOfFraud=ParcelScam **AND** MaritalStatus=Married **THEN** Gender=Female	0.80	412
11	**IF** Races=Malay **AND** TypeOfFraud=ParcelScam **AND** Age=46< **THEN** Gender=Female	0.80	231
12	**IF** Races=Malay **AND** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **THEN** Gender=Female	0.77	282
13	**IF** TypeOfFraud=ParcelScam **AND** Occupation=PrivateSector **AND** MaritalStatus=Married **THEN** Gender=Female	0.76	431
14	**IF** TypeOfFraud=ParcelScam **AND** Age=46< **AND** MaritalStatus=Married **THEN** Gender=Female	0.75	305
15	**IF** Races=Chinese **AND** Gender=Female **AND** TypeOfFraud=ParcelScam **THEN** Occupation=PrivateSector	0.7	382
16	**IF** Gender=Female **AND** TypeOfFraud=ParcelScam **AND** Occupation=UnEmployed **THEN** MaritalStatus=Married	0.76	257
17	**IF** Gender=Female **AND** TypeOfFraud=ParcelScam **AND Age=36-45 THEN** MaritalStatus=Married	0.74	320
18	**IF** Gender=Female **AND** TypeOfFraud=ParcelScam **AND** Age=46< **THEN** MaritalStatus=Married	0.72	305
19	**IF** Gender=Female **AND** MaritalStatus=Single **AND** Age=25-30 **THEN** TypeOfFraud=ParcelScam	0.88	249
20	**IF** Gender=Female **AND** Age=25-30 **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.88	337
21	**IF** Gender=Female **AND** MaritalStatus=Married **AND** Age=25-30 **THEN** TypeOfFraud=ParcelScam	0.88	230
22	**IF** Gender=Female **AND** MaritalStatus=Single **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.86	288
23	**IF** Gender=Female **AND** MaritalStatus=Married **AND** Occupation=UnEmployed **THEN** TypeOfFraud=ParcelScam	0.84	257
24	**IF** Races=Malay **AND** Gender=Female **AND** MaritalStatus=Married **THEN** TypeOfFraud=ParcelScam	0.83	412
25	**IF** Races=Malay **AND** Gender=Female **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.82	282
26	**IF** Races=Malay **AND** Gender=Female **AND** Age=46< **THEN** TypeOfFraud=ParcelScam	0.81	231
27	**IF** Races=Chinese **AND** Gender=Female **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.81	382
28	**IF** Gender=Female **AND** MaritalStatus=Married **AND** Age=36-45 **THEN** TypeOfFraud=ParcelScam	0.81	320
29	**IF** Gender=Female **AND** MaritalStatus=Married **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.81	431
30	**IF** Gender=Female **AND** Age=36-45 **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.78	228
31	**IF** Gender=Female **AND** MaritalStatus=Married **AND** Age=46< **THEN** TypeOfFraud=ParcelScam	0.77	305
32	**IF** Races=Chinese **AND** Gender=Female **AND** MaritalStatus=Married **THEN** TypeOfFraud=ParcelScam	0.77	294
33	**IF** Races=Chinese **AND** MaritalStatus=Married **AND** Occupation=PrivateSector **THEN** TypeOfFraud=ParcelScam	0.76	292

In this study, rule generation is controlled by parameter setting, such as minimum support level and minimum confidence level. Determining the value of support and minimum confidence levels is a complex task as it affects the quality of the generated rules. Normally, rules were generated with high confidence value and it is a top priority for rules selection because they are considered strong, but this method does not provide an opportunity for odd cases.

Therefore, this study sets the parameter of confidence value from 0.3 to 1.0 and the minimum support value is set as 0.1. This is to ensure that the rules with frequent and meaningful attributes at low confidence and support values can be generated and discovered. In this section, the results of the study were broken down into seven experiments where each attribute was given the opportunity to become a class attribute on the data record. Then, meaningful rules will be selected through two criteria:

- Confidence and support values are greater than other.
- High support value and low confidence value, but rules are generated earlier that the other rules.

Through the study, Apriori's based Association Rules algorithm has been selected based on its effectiveness in producing interesting rules. The choice of meaningful rules is obtained through the process of repeated analysis. Table 6 shows the selected rules which minimum confidence is above 0.7 and minimum support is 0.1.

Overall, experimental results have shown the uniform pattern for the association between all selected attributes despite different attributes class. This shows that the data used is able to produce interesting and stable pattern. Therefore, based on the rules obtained and the interpretations made from several analyses, it can be concluded that Chinese and Malay women were likely easier to become a victim. This may be due to the large ratio of Chinese and Malay population in Malaysia. Those between the ages of 25 and 45 years were likely to be the victims of cyber romance scams. In addition, this research shows that married people are more likely to become a victim of cyber romance scams. This is contrary to the study's hypothesis that single individuals are more likely to become a victim. Unemployed person also can be a victim of cyber romance scam probably because they have a lot of time to go online. Lastly, the scammer usually will ask the money from the victim around RM3025 until RM5490. The model of cyber romance scam based on extracted rules can be illustrated as figure 2.

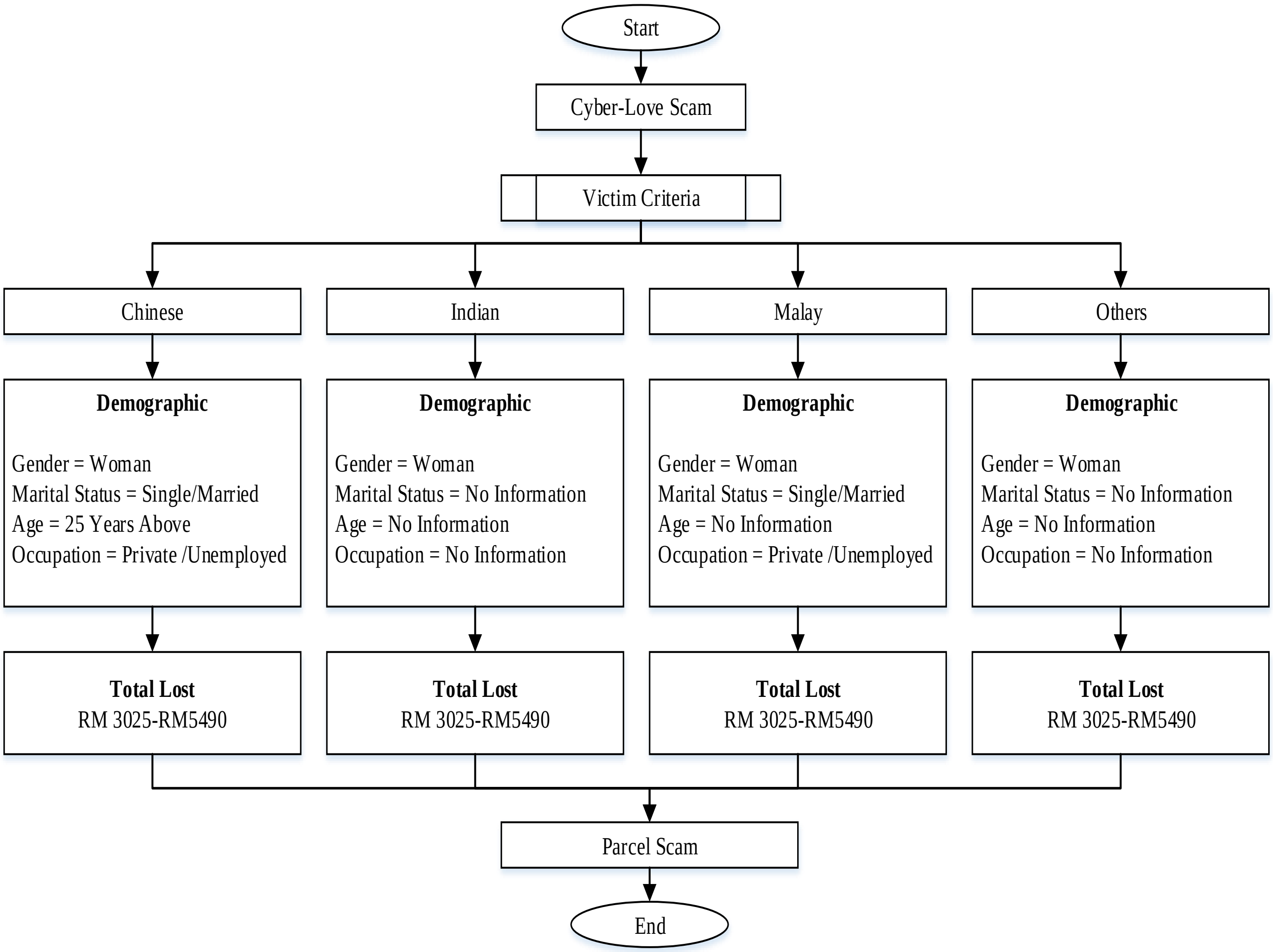


Fig. 2. Cyber Romance Scam Models using Apriori Techniques.

IV. Conclusion

The main objective of this study is to analyze the factors that influence the susceptibility to become a cyber romance scam victim in Malaysia using the information from the individuals who have been victims of cyber romance scams and lodged the police reports in the state of Selangor, Malaysia. The lack of research in this area as well as the growing case of cyber romance scam in Malaysia yearly has been a driving force in publishing this research. RAT Theory is used as a theoretical basis and this study will help to improve understanding of this theory in the context of cyber crime because this theory is more synonym and often used in the context of ordinary street crime. The framework of this study examines demographic variables such as Age, Education Level, Marital Status, Monthly Income, and manipulated variables such as level of computer skills and the level of cyber crime awareness. Both of these manipulated variables are hypothesized to have a relationship with the tendency to become a victim of cyber romance scam in Malaysia.

Overall, the study found that those between the ages of 25 and 45 years were likely to be victims of cyber romance scams in Malaysia. The majority of the victims are educated and having a Diploma, the remaining have a degree, STPM and SPM / SPMV. In addition, this research shows that both single and married people can become victims of cyber romance scams. Furthermore, studies show that individuals who are not earning are also vulnerable to being a victims because the study shows that 17 percent of respondents who are victims are from this group. However, as expected, those who work and have monthly income between RM2001 and above are more likely to be targeted and victims of cyber romance scams. The study also shows that those who lack computer skills and less awareness of cyber-fraud are more likely to be victims of cyber romance scams. The findings of this study are useful for policy makers and enforcement agencies to protect Internet users in Malaysia. Based on the analysis, it can be concluded that this research has found strong characteristics of cyber romance scam victimization in Malaysia.

Furthermore, this study is based on RAT theory, so this study confirms the fundamental principle of this theory and the usability of the theory in cyberspace environment. This theory is not limited to ordinary street crime but is relevant for application in cyber-crime. Indirectly, this study contributes to an increased understanding of RAT Theory and its usability in different contexts.

For future work, the researchers should investigate the extent of financial loss suffered by the victim and take the environment, motives or major motivation of the cyber criminals in choosing their target into account. Overall, the findings of this study may useful for policy makers in creating internet-related policies to protect the Internet users in Malaysia.

Acknowledgment

This research was funded by Ministry of Higher Education using grant AP2017-005/2.

References

[1] E. V. Garrett, “Exploring internet users’ vulnerability to online dating fraud: Analysis of routine activities theory factors,” 2014.

[2] A. Salman, S. Saad, and M. N. Shahizan Ali, “Dealing with ethical issues among internet users: Do we need legal enforcement?,” Asian Soc. Sci., 2013.

[3] C. Kopp, R. Layton, J. Sillitoe, and I. Gondal, “The role of love stories in Romance Scams: A qualitative analysis of fraudulent profiles,” Int. J. Cyber Criminol., 2016.

[4] N. A. Manap, A. A. Rahim, and H. Taji, “Cyberspace Identity Theft: The Conceptual Framework,” Mediterr. J. Soc. Sci. , 2015.

[5] M. T. Whitty, “Anatomy of the online dating romance scam,” Secur. J., vol. 28, no. 4, pp. 443–455, 2015.

[6] James Lyne, “Cybersecurity in 2015,” sophos, 2015.

[7] M. Riek, R. Böhme, and T. Moore, “Measuring the Influence of Perceived Cybercrime Risk on Online Service Avoidance,” IEEE Trans. Dependable Secur. Comput., 2016.

[8] M. A. Bin Pitchan, W. A. W. Mahmud, S. N. Sannusi, and A. Salman, “Control and freedom of the Internet: Challenges faced by the government,” J. Asian Pacific Commun., vol. 25, no. 2, pp. 243–252, 2015.

[9] C. Barclay, “Using Frugal Innovations to Support Cybercrime Legislations in Small Developing States: Introducing the Cyber-Legislation Development and Implementation Process Model (CyberLeg-DPM),” Inf. Technol. Dev., 2014.

[10] L. E. Cohen and M. Felson, “Social Change and Crime Rate Trends: A Routine Activity Approach,” Am. Sociol. Rev., 1979.

[11] F. Ngo and R. Paternoster, “Cybercrime Victimization: An examination of Individual and Situational level factors.,” Int. J. Cyber, 2011.

[12] J. van Wilsem, “Worlds tied together? online and non-domestic routine activities and their impact on digital and traditional threat victimization,” Eur. J. Criminol., 2011.

[13] P. Yuan, D. Chen, T. Wang, S. Cao, Y. Cai, and L. Xue, “A compensation method based on extreme learning machine to enhance absolute position accuracy for aviation drilling robot,” Adv. Mech. Eng., vol. 10, no. 3, p. 1687814018763411, 2018.

[14] P. Pravallika and K. Narendra, “Analysis on Medical Data sets using Apriori Algorithm Based on Association Rules,” 2018.

[15] R. Wadhawan, “Prediction of Coronary Heart Disease Using Apriori algorithm with Data Mining Classification,” Int. J. Res. Sci. Technol., vol. 3, no. 1, pp. 1–15, 2018.

[16] R. A. A. Rashid, P. N. E. Nohuddin, and Z. Zainol, “Association Rule Mining Using Time Series Data for Malaysia Climate Variability Prediction,” in International Visual Informatics Conference, 2017, pp. 120–130.

[17] Z. A. Othman, N. Ismail, and M. T. Latif, “Association pattern of NO 2 and NMHC towards high ozone concentration in klang,” in Electrical Engineering and Informatics (ICEEI), 2017 6th International Conference on, 2017, pp. 1–6.

Deep Learning-Based Model Architecture for Time-Frequency Images Analysis

Haya Alaskar
Computer Science Department
Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia

***Abstract*—Time-frequency analysis is an initial step in the design of invariant representations for any type of time series signals. Time-frequency analysis has been studied and developed widely for decades, but accurate analysis using deep learning neural networks has only been presented in the last few years. In this paper, a comprehensive survey of deep learning neural network architectures for time-frequency analysis is presented and compares the networks with previous approaches to time-frequency analysis based on feature extraction and other machine learning algorithms. The results highlight the improvements achieved by deep learning networks, critically review the application of deep learning for time-frequency analysis and provide a holistic overview of current works in the literature. Finally, this work facilitates discussions regarding research opportunities with deep learning algorithms in future researches.**

Keywords—Convolutional neural network; time-frequency; spectrogram; scalograms; Hilbert-Huang transform; deep learning; sound signals; biomedical signals

I. Introduction

Time-frequency analysis has been considered for pattern recognition and fault diagnosis. It is usually known as an initial step for signal preprocessing. It provides a suitable tool for analyzing signals in many fields of engineering, biomedicine, finance and speech [1]–[7]. Recently, the importance of discovering powerful signal processing tools has become essential to the analysis of signals. The first time-frequency representation was addressed in the early development of quantum mechanics by H. Weyl, E. Wigner, and J. von Neumann in approximately 1930 [8]. Since then, there have been numerous implementations of time-frequency representation to address signal processing. The early time-frequency analysis system was based on handcrafted techniques. These systems were followed by time-frequency analysis systems based on feature-extraction and machine learning [9]–[11]. Unfortunately, according to a scientist's point of view, preprocessing and feature extraction in any time series signal is not an easy task. There are a number of feature sets that can be extracted from time-frequency domains. Determining the ideal features from such domains requires time for examination and investigation [4], [12]. Furthermore, identify a particular pattern or s of a time-frequency representation is usually unknown [13]. Therefore, effective and reliable tools need to be considered to solve the task. In recent years, studies have been performed to find alternative tools to analyze and identify the pattern directly from a time-frequency image. Starting with the [14] paper, they extracted time-frequency images from the sound signals and used them as input to a deep learning network architecture for classification. Since then, various deep learning network architectures have been proposed, typically based on some form of convolutional neural network (CNN) [10], [11], [13], [15], [16]. In these studies, the CNNs obtain better results than traditional machine learning. Such approaches are attractive since they typically do not need domain knowledge expertise. In fact, CNNs rival human accuracies for the same tasks [17], [18].

Recently, deep learning has proved to be successful in all areas of science, such as successes in image recognition [19], handwriting, manufacturing [13], disease diagnosis [15], [20], [21] and speech processing [22]. The results of these studies have proven the benefits of CNNs in image and signal analysis, which emphasize that CNNs have the capability of addressing diagnosis and classification tasks. Therefore, in the literature, deep learning networks have received considerable attention from researchers; especially, the convolutional neural network. CNNs are able to address data directly without requiring complex preprocessing steps. CNN models are advantageous because of their high levels of expert information processing and can propose much more effective models for complex high dimensional datasets. Therefore, it is important to highlight recent advances techniques of time-frequency analysis, especially recent deep learning architectures, which have outperformed state-of-the-art approaches.

This paper introduces a comprehensive survey of current applications to train a deep learning network in the time-frequency domain in order to classify or diagnose patterns. It will contrast these techniques and compare them among the traditional machine learning applications. To the best of knowledge, this is the first survey that focuses on the use of deep learning with time-frequency analysis and compares it to previous feature-based systems.

The main aim of this survey is two-fold. First, it documents the background knowledge about how the time-frequency domain has been used to address signal processing in the past few years.

Second, it critically reviews the application of deep learning with the time-frequency domain and offers a general overview of the existing literature. In the process of achieving these aims of the paper, the following research questions should be addressed

- Can deep learning be used to classify time-frequency representations of signals?
- Does the deep learning network alter the results of a time-frequency analysis?
- If so, which time-frequency representation of the signal yields the best results?

First, a discussion of the time-frequency representation types and the challenges raised for analyzing the time-frequency domain are presented in section 2. A brief discerption of deep learning networks, especially the CNN, is introduced in section 3. Then, the selection criterion and methodology for selecting which systems to review are explained in section 4. A literature review is highlighted in section 4, and a brief discussion is addressed in section 5.

II. Background

A. Time-Frequency

The time-frequency approach can provide suitable outputs for the discovery of complex, high-dimensional and nonstationary properties. Time-frequency characterization simultaneously represents a signal in both the time and frequency domain. The most popular visual representations of the time-frequency domain are spectrograms and scalograms. This type of representation methods are able to extract particular patterns, for example, the professional extraction of sensitive fault patterns [1]. In medical applications, this type of representation can help to identify an abnormal pattern in biomedical signals. Their success is reported in a number of applications [2]–[7]. Time-frequency methods were also integrated with other advanced algorithms, such as neural networks [5] and support vector machines [8]. In the next sections, a brief introduction is provided about the three types of time-frequency representations.

1) Spectrograms:- a spectrogram is generated using the short-time Fourier transform (STFT). The axis on STFT shows time and frequency, and the color scale of the spectrogram image represents the amplitude of the frequency. The basis for the STFT representation is known as a series of sinusoids.

2) Scalograms:- scalograms are a generated by using the wavelet transform (WT). WTs are a linear time-frequency representation. The basis for the WT representation is a wavelet basis function, which depends on the frequency resolution. The signal is decomposed with different resolutions at different time and frequency scales by scaling and translating the wavelet function.

There are many wavelets types such as the Gaussian, Morlet, Shannon, Meyer, Laplace, Hermit, or the Mexican Hat wavelets. There are differences between each type in both simple and complex functions. There have been many studies to address the effectiveness of each wavelet type. Currently, there is not a clear technique for finding the most suitable wavelet.

3) Hilbert-Huang transform:- the Hilbert-Huang transform (HHT) is considered an adaptive nonparametric representation. It is different from the previous methods such as STFT and WT, which are based on set of basic functions. In contrast, HHT does not need to make assumptions on the basis of the data. It just uses the empirical mode decomposition (EMD) to decompose the signal into a set of elemental signals named intrinsic mode functions (IMFs). The HHT methodology is depicted in Figure 3.

The HHT involves two steps, namely, EMD of the time series signal and the Hilbert spectrum construction. HHTs are particularly useful for localizing the properties of arbitrary signals. For more explanation, see [9].

The HHT does not divide the signal at fixed frequency components, but the frequency of the different components (IMFs) adapts to the signal. Therefore, there is no reduction of the frequency resolution by dividing the data into sections, which gives HHT a higher time-frequency resolution than spectrograms and scalograms.

B. Challenges of Analyzing Time-Frequency Domain

Despite numerous applications using time-frequency representations, analyzing signals have some limitations [10]. Signals usually suffer from several causes of extensive noise, including recording devices, power interference and baseline drift [11]. Hence, the analysis of these signals requires addressing noise and filtering signals.

On the other hand, the features extracted from time-frequency representations need appropriate techniques. Some features can be insufficient to describe the time-frequency domain and will lead to in information loss. In fact, feature selection and extraction expressively need expert knowledge. Furthermore, analyzing time-frequency images to detect features or patterns cannot be accomplished by examining images one by one [1]. Actually, it is very unrealistic to identify a large number of time-frequency images by manual methods. To intelligently and automatically identify the features from many time-frequency images, the prevalent deep learning networks show professional serviceability.

Deep learning is a promising technique for large-scale data analytics[12]. In the literature, they have been used in biomedical signal analysis such as EEG [13], ECG [11], [14]–[16] and EMG [17], [18].

Deep learning networks achieved remarkable result compared with the traditional hand-crafted features. Moreover, once a large size of datasets is available, CNNs are a good method and usually beat human agreement rates. The appearance of deep learning networks has made the analysis of the signals simpler than before.

III. Deep Learning Network (DNN)

DNN is a branch of machine learning tools that has shown significant success in various fields in medicine, business, industry sectors, etc. It attempts to model data hierarchically and classifies patterns using multiple nonlinear processing layers. There are several variants of deep learning such as autoencoders, deep belief networks, deep Boltzmann machines, convolutional neural networks and recurrent neural networks. Since current works have established the success of

CNN deep learning models in the application of time frequency analysis, the concentration of this paper is limited to reviewing the past literatures related to CNN models.

A. Convolutional Neural Network (CNN)

The most successful model of DNN is convolutional neural networks (CNNs). Despite, the CNN was first designated by LeCun et al. in 1998[39]. The golden age of deep learning revolution started when Krizhenvsky et al. [19] won the ImageNet competition by a considerable margin. Since then, only convolutional neural networks have won this ImageNet competition [20], [21].

The differentiation between CNN and the simple multilayer network (MLP) is that MLPs only use input and output layers, and, at most, a single hidden layer, whereas in the DNNs there are a number of layers, including input and output layers [22]. Fig. 1 shows the difference between a simple MLP and a CNN. Each block in the CNN model holds a number of layers.

The CNN contains one or more convolutional and max pooling layers followed by one or more fully connected layers, which perform as the classification layer. Different CNNs employ various algorithms in the convolution layer and subsample layer and different network structures. Finally, the fully connected layers are at the end of the network. In the fully connected layer, weights are no longer shared with the conventional layer. These layers are similar to MLPs, where in the final layer, a SoftMax function is used to generate a distribution over classes.

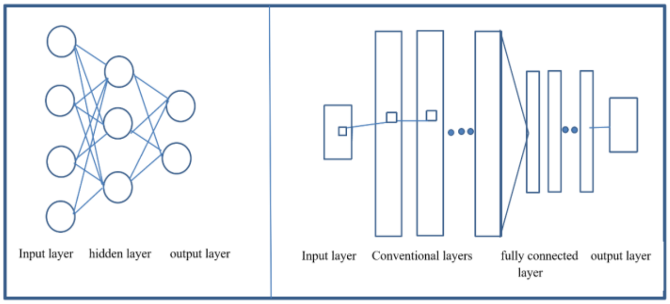


Fig. 1. The Differences in Architecture between a Simple MLP and a CNN.

The significant features of CNNs are that the tasks of preprocessing and feature extraction are not essential in CNNs. In contrast, CNN can automatically identify more complex features because of the number of conventional layers it contains. Furthermore, they are self-learned networks without the need for supervision [35]. This function of DNNs supports the ability of the network to handle large, high-dimensional data that contain a large number of features [36]. This is a beneficial feature of CNNs that reduces the liability during training and helps to select the best features that discriminate classes in the dataset.

IV. Methodology

A. Search Strategy and Selection Process

A database search through online databases such as Google, Google Scholar, and IEEE Explore were used as recommended by [23]. In addition, online databases such as Elsevier, ScienceDirect and ACM, which are the most popular sources for finding scientific papers, were searched. The query terms included time-frequency, DNN based on time-frequency analysis, DNN in signals or time series classification and analysis, etc. also articles that implemented these systems for different languages or domains are included. In total, 154 articles were reviewed and 83 articles were selected for the survey.

B. Literature Sources

The investigation of the applications of DNN with the time-frequency domain was addressed, and articles published in the domain were analyzed.

Most of the selected articles were collected from the publishers, as presented in Table 1, so that the integrity of this review paper is not compromised. However, there is an extensive variety of other sources that are also suitable for this survey.

TABLE I. The Main Sources of the Selected Articles

Publisher
IEEE
Elsevier
bioRxiv
Bioengineering
Springer
Hindawi

C. Data Collection Process

The data collection process involved extensive research of papers that addressed the applications of DNN with time-frequency analysis. These papers were downloaded and studied for collecting suitable information on the subject. The type of results in this paper are qualitative, and the main motivation is to provide a survey of the applications of DNNs and attempt to answer the research questions listed in the introduction section. Overall, the data collection process comprised three main phases

Phase 1: Searching for papers in reliable journals. This phase was completed using some keywords.

Phase 2: Papers are selected and categorized in order to serve the aim of the survey. Then, the qualified papers are examined critically.

Phase 3: Qualitative data were collected and notes were taken to briefly present the data in the results section of this paper. Data were gathered regarding the type of time-frequency domain methods employed.

V. Literature Review

The extensive investigation of the application of DNNs with time-frequency images showed that most of the papers and studies were published after 2016, as represented in Table 2. Most of the papers used the conventional neural network to address this type of image. The next three sections will briefly introduce the applications on DNNs.

TABLE II. THE PUBLISHED PAPERS ON THE APPLICATION OF CNNS ON EEG, ECG AND EMG

References	***Signals and types of representation***	***Deep learning***	***Result***
[24]	EEG, spectrogram	CNN	74% accuracy
[24]	EEG , scalogram	CNN	F1-score 81%
[25]	EEG spectrogram	CNN	96% accuracy
[26]	ECG, spectrogram	DNN	97.5% accuracy
[27]	ECG, Spectrogram	16 CNN	90% sensitivity
[28]	EEG, spectrogram	VGG15	89% accuracy
[29]	facial videos, Spectrogram	VGG15	RMSE was 4.27
[30]	Gait signals, scalogram	CNN	97.06% accuracy
[31]	EMG, spectrograms	CNN	69.23 % accuracy
[32]	ECG, Spectrogram and Hilbert spectrum	CNN	98.3% accuracy
[33]	EEG, spectrogram	CNN	96.16% accuracy
[34]	PPG scalogram	GoogLeNet	92.55% F1
[35]	PCG scalogram	VGG16+SVM	56.2% MAcc
[18]	EMG spectrum	RCNN	90.6 % in R2
[36]	EEG spectrograms	CNN	80% accuracy
[35]	PCG, scalogram image	VGG16	56.2 % accuracy
[37]	EEG, EOG, EMG Spectrogram, scalogram	CNN	95% accuracy
[38]	Sound log-mel spectrogram	CNN	EER was 2.7%
[39]	Sound spectrogram	CNN	71% accuracy
[40]	Sound, spectrograms	VGG	85.36 accuracy
[41]	Sound, spectrograms, scalogram	CNN	74.66 % accuracy
[42]	Sound, spectrogram	CNN	AUC is 0.970
[43]	Fault diagnosis Scalogram	CNN	96% accuracy
[44]	Fault diagnosis, spectrograms	CNN	98%-99%
[45]	Fault diagnosis spectrograms	DNN	95.68% accuracy
[46]	Fault diagnosis Spectrogram	CNN	93.61 % accuracy
[47]	Fault diagnosis Spectrogram, scalogram and Hilbert-Huang.	CNN	81.4%, 99.7% and 75.7% respectively.
[10]	Fault diagnosis, scalogram	PSPP with CNN	99.11% accuracy
[1]	Fault diagnosis Spectrogram	DCNN	96.78% accuracy

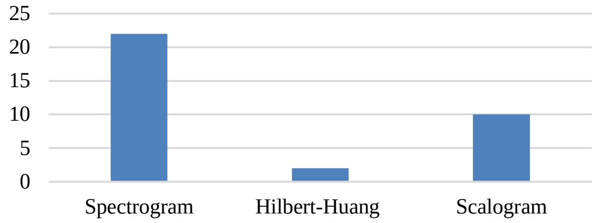


Fig. 2. The Number of Papers used (Spectrogram, Hilbert-Hyang and Scalogram) Type.

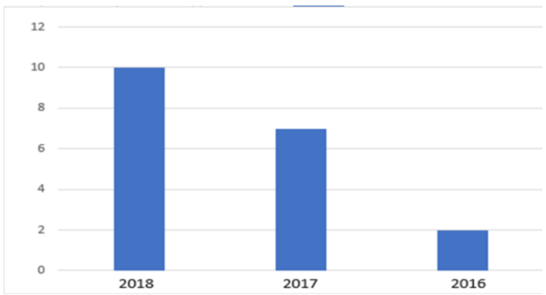


Fig. 3. The Numbers of Papers used Applied the Deep Learning with Time-Frequency Domain on Medical Signals.

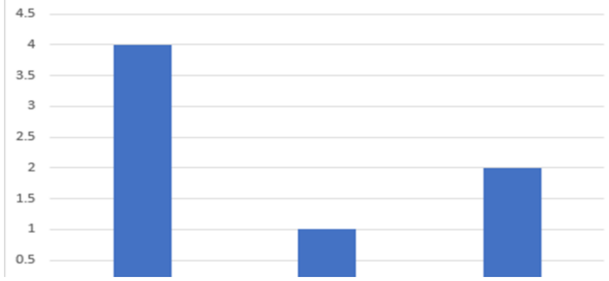


Fig. 4. The Numbers of Papers used Applied the Dnns with Time-Frequency Domain For Fault Diagnosis.

From Fig. 2, it can be noticed that spectrogram has been considered in numbers of studies compared with other type of time-frequency methods. In term of the years of publications. From Fig. 3 and 4, it can be observed that, from 2016 until 2018, considerable effort was undertaken to study and embed the conventional neural network into approaches using these types of data. From the analyzing of different articles , VGG has been selected five times from 31 articles where GoogLeNEt was used only in two papers.

A. Application of CNNs for Fault Diagnosis

Vibration signals are extensively used to diagnose rotating machinery. Researchers attempted to develop automatic and intelligent fault diagnosis tools based on CNN. They extracted the time-frequency representation of vibration signals and fed them directly into a CNN to classify the different kinds of fault features of the rotating machinery. For example, Wang et al. [52] investigate the use of scalogram images as an input to a CNN to predict faults in a set of vibrational data. They used a series of 32 × 32 scalogram images. The highest result they achieved was 96% accuracy.

Lee et al. [53] explored corrupted signals with noise by using a CNN. A short-time Fourier transform was used to generate images from The MFPT data and the Case Western dataset. The trained CNN was able to detect patterns in signals with 98% and 99%.

Janssens et al. [55] incorporated shallow CNNs with the amplitudes of the discrete Fourier transform vector of the raw signal as an input. Pooling, or subsampling, layers were not used. Liu et al.[54] used spectrograms as input vectors into sparse and stacked autoencoders. They attempted to recognize the faults from the normal, inner-race fault, outer-race fault, and rolling bearing parts of fault bearings. The experimental result obtained a good recognition performance on four fault modes with 95.68% accuracy.

Verstraete et al. [18] used a CNN based on time-frequency image analysis for fault diagnosis of rolling element bearings. The CNN consisted of two consecutive convolutional layers without a pooling layer between them. For CNN image inputs, three types of time-frequency transformations are used: short-time Fourier transform spectrogram and wavelet transform (WT) scalogram and Hilbert-Huang transformation (HHT). Their accuracy was 81.4%, 99.7% and 75.7% respectively.

Other study [25] used the Morlet wavelet method to discompose vibration signals of rotating machinery. They used the Pythagorean spatial pyramid pooling (PSPP) layers in the front of the CNN. Hence, the features extracted by the PSPP layer were passed into the convolutional layers for more feature extraction. The evaluation of this model was carried out on two datasets of constant rotating speed signals and variable rotating speed signals. The experiment showed that PSPP CNN was able to achieve 99.11% accuracy.

Another more recent approach in the same manner was proposed in [13]. Xin et al., developed a new CNN to detect different kinds of fault features from the time-frequency representation. The vibration signals were collected from bearings and gears. While the gearbox datasets contain different kinds of faults under the operating conditions, the bearing signals datasets have different fault locations and diameters under several working loads. Those signals are separated into several segments and the time-frequency images are generated by using STFT. These images are treated by the sparse autoencoder method with a linear decoding to expand the sparsity. The proposed DCNN achieved the highest accuracy, with 96.78% compared with the CNN at 89.72% and the LSSVM at 78.33% [13].

B. *Application of CNNs for Sound Signals*

CNN implementations are becoming more common models in the ASC research domain, where Weiping et al., [50] attempted to use the DCNN for the acoustic scene classification. A CNN model is presented which is similar to the VGG style. They use two types of spectrograms; the first was a generated STFT from raw audio frames, and the second was a CQT spectrogram. The highest result achieved by using the STFT spectrograms images was 0.8536, and the one using the CQT spectrograms images was 0.8052. Weiping et al. conclude that the performance of the CNN could be improved by fine tuning the parameters, normalizing the spectrograms in the training of the DCNN and utilizing the temporal feature.

To better describe sounds that are quite different from speech, Espi et al., [49] used high resolution spectrogram images. These images were directly used as input to a CNN.

However, Thomas et al. [14] used the log-mel spectrogram with its delta and acceleration coefficients to train a CNN. The CNN was evaluated in terms of the SAD accuracy on noisy radio recorded by the Linguistic Data Consortium (LDC) for the DARPA RATS program. Most of the RATS data gained by retransmitting existing audio collections, such as the DARPA EARS Levantine/English Fisher communication telephone speech (CTS) corpus, are broadcast over eight radio channels. In addition, telephone recordings in Arabic Levantine, Pashto and Urdu provided an extensive variety of radio channel broadcast effects.

Other studies conducted to address the efficiency of fusing the mel-scaled short-time Fourier transform spectrogram to train a CNN in [18] determined that using a CNN with the log-mel filter bank energy extracted from the mel-scaled STFT spectrogram outperformed other classifiers. The conclusion of this result was that the log-mel filter bank energy feature possesses fewer coefficients per frame compared to the linear-scaled STFT spectrogram and mel-scaled STFT spectrogram, resulting in a decreased requirement of the parameters of the CNN architecture. In [16], it was asserted that representing audio as images using mel-scaled STFT spectrograms achieved better performance than that achieved with linear-scaled STFT spectrograms, the constant-Q transform (CQT) spectrogram and the continuous wavelet transform scalogram when used as inputs to CNNs for audio classification tasks. The dataset was the ESC-50 dataset, which contains 2000 short (5 second) environmental recordings divided equally into 50 classes. Classes were extracted from five groups, namely, human nonspeech sounds, animals, natural soundscapes and water sounds, exterior/urban noises and interior/domestic sounds. Four frequency-time representations were extracted, namely, linear-scaled STFT spectrogram, Mel-scaled STFT spectrogram, CQT spectrogram, CWT scalogram and MFCC spectrogram. The highest result was obtained by using the mel-scaled STFT spectrogram images, achieving 74.66±3.39 accuracy.

Another novel approach for sound classification of free-flying mosquitoes was proposed by [51]. Their motivation was to detect the existence of a mosquito from its sound signature. A CNN was trained on a wavelet spectrogram. They showed that the CNN performance was better than traditional machine learning classifiers. The result of the ROC analysis was 0.970. The authors concluded that the CNN result was remarkable when compared with traditional feature extraction methods.

C. *Application of CNNs for Biomedical Signals*

CNN approaches with time-frequency analysis have also been utilized for medical applications. They were employed to serve as decision makers to detect abnormalities in biomedical signals. For example, Hsu et al., [42] used spectrogram images to train a CNN for heart rate estimation based on facial videos.

they have used the GG15 CNN. They claimed that their approach was a novel work that used a DNN network on real-time pulse estimation. They developed a pulse database, named the pulse from face (PFF), and used it to train the CNN.

In [40], spectrogram images were employed to train a CNN for automatic AF detection. The 16-layer CNN was used and achieved 82% accuracy. The CNN recognized normal rhythm, AF and other rhythms with an accuracy of 90%, 82% and 75%, respectively. The conversion of ECG signals to time-frequency images has improved the CNN’s ability to automatically perform ECG signal classification, and further, it can also possibly aid robust patient diagnosis.

In this study [39], the time-frequency representations for the heartbeat signal was obtained by using an adapted frequency slice wavelet transform (MFSWT). Features were automatically extracted by the stacked denoising autoencoder (SDA) from the time-frequency image. The DNN classifier was used to identify different pattern on heartbeats. The experiments were applied on the MIT-BIH arrhythmia database. The proposed method gained an accuracy of 97.5%.

Other study [46] investigated if CNNs are able to provide better performance for hypertension risk stratification compared with the traditional signal processing methods. Liang et al., used photoplethysmography (PPG) signals for this investigation. The signals were treated by the continuous wavelet transform via the Morse method to create scalogram images. These images were used to train a pretrained GoogLeNet. The signals included 121 samples from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) Database, and each had arterial blood pressure (ABP) and photoplethysmography (PPG) signals. The classification will be based on blood pressure levels which were normotension (NT), prehypertension (PHT), and hypertension (HT) classes. The experiment was run for the following three trials: NT vs. PHT, NT vs. HT, and (NT + PHT) vs. HT. For the purpose of fitting GoogLeNet, each subject signal was divided into 24 five-second windows. Therefore, 2904 scalogram images were extracted from 2904 signal segments. The F-score obtained to classify NT vs PHT was 80.52%, whereas the approach achieved 92.55% for classifying NT vs HT. The results showed that using a pretrained CNN with scalogram images achieved higher accuracy than that achieved with traditional feature extraction methods.

In [47], the authors examined the ability to train the pretrained VGG16 with scalogram images to classify phonocardiogram (PCG) signals for normal/ abnormal heart sounds. First, the PCG files are segmented into chunks of equal length. Scalogram images are generated using the Morse wavelet transformation. The experimental results showed that the CNN model achieved the highest accuracy at 56.2%, whereas the traditional feature processing with a support vector machine achieved 46.9% accuracy. In total, 3240 PCG signals were collected from 947 pathological patients and healthy subjects.

Gurve and Krishnan [56] employed a CNN on EEG data for classification of the eye state. The spectrogram of the EEG signal is created and fed into a CNN with the NMF features. The implementation of this approach has achieved a good result of 96.16% compared to existing methods for eye state detection.

Eltvik [15] has also applied CNN to analyze the time-frequency domain from EEG signals. He used three types of time-frequency domains. The evaluation of this method involved testing it on two different datasets. The first was an artificial dataset created by simulating a nonstationary and noisy method. The second dataset was real EEG signals made available through the BCI Competition III. It was composed 1,400 EEG signals involving a duration of 3.5 seconds, where each subject was asked to imagine movement in either the right hand or in the left foot. The main task is to identify if the subject was imagining during the experiment. Four different CNN architectures were evaluated using k-fold cross-validation with each of the three representations. The resulting spectrogram and Hilbert spectrum representation of the synthetic data achieved accuracies of 98.3% and 88.19%, respectively. In contrast, the scalogram representation obtained a very poor result of 59.29%. In the real data case, the highest accuracy achieved when classifying the EEG spectrograms was 72.50%. For Hilbert spectra, it was 58.00%, and for scalograms, it was 55.93%.

Ruffini et al. [44] explain how to use a CNN for the REM sleep behavior disorder (RBD) prognosis and diagnosis from an EEG. The EEG data were recorded from 121 idiopathic RBD patients and 91 healthy controls. The signals were taken after a few minutes of being in an eyes-closed resting state. After 2 to 4 years of EEG collecting, 19 of these patients were found to develop Parkinson disease PD and 12 of them had dementia with Lewy bodies, whereas the rest remained idiopathic RBD. Ruffin et al. used a CNN trained with stacked multichannel spectrograms. The performance of a DCNN network reached 80% classification accuracy to classify healthy and PD subjects.

Yuan and Cao [38] attempted to analyze EEGs via spectrogram images by using a CNN. Their motivation was to prove the clinical brain death diagnosis. In this paper Caffe network [57]was used to design a CNN. The EEG signals were acquired from the patients with brain damage. The EEG datasets contained 36 patients, including 19 coma subjects and 17 brain-dead subjects. Spectrogram images were generated from these signals using STFT. In addition, in order to increase the number of created images, six channels of the EEG signals were used to create spectrogram images. In addition, every window of STFT overlapped 20% with the adjacent windows. One hundred spectrogram images were extracted from the EEG data. Based on the experimental result, the CNN was able to distinguish between the coma and brain-dead classes with 96% and 94% accuracy, respectively.

Other researchers shed a light onto how CNNs are able to discriminate sleep stages. For example, [41] used the time-frequency domain of EEG signals in order to classify sleep stages. To reduce the bias and variance in spectrogram images, multitaper spectral estimation was utilized. The dataset included signals collected from 20 young healthy subjects. VGGNet was used with to extracted features by employ VGG-FE. VGG-FE achieved the highest accuracy with 89%, where most of sleep stages correctly detected slow wave sleep with (89%), rapid eye movement stage (81%), wake stage (78%) and N2 (75%) sensitivity. However, the N1 stage was incorrectly classified with 44% sensitivity.

An analogous study was directed using a CNN for sleep stage detection based on EEGs [37]. In this study, EEGs of 20 healthy young adults were recorded for evaluation. Morlet wavelets were used to produce a time-frequency representation. They achieved a high mean F1-score of 81%, where the accuracy over all sleep stages was 74%.

Andreotti et al. [48] proposed a simple CNN architecture that is trained from scratch using a large publicly available database. They provide EEG, EOG and EMG signals as an input to the CNN. The guided gradient-weighted class activation maps were used for visualizing this network's weights. A large publicly available dataset comprising single night PSG recordings of 200 healthy participants with (STFT). They generated time-frequency transforms for each epoch and modality of the signals. The continuous wavelet transforms (CWT) with a Morlet basis function was used to extract time-frequency images.

Another study was constructed to identify the human gait using the time-frequency representation with a CNN of human gait cycles. For example, [43] used the same approach to detect joint 2-dimensional (2D) spectral and temporal patterns of gait cycles. The signals were acquired from 10 subjects. Each signal was obtained from five inertial sensors that were worn and placed at the lower-back, right hand wrist, the chest, right knee, and right ankle. The experimental results were 91% subject identification accuracy. In this study, they conducted another experiment to improve the gait identification generalization performance by using two methods for an input level and decision score level multisensor combination. The performance improved and the accuracy reached 93.36% and 97.06%, respectively.

Another study attempted to improve CNN performance by combining it with an RNN in order to extract the movement pattern of the upper limb from EMG signals. Xia et al., 2018 [21]. The EMG signals were collected from eight subjects. These signals were recorded in six sessions for each subject and were converted to time-frequency spectrum images and used to train a one-dimensional CNN. The CNN included two recurrent layers in order to develop an RCNN. The experimental result proved that the CNN with the RNN achieved higher accuracy compared with that obtained by using CNNs alone. The authors claimed that these combinations can help to represent the features of EMG signals in the time and frequency domain in a better way. Based on their experimental results, the RCNN model can estimate limb movement with sufficient accuracy, and it was able to extract the features in the frequency domain and was robust against noises.

In this study [48] , the authors proposed the use of the CWT to represent the breathing cycles using scalogram images. The experiment attempted to identify the presence of wheezes and or crackles in breath. The CNN was trained to distinguish the scalograms from different classes. The result showed that the model achieved 84% and 87% accuracy of the class of crackles and wheezes, respectively.

VI. Discussion

The main motivation of this paper was to review various studies and papers that addressed the application of the DNN with the time-frequency representation. After analyzing more than 70 articles, 31 were further examined, and the results of each article were addressed. First, a number of findings were identified, and most of the studies were published during the last three years. In addition, convolutional neural networks, especially CNN that were pretrained, were the most commonly utilized. Furthermore, spectrogram and scalogram images were the most regularly used to train CNNs.

It can be observed that there is a large variety in the type of CNN applications that are used to learn patterns and features from the time-frequency domain automatically. All of the studies have investigated the ability of this approach in medical and manufacturing applications. Each of these studies has confirmed that CNN can extract the optimal information in order to address the required task. Most of these articles' results are comparable to state-of-the-art methods. CNNs are proven to be highly successful in analyzing any signal. Previously, reported studies mainly addressed medical signal analysis and diagnosis with the application of expert-designed features.

For example, a CNN using the time-frequency domain of the presented signals has already been shown to be competitive to traditional approaches. Traditional approaches usually extract a set of features from single or multiple channel signals based on human expertise. Therefore, this could be a difficulty for nondomain experts. Furthermore, traditional feature extraction methods are not capable of utilizing correlation information between different channels. CNNs are very powerful for learning features directly from the time-frequency domain without the need for signal processing and feature extraction methods [49].

Several significant points can be drawn from this survey. Most of the articles obtained their best result without any human intervention. Furthermore, they did not need to have domain knowledge for the analysis of signals. Based on the results of each article, deep learning can be considered as a sound basis for further optimization toward a competitive, fully automated feature extraction method to analyze signals. The potential of directly training a CNN using the time-frequency domain rather than only the time or the frequency domain, for example, in sound signals studies, has been claimed to be related to the time-frequency domain's very detail-rich but sufficiently sparse features that address complex characterization with overlapping sounds [49].

Another important point of this survey was the selection of the STFT-based images to train the CNN, However, studies confirmed that using sclogram is the usually obtained a good result. They motivated by the fact that the scalogram could better represent the nonstationary aspect of any type of signal unlike the STFT. In fact, wavelets are known to provide a robust time-frequency representation for different type of signals as they are localized both in time and frequency.

Therefore, their time–frequency domain information is rich and various [46]. Furthermore, [25] asserted that the wavelet transform is a time-frequency domain analysis tool that offers the best local features of the signal. Because of this, it is frequently used in denoising, feature extraction, and fault diagnosis. Hence, scalogram as input to the CNN can more accurately represent the nature of signals, which improves CNN feature encoding.

VII. Conclusion

This paper is presented to describe the background knowledge of how deep learning has been considered for the field of signal analysis and how it has transformed that field. Then, the state-of-art applications of CNN deep learning models for different types of tasks are identified. Finally, 35 articles from the literature that are related to the field of the study are considered, most of which were recently published since 2016. These articles from the literature are critically studied to provide a general overview on the performance of deep learning models with a time-frequency representation for signal analysis. From the reviews of the outcomes from these studies, it can be concluded that deep learning is able to learn features and patterns directly from time-frequency images. Thus, the brief nature of this survey can make a small but meaningful contribution to the current literature. In addition, it can provide insight on research challenges and future opportunities in the field of signal analysis. Moreover, CNN models generally outperform feature-engineered models.

References

[1] H. Alaskar, A. J. Hussain, F. H. Paul, D. Al-Jumeily, H. Tawfik, and H. Hamdan, “Feature Analysis of Uterine Electrohystography Signal Using Dynamic Self-organised Multilayer Network Inspired by the Immune Algorithm,” in International Conference on Intelligent Computing, 2014, pp. 206–212.

[2] H. Alaskar and A. J. Hussain, “Data Mining to Support the Discrimination of Amyotrophic Lateral Sclerosis Diseases Based on Gait Analysis,” in International Conference on Intelligent Computing, 2018, pp. 760–766.

[3] T. Balli and R. Palaniappan, “Classification of biological signals using linear and nonlinear features.,” Physiol. Meas., vol. 31, no. 7, pp. 903–20, Jul. 2010.

[4] X. Chen, X. Zhu, and D. Zhang, “A discriminant bispectrum feature for surface electromyogram signal classification.,” Med. Eng. Phys., vol. 32, no. 2, pp. 126–35, Mar. 2010.

[5] B. Liu, M. Wang, H. Yu, L. Yu, and Z. Liu, “Study of Feature Classification Methods in BCI Based on Neural Networks.,” in Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2005, vol. 3, pp. 2932–5.

[6] L. O. Machado, “Medical Application of Artificial Network Connectionist Models of Survival,” Stanford University, 1996.

[7] N. K. Orphanidou, A. Hussain, R. Keight, P. Lishoa, J. Hind, and H. Al-Askar, “Predicting Freezing of Gait in Parkinsons Disease Patients Using Machine Learning,” in 2018 IEEE Congress on Evolutionary Computation (CEC), 2018, pp. 1–8.

[8] K. Gröchenig, Foundations of time-frequency analysis. Springer Science & Business Media, 2013.

[9] H. M. Alaskar, “Dynamic self-organised neural network inspired by the immune algorithm for financial time series prediction and medical data classification,” PhD Thesis, Liverpool John Moores University, 2014.

[10] W.-L. Zheng and B.-L. Lu, “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” IEEE Trans. Auton. Ment. Dev., vol. 7, no. 3, pp. 162–175, 2015.

[11] H. Zhou et al., “Towards Real-Time Detection of Gait Events on Different Terrains Using Time-Frequency Analysis and Peak Heuristics Algorithm,” Sensors, vol. 16, no. 10, Oct. 2016.

[12] T. Verplancke et al., “A novel time series analysis approach for prediction of dialysis in critically ill patients using echo-state networks,” BMC Med. Inform. Decis. Mak., vol. 10, no. 1, p. 4, 2010.

[13] Y. Xin, S. Li, C. Cheng, and J. Wang, “An intelligent fault diagnosis method of rotating machinery based on deep neural networks and time-frequency analysis,” J. Vibroengineering, vol. 20, no. 6, pp. 2321–2335, 2018.

[14] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, “Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 2519–2523.

[15] A. Eltvik, “Deep Learning for the Classification of EEG Time-Frequency Representations,” Master’s Thesis, NTNU, 2018.

[16] M. Huzaifah, “Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks,” ArXiv Prepr. ArXiv170607156, 2017.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[18] D. Verstraete, A. Ferrada, E. L. Droguett, V. Meruane, and M. Modarres, “Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings,” Shock Vib., vol. 2017, 2017.

[19] S. A. Khan and S.-P. Yong, “An Evaluation of Convolutional Neural Nets for Medical Image Anatomy Classification,” in Advances in Machine Learning and Signal Processing, Springer, 2016, pp. 293–303.

[20] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, “Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals,” Inf. Sci., vol. 415, pp. 190–198, 2017.

[21] P. Xia, J. Hu, and Y. Peng, “EMG-Based Estimation of Limb Movement Using Deep Learning With Recurrent Convolutional Neural Networks,” Artif. Organs, vol. 42, no. 5, pp. E67–E77, 2018.

[22] G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.

[23] S. Shetty and Y. S. Rao, “SVM based machine learning approach to identify Parkinson’s disease using gait analysis,” in Inventive Computation Technologies (ICICT), International Conference on, 2016, vol. 2, pp. 1–5.

[24] N. E. Huang, “Introduction to the Hilbert–Huang transform and its related mathematical problems,” in Hilbert–Huang transform and its applications, World Scientific, 2014, pp. 1–26.

[25] S. Guo, T. Yang, W. Gao, and C. Zhang, “A Novel Fault Diagnosis Method for Rotating Machinery Based on a Convolutional Neural Network,” Sensors, vol. 18, no. 5, 2018.

[26] K. Kim, “Arrhythmia Classification in Multi-Channel ECG Signals Using Deep Neural Networks,” 2018.

[27] M. Längkvist, L. Karlsson, and A. Loutfi, “Sleep stage classification using unsupervised feature learning,” Adv. Artif. Neural Syst., vol. 2012, p. 5, 2012.

[28] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, “Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals,” Comput. Biol. Med., vol. 100, pp. 270–278, 2018.

[29] F. Andreotti, O. Carr, M. A. Pimentel, A. Mahdi, and M. De Vos, “Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG,” Computing, vol. 44, p. 1, 2017.

[30] S. Kiranyaz, T. Ince, and M. Gabbouj, “Real-time patient-specific ECG classification by 1-D convolutional neural networks,” IEEE Trans. Biomed. Eng., vol. 63, no. 3, pp. 664–675, 2016.

[31] G. Biagetti, P. Crippa, S. Orcioni, and C. Turchetti, “Surface EMG fatigue analysis by means of homomorphic deconvolution,” in Mobile Networks for Biometric Data Analysis, Springer, 2016, pp. 173–188.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[33] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning (Book in preparation) -. MIT press, 2016.

[34] C. Szegedy et al., “Going deeper with convolutions,” 2015.

[35] S. Savalia and V. Emamian, “Cardiac Arrhythmia Classification by Multi-Layer Perceptron and Convolution Neural Networks,” Bioengineering, vol. 5, no. 2, p. 35, 2018.

[36] U. R. Acharya et al., “Automated characterization of arrhythmias using nonlinear features from tachycardia ECG beats,” in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 000533–000538.

[37] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, “Automatic sleep stage scoring with single-channel EEG using convolutional neural networks,” ArXiv Prepr. ArXiv161001683, 2016.

[38] L. Yuan and J. Cao, “Patients’ EEG Data Analysis via Spectrogram Image with a Convolution Neural Network,” in International Conference on Intelligent Decision Technologies, 2017, pp. 13–21.

[39] K. Luo, J. Li, Z. Wang, and A. Cuschieri, “Patient-specific deep architectural model for ecg classification,” J. Healthc. Eng., vol. 2017, 2017.

[40] Z. Xiong, M. K. Stiles, and J. Zhao, “Robust ECG Signal Classification for Detection of Atrial Fibrillation Using a Novel Neural Network,” Computing, vol. 44, p. 1, 2017.

[41] A. Vilamala, K. H. Madsen, and L. K. Hansen, “Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring,” ArXiv Prepr. ArXiv171000633, 2017.

[42] G.-S. Hsu, A. Ambikapathi, and M.-S. Chen, “Deep learning with time-frequency representation for pulse estimation from facial videos,” in Biometrics (IJCB), 2017 IEEE International Joint Conference on, 2017, pp. 383–389.

[43] O. Dehzangi, M. Taherisadr, and R. ChangalVala, “IMU-Based Gait Recognition Using Convolutional Neural Networks and Multi-Sensor Fusion,” Sensors, vol. 17, no. 12, p. 2735, 2017.

[44] G. Ruffini et al., “Deep learning with EEG spectrograms in rapid eye movement behavior disorder,” bioRxiv, p. 240267, 2018.

[45] G. Vrbancic and V. Podgorelec, “Automatic Classification of Motor Impairment Neural Disorders from EEG Signals Using Deep Convolutional Neural Networks,” Elektron. Ir Elektrotechnika, vol. 24, no. 4, pp. 3–7, 2018.

[46] Y. Liang, Z. Chen, R. Ward, and M. Elgendi, “Photoplethysmography and Deep Learning: Enhancing Hypertension Risk Stratification,” Biosensors, vol. 8, no. 4, p. 101, 2018.

[47] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, “Learning Image-based Representations for Heart Sound Classification,” in Proceedings of the 2018 International Conference on Digital Health, 2018, pp. 143–147.

[48] F. Andreotti, H. Phan, and M. De Vos, “Visualising Convolutional Neural Network Decisions in Automated Sleep Scoring⋆,” in ICML Workshop, 2018, pp. 1–12.

[49] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Exploiting spectro-temporal locality in deep learning based acoustic event detection,” EURASIP J. Audio Speech Music Process., vol. 2015, no. 1, p. 26, 2015.

[50] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, “Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion,” in Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), 2017.

[51] I. Kiskin et al., “Mosquito detection with neural networks: the buzz of deep learning,” ArXiv Prepr. ArXiv170505180, 2017.

[52] J. Wang, J. Zhuang, L. Duan, and W. Cheng, “A multi-scale convolution neural network for featureless fault diagnosis,” in Flexible Automation (ISFA), International Symposium on, 2016, pp. 65–70.

[53] D. Lee, V. Siu, R. Cruz, and C. Yetman, “Convolutional neural net and bearing fault analysis,” in Proceedings of the International Conference on Data Mining series (ICDM) Barcelona, 2016, pp. 194–200.

[54] H. Liu, L. Li, and J. Ma, “Rolling bearing fault diagnosis based on STFT-deep learning and sound signals,” Shock Vib., vol. 2016, 2016.

[55] O. Janssens et al., “Convolutional neural network based fault detection for rotating machinery,” J. Sound Vib., vol. 377, pp. 331–345, 2016.

[56] D. Gurve and S. Krishnan, “Deep Learning of EEG Time-Frequency Representations for Identifying Eye States,” Adv. Data Sci. Adapt. Anal.

[57] Y. Jia et al., “Caffe: Convolutional architecture for fast feature embedding,” in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675–678.

Construction of TVET M-Learning Model based on Student Learning Style

Azmi S[1]
Department of Skills Development
Ministry of Human Resources,
62530 Putrajaya, Malaysia

Mat Noor S.F[2], Mohamed H[3]
Faculty of Information Science and Technology
Universiti Kebangsaan
43600 Bangi, Selangor Malaysia

***Abstract*—Mobile learning or m-learning is emerging as the innovation of virtual learning that used mobile devices for teaching and learning which can be accessed readily at hand anywhere either in classroom or group. Whereas preliminary study showed that Technical and Vocational Education and Training (TVET) institution were still using the conventional learning, where the students were not getting much exposure at all towards m-learning. In fact, this research discussed about the development and validating the usability of TVET m-learning model based on the user requirements that categorized into three main aspects: devices, users and social. The research scope focused on TVET students as the target users. While, user-centered design (UCD) method has been used in this research through four phases which were analyzing the user requirements, designing model, developing prototype and evaluating usability. Based on the usability evaluation results showed that TVET m-learning model is acceptable and compatible as a guideline of m-learning development for the TVET students. This TVET m-learning model brings benefits in improving the quality of teaching and learning in TVET institutions especially the public training skills institutions to achieve the nation goals in order to become a successful developing country and produce skilled workers in the future as well.**

Keywords*—M-learning; technical and vocational education and training (TVET); user-centered design (UCD)

I. Introduction

Malaysia has been forged ahead to become a successful developing country in all kind of sectors with evolving of latest mobile technology. In order to be more successful in TVET sector, the innovation of mobile technologies is applied for the rising quality of teaching and learning. In particular, it helps in exposing the students more on m-learning and life-long learning. The positive implications indirectly affect the TVET students in terms of careers, personalities and knowledge that strengthening them to be hired as well as adapting themselves in industrial work and environment [19], [21].

Based on previous studies, the implementation of m-learning in learning process brought positive impacts to the students and teachers. For instance, maximizing the sharing and exchanging of information activities, the use of mobile learning regardless of time and place and ease the delivery of information in learning [1]. The students benefit from m-learning to get the full information and knowledge easily, can access the content anywhere they want to. This paper focused on the model of TVET m-learning to clearly show the development of functionality and features of the system that were obtained based on user requirements and used as the guideline of m-learning application development.

Conventional learning or traditional learning is a pioneer of learning process that involves teaching and learning process between the students and teachers in classroom. Beyond that, the most common role of teachers is to convey knowledge while the students are just following every actions and words at the same time [20]. This process normally caused by one-way communication. In parallel with the evolving of mobile devices, the transformation of learning process towards mobile devices from computer has changed and well known as m-learning. Brown [2] stated that there were two types of flexible learning that were face-to-face learning and long-distance learning. While long distance learning is categorized into two parts that are e-learning and long-distance learning paper-based. Meanwhile, m-learning and online learning are specifically classified in e-learning.

M-learning is one of innovation of e-learning that have learning features which are embedded in e-learning such as accessing the learning materials, tasks, quizzes, forum discussion, user messenger, calendar and current notice activities [3]. Advances of mobile devices as the medium of interaction in technology that provide a lot of features such as short messaging service (SMS) and multimedia, sending and receiving emails, and applications that have the combination of short messaging service (SMS) and multimedia elements such as sound, image, animation and video [4]. In fact, the advances in technologies help the efficiency of learning. According to Bogdanović et al. [5], the implementation of quiz evaluation using m-learning among the students who familiar with mobile devices proved that m-learning itself was more effective, improving the motivation and students' performances.

M-learning provides unlimited excessing online learning regardless of time and place, the students could easily bring the mobile devices anywhere [6]. Besides that, m-learning enables teaching and learning opportunities to new learners by giving them chances to use active learning process in self-learning context [7]. The comparisons between previous models of m-learning focused on three main aspects which were devices, users and social [8]. These comparisons absolutely helped in determining the additional requirements needed in developing the model of TVET m-learning.

Those requirements of physical, technical or functionality features in the devices were actually needed to be compatible with the users. However, from the perspective of the users, Abdullah et al. [9] and Irwan Mahazir et al. [10] stated that mastery learning and process of learning were synced with the user mental model in Vygotsky – ZPD and model of Rudric and Krulik problem solving. Therefore, the characteristics of learning could be as reference to define the tendency of users' learning styles towards the m-learning model. While, Sha et al. [11] was focusing on self-regulated learning where the users could control the learning activities that specified on goals and achievements. These characteristics were defined to be integral in m-learning as the activity control.

Hence, to design the model of TVET m-learning completely, there were additional requirements to be included in designing the model. For device aspect, the device specifications must be compatible with the users. While, for user aspect, instead of focusing on learning style only there were also defining on digital technology skills to ensure that the model developed was suitable with the users' skills. For social aspect, beside the activity control, data security system, user guide, m-learning activity, automatic notification, messaging and file storage which were additional features in the TVET m-learning model.

Besides, usability elements were also important in order to validate the model of m-learning. Chang et al. [12] and Park et al. [13] stated that the basic elements of usability such as usefulness, easy to use and user satisfaction as the factors of user's acceptance with the used of m-learning. These three elements act as main elements of usability to validate the TVET m-learning model. According to Nielsen [14], the usability of mean score which is 4.00 means that the usability of system or prototype is at the high level (adequately usable).

II. Method

This study used user-centered design (UCD) to develop mobile learning for TVET. UCD is commonly used in designing process of human computer interaction research [15], [23]. In general, this method is an approach of the whole development processes based upon an explicit activity that involve the users.

The best way to develop the system continuously is defining every activity that involved real users in the whole processes of development [16]. Thus, the system developer will focus more towards the users' requirements which is the main target to make the system easy to use and compatible to the user.

This method involved four phases as shown in Figure 1. These four phases are requirements analysis, conceptual design, implementation and usability evaluation.

A. Phase I

As stated in Figure 1, the first phase started with data collection to acquire the user requirements through questionnaire which consists of five sections referring to the previous studies as specifications of device, learning styles, digital technology skills, m-learning features and user perceptions towards mobile technology.

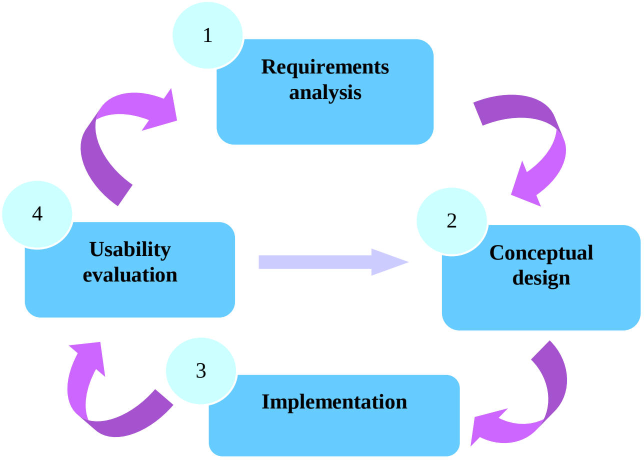


Fig. 1. Process of the user-Centered Design (UCD) Method.

B. Phase II

The second phase was designing the TVET m-learning model by referring to the data that obtained from the first phase. Designing the model involved three aspects which were device, user and social. Regarding the device aspect, the specifications of device used by the users are defined. Meanwhile, the user aspect defined the user's learning style and digital technology skills. Social aspect defined the features of m-learning. As a result, throughout this phase the TVET m-learning model was developed. So as to validate the model, prototype of mobile learning was developed.

C. Phase III

The third phase showed the development of prototype through two phases which were low-fidelity and high-fidelity prototype. The low-fidelity prototype was developed using the application software of Prototyping on Paper (POP) which the functions provided were the same as manual sketching in storyboard. While the high-fidelity prototype was developed using Android Studio software and developed based on data analysis of low-fidelity prototype through cognitive walkthrough method.

D. Phase IV

The last fourth phase was evaluating the usability of TVET m-learning prototype to test on the users in determining the element of usefulness, satisfaction, easy to use and learnability. The elements of usefulness consisted the useful of prototype towards users, activities provided were easy to complete, more effective and productive activities, activities can be done as expected and save time while activities were conducted. The element of satisfaction consisted of smoothness while using the system, good functionality of the system, entertaining and impressive system. In addition, the elements that were easy to use consisted of items where it can be used easily in the learning activities, user-friendly system, required fewest steps to accomplish activities, flexible interaction, clear instruction and guide, consistency, easy to recover the mistakes and successfully used by the users. While the element of learnability consisted of fastest way of learning on how to use the system, easy to learn the system, easy to remember on how to operate the system, quickly become skillful to use the system and not complicated to learn the user guide. These four elements of usability could be measured by using Lund [17] questionnaire on measurement of Likert Scale.

In order to evaluate the prototype specifically, there were two types of techniques used. The combination of observation and questionnaires techniques were used to measure the usability of the prototype. There were five steps in the usability evaluation. The first one was the evaluation of preparation by providing the experiment room and few tools such as video camera, mobile devices, note book, questionnaire instruments and the list of tasks in learning activities. This evaluation was done individually so that the researcher could focus on observing the users in every detail of activities during testing.

Before the evaluation process started, the researcher briefly explained about the evaluation. Next, user will follow the task list of activities in using m-learning while the researcher will observe and note-down every single action in case the activities could not be performed correctly. These activities and users' actions were recorded using video camera and at the same time those actions that had been done in the device were also recorded by using Mobizen application. The video recording was used as a backup of previous testing activities as a reference to researcher. After the learning activities using the TVET m-learning prototype done, the user was given the usability questionnaire to fill in.

The results of the usability testing were obtained by analysing using t-test which is used to validate the TVET m-learning model. These phases were summarized as shown in Figure 2.

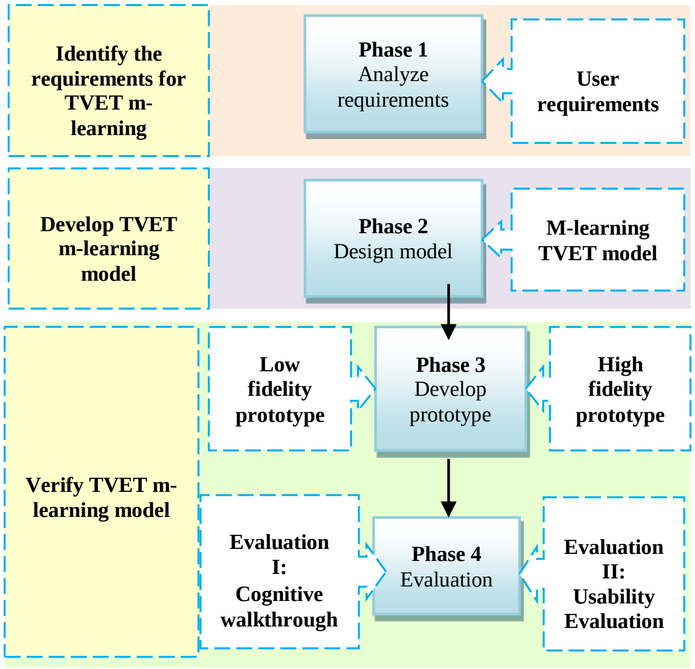

Identify the requirements for TVET m-learning
Phase 1
Analyze requirements
User requirements
Develop TVET m-learning model
Phase 2
Design model
M-learning TVET model
Low fidelity prototype
Phase 3
Develop prototype
High fidelity prototype
Verify TVET m-learning model
Evaluation I: Cognitive walkthrough
Phase 4
Evaluation
Evaluation II: Usability Evaluation

Fig. 2. User-Centered Design Method.

III. Result and Discussion

The TVET m-learning model was developed based on the analysis of user requirements which was focusing on three main aspects of m-learning: device, user and social. While the categories formation of the model was specified as device specifications, learning style, digital technology skills and m-learning features.

A. *TVET M-Learning Model*

The specifications of device were divided into four parts that were device, platform, features and functionality. Based on the user requirement, the device used was smartphone using the Android platform. In addition, the features in the device such as internet, 3G service and memory card are used for m-learning. The device also has the function of reading and opening digital files.

On the other hand, learning styles of TVET students were identified in the first dimension and third dimension of Felder and Silverman learning style [18]. The first dimension is about on how to process the information using the system. The learning style of this dimension was divided into two styles which were active and reflective. TVET students tend to be on active learning where they focused more on learning within groups rather than reflective or individual learning. Hence, the feature that adapted in TVET m-learning was learning within group instead of the individual learning. Meanwhile, third dimension refers to the tendency of students in memorizing so that they would be easy to use m-learning. Learning styles of this dimension were divided into visual and verbal. The tendency of TVET students on memorizing was easy in the form of visual rather than verbal. Therefore, the interface design of m-learning should be more graphically or icon used that can attract the TVET students' attention compared with to the interface design with text provided only.

Digital technology skills are the basic requirements for users to operate the system. With the level of skills, the developed system should be convenient for the users to operate. Therefore, the level of digital technology skills of TVET students were towards more on operating devices, social and mobile used which means they have the basic skills to operate the system well.

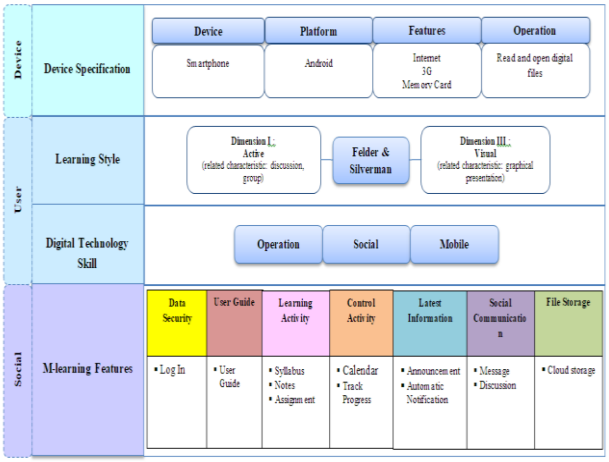


Fig. 3. TVET M-Learning Model.

By referring to the analysis of user requirements obtained, the features of TVET m-learning were identified by categorizing them in data security, user guide, learning activity, activity control, latest information, social communication and files storage. These features were embedded in the prototype development.

The category of data security, the students needed to log in the system for assuring the safety of data and information. While the user guide was provided to ease the users and act as a guideline on how to use TVET m-learning. Besides, the category of learning activities including the basic activities that used by the students daily such as accessing the syllabus, notes and tasks from the system.

As shown in Figure 3, TVET m-learning model was developed based on the analysis of user requirements. By using this model, the low-fidelity and high-fidelity prototype were developed. These prototypes will be evaluated through cognitive walkthrough and usability evaluation to validate the TVET m-learning model for TVET students.

Students could control the activities by adding reminders in the calendar and tracking their activities progress. The students would also be receiving the latest information which will be announced by the teacher and automatically notified them in the m-learning system. So that, the students were always be ready by any cause when receiving information and doing the learning activity. If the students needed help from the teachers or friends, they could use the m-learning messaging service to communicate individually or among group. For digital files storage, memory card is the basic requirement for the mobile device. Thus, students should not be worried with the limited size of memory card because the TVET m-learning could link directly to the cloud storage throughout the learning activities.

B. TVET M-Learning Usability

In usability evaluation, there were 30 students were chosen from Multimedia Software Technology from ILP Kuala Langat to evaluate the TVET m-learning prototype. The questionnaires were given to the students in determining four usability elements which were system usefulness, easy to use, learnability and user satisfaction. The results of the analysis were shown in Table 1. Overall, average of the mean score for each element through bar chart was shown in Figure 4.

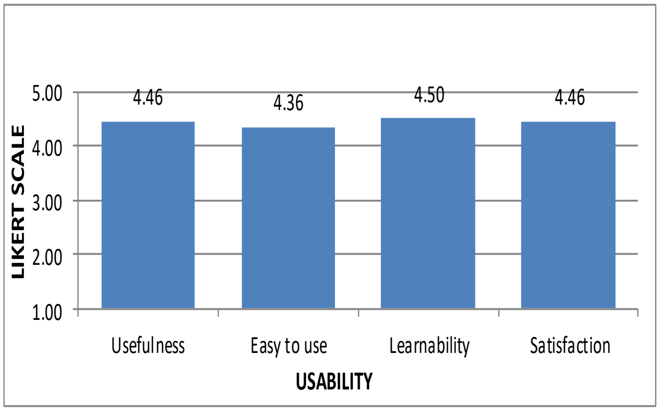


Fig. 4. Mean Score of TVET M-Learning Prototype Usability.

TABLE I. USABILITY EVALUATION OF TVET M-LEARNING PROTOTYPE

USABILITY OF TVET M-LEARNING PROTOTYPE							
Usefulness		***Easy to use***		***Learnability***		***Satisfaction***	
Item	***Mean Score***	***Item***	***Mean Score***	***Item***	***Mean Score***	***Item***	***Mean Score***
S1	4.43	S8	4.53	S16	4.63	S23	4.37
S2	4.57	S9	4.33	S17	4.57	S24	4.43
S3	4.40	S10	4.27	S18	4.50	S25	4.50
S4	4.30	S11	4.40	S19	4.40	S26	4.23
S5	4.33	S12	4.47	S20	4.43	S27	4.53
S6	4.57	S13	4.30	S21	4.47	S28	4.60
S7	4.60	S14	4.23	S22	4.53	S29	4.57
		S15	4.37				
Mean	**4.46**	**Mean**	**4.36**	**Mean**	**4.50**	**Mean**	**4.46**

The mean scores of usefulness, easy to use, learnability and satisfaction were 4.46, 4.36, 4.50 and 4.46 respectively. These values showed that the TVET m-learning were adequately usable and compatible to the users.

IV. Conclusions

This study achieved the objectives by developing and validating of TVET m-learning model. Based on the data of usability evaluation showed that the mean score values of usability elements were at the high level which means that the TVET m-learning is adequately usable and compatible to TVET students in teaching and learning process.

The main contributions in this study were divided into two fields: m-learning and TVET. The first field is m-learning where, the study developed the model which, involving the users from the beginning of the development process and until the final phase of validating the model. Therefore, the details of user requirements have been obtained as the main focus in developing TVET m-learning model. The usability of TVET m-learning was highly effective based on the usability elements observed such as usefulness, user satisfaction, ease of use and learnability.

This study also contributes in TVET learning style where TVET m-learning model will be a guideline for the system developer in designing m-learning specifically for TVET students. With the features offered in m-learning will provide further improvements in teaching and learning, therefore bringing the excellence to TVET institutions.

ACKNOWLEDGMENT

This study was supported by the Grant of GUP-2016-004, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia.

REFERENCES

[1] Hwang, G.-J., & Wu, P.-H., “Applications, impacts and trends of mobile technology-enhanced learning: A Review of 2008-2012 publications in selected SSCI journals,” *International Journal of Mobile Learning and Organisation, 8(2),* 83-95, https://doi.org/10.1504/IJMLO.2014.062346, 2014.

[2] Brown, T.H., “Towards a model for M-Learning in Africa,” *International Journal on E-Learning, 4(3),* 299-315, 2005.

[3] Kumar, S., Gankotiya, A.K., & Dutta, K., “A comparative study of moodle with other e-learning systems,” *ICECT 2011 – 2011 3rd International Conference on Electronics Computer Technology, 5,* 414-418, https://doi.org/10.1109/ICECTECH.2011.5942032, 2011.

[4] Rashid, Z. A., Kadiman, S., Zulkifli, Z., Selamat, J., Hisyam, M., & Hashim, M., “Review of Web-Based Learning in TVET: History, Advantages and Disadvantages. *International Journal of Vocational Education and Training Reseaarch,”* 2(211), 7-17, https://doi.org/10.11648/j.ijvetr.20160202.11, 2016.

[5] Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B., “Evaluation of mobile assessment in a learning management system. *British Journal of Educational Technology*,” *45*(2), 231–244, https://doi.org/10.1111/bjet.12015, 2014.

[6] Ally, M., & Prieto-Blázquez, J., “What is the future of mobile learning in education?” *RUSC, Universities and Knowledge Society Journal*, *11*(1), 142, https://doi.org/10.7238/rusc.v11i1.2033, 2014.

[7] Cochrane, T., & Antonczak, L., “Implementing a Mobile Social Media Framework for Designing Creative Pedagogies,” *Soc. Sci,3* (SEPTEMBER 2014), 359–377, https://doi.org/10.3390/socsci3030359, 2014.

[8] Koole, M. L., “A model for framing mobile learning. Mobile Learning: Transforming the Delivery of Education and Training,” 39, 2009.

[9] Abdullah, M. R. T. L., Hussin, Z., Asra, & Zakaria, A. R., “MLearning scaffolding model for undergraduate English language learning: Bridging formal and informal learning,” *Turkish Online Journal of Educational Technology*, *12*(2), 217–233, 2013.

[10] Irwan Mahazir, I., Norazah, M. N., Din, R., Abdul Rahim, A. A., & Che’ Rus, R., “Design and Development Performance-based into Mobile Learning for TVET,” *Procedia – Social and Behavioral Sciences, 174 (2015), 1764-1770,* https://doi.org/10.1016/j.sbspro.2015.01.835, 2015.

[11] Sha, L., Looi, C. K., Chen, W., & Zhang, B. H., “Understanding mobile learning from the perspective of self-regulated learning. *Journal of Computer Assisted Learning*,” *28*(4), 366–378, https://doi.org/10.1111/j.1365-2729.2011.00461.x, 2012.

[12] Chang, C.-C., Yan, C.-F., & Tseng, J.-S., “Perceived convenience in an extended technology acceptance model: Mobile technology and English learning for college students,” *Australasian Journal of Educational Technology*, *28*(5), 809–826, https://doi.org/10.14742/ajet.v28i5.818, 2012.

[13] Park, S. Y., Nam, M.-W., & Cha, S.-B., “University students’ behavioral intention to use mobile learning: Evaluating the technology acceptance model,” *British Journal of Educational Technology*, *43*(4), 592–605, https://doi.org/10.1111/j.1467-8535.2011.01229.x, 2012.

[14] Nielsen, J., “Usability Engineering. Morgan Kaufman Pietquin O and Beaufort R,” 1st Edition. Academic Press, 1993.

[15] Ritter, F. E., Baxter, G. D., & Churchill, E. F., “*Foundations for Designing User-Centered Systems*,” London: Springer London, https://doi.org/10.1007/978-1-4471-5134-0, 2014.

[16] Preece, J., Sharp, H., & Rogers, Y., “*Interaction Design: Beyond Human-Computer Interaction* (4th ed.),” United Kingdom: Wiley, 2015.

[17] Lund, A. M., “Measuring usability with the USE questionnaire,” *Usability Interface*, *8*(2), 3–6, 2001.

[18] Graf, S., Viola S., R., Lea, T., & Kinshuk., “In Depth Analysis of the Felder-Silverman Learning Style Dimensions,” *Reseach on Technology in Education*, *40*(1), 79–93, 2007.

[19] Rasul, M. S., Hilmi, Z., Ashari, M., Azman, N., Amnah, R., & Rauf, A., “Transforming TVET in Malaysia : Harmonizing the Governance Structure in a Multiple Stakeholder Setting,” *TVET-Online.Asia*, (4), 1–13, 2015.

[20] Hanimastura, H., Hairulliza, M. J., & Tengku Siti Meriam, T. W., “Success Factors for Knowledge Sharing Among Tvet Instructors,” *Journal of Theoretical and Applied Information Technology*, *85*(1), 12–20, 2016.

[21] Mat Yaacob, N., & Hussin, M., “Pendidikan Dan Latihan Teknik dan Vokasional (TVET) Dalam Konteks Memacu Pertumbuhan Ekonomi Malaysia,” *Proceeding International Conference On Global Education V*, pp. 1147-1159, 2017.

[22] Sohimi, N.E, Affandi, H.M, Fadzil, H., & Mohd Sattar, R., “Exploring The Malaysian Qlassic Practicality,” *Proceedings 4th International Conference on Vocational Education and Training (ICVET)*, *79*, 16–23, 2016.

[23] Wook, T. S. M. T., Mohamed, H., Judi, H. M., & Ashaari, N. S., “Applying cognitive walkthrough to evaluate the design of SPIN interface,” *Journal of Convergence Information Technology*, *7*(4), 106–115. https://doi.org/10.4156/jcit.vol7.issue4.13, 2012.

Recurrence Relation for Projectile Simulation Project and Game based Learning

Humera Tariq[1], Tahseen Jilani[2],
Usman Amjad[5]
Department of Computer Science
University of Karachi
Karachi, Pakistan

Ebad Ali[3]
School of Computing
National College of Ireland
Dublin, Ireland

Syed Faraz[4]
Center for Intelligent Signal and
Imaging Research (CISIR)
Universiti Teknologi Petronas
Perak, Malaysia

***Abstract*—Huge Gap has been observed on study of projectile simulation models relating it to speed of camera or frame per seconds. The objective of this paper is to explore and investigate time driven simulation models to mimic projectile trajectory; with an intent to highlight importance of game programming on native platforms. The proposed projectile recurrence relation and extensive mathematical modeling based on Triangular Series is an innovative outcome of project and game based learning used in BSCS-514 Computer Graphics Course at Department of Computer Science (DCS) University of Karachi (UOK). Box2D Replica of Popular 2D Mobile Game Angry Bird has been created on desktop to have an in depth mathematical and programming insight of commercial physics engine and discrete event simulation. Analysis has also been performed to answer certain key questions for progressive projectile trajectory for e.g. (1) With What angle, projectile should be launched? (2) What is the maximum height it will reach? (3) How long it will take for landing? (4) What will be its velocity to reach a desired height? (4) Where it will hit? (5) How it will bounce? The above stated questions are important to answer so that projectile motion within engineering, Gaming and other CAD Applications can be taught and programmed correctly specially on native platforms like OpenGL. Besides reporting Numerical results, a successful projectile based game making has been compiled and reported to validate the significance of project based learning in classrooms and labs.**

Keywords*—*Projectile; game programming; simulation; angry birds; linear drag; trajectory; impulse

I. Introduction

Understanding projectiles in an interdisciplinary manner is exceedingly important at variety of levels to incorporate them in real life applications. Engineering applications involving projectile use deep and precise analytical equations which are difficult to grasp and requires an in depth domain knowledge [1][2][3]. Classical literature documented Physical class room experiments to study and build understanding about projectile models [4][5]. CAD and gaming applications requires the same background knowledge along with programming skill for simulation and visualization of projectile trajectory. In this paper, an attempt to unlock the projectile behavior of available game development platforms from simulation perspective on computers. Platforms are available to facilitate engineers and programmers for building these physical models without handling the pressure of complex analytical mathematics and statistics. As a senior computer science faculty, I strongly believe that understanding these mathematical foundation has played an immense important role in polishing ones' cognitive, engineering and programming skills [6] - [12]. There exist three common models to simulate projectile in computer applications for gaming and simulation; they are: (1) No drag model (2) Linear drag model and (3) Quadratic drag model. In no drag model, the motion of projectile is mainly dependent on initial velocity and the angle of launch. On the other hand, both linear and quadratic drag model taken into account, the air resistance effect on projectile trajectory. Demonstration of projectile has been done through replica of famous angry bird game using Box2D [13] [14] and figure out that which model has been used in this physics engine? Rest of the paper is organized as follows: Section II and III discuss standard No drag and linear drag trajectory models respectively. Section IV till Section VII comprise of extensive derivations based on recurrence relation, triangular numbers and Frames Per Second Criteria (FPS). The derivations are inspired by Box2D and OpenGL experiments discussed in [13] -[17]. Experiments on spread sheet and through game programming has been discussed in Section VIII which follows Results and Discussion described in Section IX. Conclusion and Future Work has been presented in the end of this paper.

II. No Drag Trajectory Model

In general, the motion produced by a body which follows projectile trajectory reached a certain elevation and then allow to descending as a mirror of elevation. The vector of the motion can be divided into two components, 'x': the horizontal component and 'y': the vertical component of the motion. The force at the horizontal vector remains unchanged throughout the motion while the vertical forces applied at the birds frequently changes due to the effect of gravity and height of the bird. For computing the velocity of the birds during the flight time, following equation can be used.

$$V_x = V_0 Cos\theta \tag{1}$$

$$V_y = V_0 Sin\theta - gt \tag{2}$$

The acceleration of the object, in both the components of motion remains constant throughout the flight time. Horizontal component of acceleration remains equal to zero $\boldsymbol{a}_{x} = \mathbf{0}$ and negative gravitational value at vertical $\boldsymbol{a}_{Y} = -\boldsymbol{g}$. During flight time, highest point of the bird can achieve with the given initial angle and velocity applied

$$h = \frac{V_0^2 Sin^2\theta}{2g} \tag{3}$$

Equation (3) will compute the height of the flight with respect to the ground. Range of the projectile launched is also important to calculate because this information will enable us to determine the landing position of the object. The mathematical model for computing total range of the bird is mentioned below

$$R = \frac{V_x^2 Sin^2 2\theta}{g} \tag{4}$$

Range of the bird depends on two independent variables in the projectile motion of a game, first is the adjusted angle and second is the adjusted velocity of the bird. The angles of the launch greatly affect the range of the bird due to the trigonometric function of Sine on the angle. It is computed that at any velocity the bird will achieve the maximum range value if the angle adjusted is 45^0 because of $\boldsymbol{Sin}(\ \mathbf{2}\ \Theta)$ in Eq (4). When the value of Θ is equal to 45 then $\mathbf{2}\Theta$ becomes equal to 90 degrees and thus maximum range will be

$$R = \frac{V_x^2}{g} \tag{5}$$

From simulation perspective, total time of flight is calculated by manipulation of Eq. (2) and Eq. (3) and finally putting h =0 to attain maximum range at ground. It is not difficult to have following equation for computation of total time of flight 'T' in advance as follows

$$T = 2 * \frac{V_y}{g} \tag{6}$$

Once total time of flight is calculated, computation of points (X, Y) on trajectory will become possible by dividing time T into small equal space intervals t +delT as follows

$$X(t) = V_x * t \tag{7}$$

$$Y(t) = X(t)\tan\theta - \frac{1}{2g}\left(\frac{1}{V_0^2 Cos^2\theta}\right)X(t)^2 \tag{8}$$

Eq (7) and Eq (8) are used to determine instantaneous position of projectile given initial Velocity as 'V' and initial angle 'theta'. It is important to note that Y(t) depends on X(t).

III. Linear Drag Trajectory Model

Linear impulse is defined as a linear force applied on anybody i.e. $F = ma$. In many games, an external backward force is applied on body before launching it as projectile for e.g. drag of the bird in game "Angry bird" is applied through a slingshot controlled by user. This drag force will of course be decomposed into its components along x and y-axis. The area for dragging the object is specified with reference to the initial position of the projectile. The drag backward is restricted on x-axis; the same will be true for y-axis. Drag displacement will be directly proportional to drag force. F. If Δd represents the change in distance due to drag then

$$F \propto \Delta d \tag{9}$$

Eq (9) can be decomposed into two independent equations with respect to x and y as follows:

$$F_x \propto \Delta d_x \quad ; \quad F_y \propto \Delta d_y \tag{10}$$

The objective is to find final velocity V_2 with the help of initial velocity V_1 so that we can determine the launch speed with which it moves along projectile trajectory. Assume fixed time incremental approach with Δt = 1. Assume that the Forces will remain constant throughout the projectile motion. The body will then move with uniform acceleration on x-axis. The y-component of launch velocity will be affected by the acceleration due to gravity g. The expression for final velocity along x-axis can be formulated as follows

$$F_x = ma_x\ ; \quad F_x = m(V_{2x} - V_{1x})$$

$$\frac{F_x}{m} = (V_{2x} - V_{1x})$$

$$\frac{F_x}{m} + V_{1x} = V_{2x}$$

$$V_{2x} = \frac{F_x}{m} + V_{1x} \tag{11}$$

Along y-axis; following simplified drag model has proposed to determine final velocity progressively

$$F_y = ma_y - gt \tag{12}$$

Aa per our assumption, the change in momentum of projectile can be compared with change in vertical forces as follows

$$F_y = m(V_{2y} - V_{1y}) - gt$$

$$F_y + gt = m(V_{2y} - V_{1y})$$

$$V_{2y} = \frac{F_y + gt}{m} + V_{1y} \tag{13}$$

IV. Recurrence Relation for Height

Our recurrence relation is inspired by Box2D description on projectile [13 [14]. To derive a general formula, that can be used to find the final velocity $V_{2y(t)}$ at any given point in time; each velocity assume to contain the sum of all the previous velocities. In a general way the velocities are in the arithmetic progression manner. If we use Eq (14) at each frame display in games, then the general expression for finding the height at any instant can be figure out by working from Eq (14) till Eq (17)

$$V_{2y(1)} = \frac{F_y + g(\frac{1}{fps})}{m} + V_{1y(0)} \qquad (14)$$

$$V_{2y(2)} = \frac{F_y + g(\frac{2}{fps})}{m} + V_{1y(1)}$$

$$V_{2y(3)} = \frac{F_y + g(\frac{3}{fps})}{m} + V_{1y(2)}$$

$$V_{2y(3)} = \frac{F_y + g(\frac{3}{fps})}{m} + \frac{F_y + g(\frac{2}{fps})}{m} + \frac{F_y + g(\frac{1}{fps})}{m}$$

$$V_{2y(3)} = \frac{1}{m}(F_y + g(\frac{3}{fps}) + F_y + g(\frac{2}{fps}) + F_y + g(\frac{1}{fps}))$$

$$V_{2y(3)} = \frac{F_y}{m}\{1 + \frac{g}{F_y}(\frac{3}{fps}) + 1 + \frac{g}{F_y}(\frac{2}{fps}) + 1 + \frac{g}{F_y}(\frac{1}{fps})\}$$

$$V_{2y(3)} = \frac{F_y}{m}\{3 + \frac{g}{F_y}(\frac{3}{fps}) + \frac{g}{F_y}(\frac{2}{fps}) + \frac{g}{F_y}(\frac{1}{fps})\}$$

$$V_{2y(n)} = \frac{F_y}{m}\{n + \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps})\} \qquad (15)$$

If we use Eq(15) to find instant launch speed and we are given the initial height of the projectile then following equation can be used to find the height at any instance of time:

$$H_n = V_{2yn} + H_0 \qquad (16)$$

This above equation can be further expanded into Eq (17) as we substitute final velocity from Eq(15). Here H_0 represents the initial height of the projectile at which it was launched

$$H_n = \frac{F_y}{m}\{n + \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps})\} + H_0 \qquad (17)$$

Where i is the counter to track time instants while projectile is flying above the ground.

V. Time of Flight as Triangular Series

The total time for the flight of the projectile can be obtained by substituting final velocity as zero because the projectile will come to rest as it hits the ground, therefore $V_{2yn} = 0$ where n is used to represent instantaneous time of flight

$$V_{2y(n)} = \frac{F_y}{m}\{n_{total} + \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps})\} \qquad (18)$$

$$0 = \frac{F_y}{m}\{n_{total} + \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps})\}$$

$$0 = \{n_{total} + \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps})\}$$

$$n_{total} = \frac{g}{F_y}(\frac{\sum_{i=1}^{n} i}{fps}) \qquad (19)$$

The sign of summation can be interchanged by $\frac{n(n+1)}{2}$ because this series is the nth partial sum of the series of triangular numbers.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad (20)$$

$$n_{total} = -\frac{g}{F_y}(\frac{\frac{n_{total}(n_{total}+1)}{2}}{fps})$$

$$-2\,fps\frac{F_y}{g} = n_{total} + 1$$

$$1 - 2\,fps\frac{F_y}{g} = n_{total}$$

$$n_{total} = \frac{g - 2\,fpsF_y}{g}$$

$$T = \frac{g - 2*fps*F_y}{g} \qquad (21)$$

VI. Relating Velocity and Height

The derivation to find the velocity at the maximum height will follows from Eq (22) till Eq (25). Substituting time of flight from Eq (20) into Eq (15) of final velocity, we will get Eq (22)

$$V_{2y(n)} = \frac{F_y}{m}\{n + \frac{g}{F_y}(\frac{n(n+1)}{2\,fps})\} \qquad (22)$$

Eq (20) provides the total time of flight but maximim height will be achieved at half time of the total therefore Eq (21) will turn into Eq (23) as follows

$$n_{1/2} = \frac{(F_y)(fps) + g}{g} \qquad (23)$$

Assuming $x = (fps)F_y + g$; Eq (23) become

$$n_{1/2} = -\frac{x}{g}$$

Substituting above fact into Eq (22), We have

$$V_{2y} = \frac{F_y}{m}\{-\frac{x}{g} + \frac{g}{F_y}(\frac{-\frac{x}{g}(-\frac{x}{g}+1)}{2\,fps})\}$$

$$V_{2y} = \frac{F_y}{m}\{\frac{g^2}{F_y}(\frac{-\frac{x}{g}(-\frac{x}{g}+1) - 2\,fps(F_y x)}{2\,fps(g)})\}$$

$$V_{2y} = \frac{1}{m}\{g^2(\frac{-\frac{x}{g}(-\frac{x}{g}+1) - 2\,fps(F_y x)}{2\,fps(g)})\}$$

$$V_{2y} = \frac{1}{m}\{g(\frac{(\frac{x^2 - xg}{g^2}) - 2\,fps(F_y x)}{2\,fps})\}$$

$$V_{2y} = \frac{1}{m}\{g(\frac{(\frac{x^2 - xg - 2\,fps(F_y x)g^2}{g^2})}{2\,fps})\}$$

$$V_{2y} = \frac{1}{m}\{(\frac{x^2 - xg - 2\,fps(F_y x)g^2}{2\,fpsg})\}$$

$$V_{2y} = \frac{x}{m}\{(\frac{x - g - 2\,fps(F_y)g^2}{2\,fpsg})\} \tag{24}$$

$$V_{2y} = \frac{(fps)F_y + g}{m}\{(\frac{(fps)F_y + g - g - 2\,fps(F_y)g^2}{2\,fpsg})\}$$

$$V_{2y} = \frac{(fps)F_y + g}{m}\{(\frac{(fps)F_y - 2\,fps(F_y)g^2}{2\,fpsg})\}$$

$$V_{2y} = \frac{(fps)F_y + g}{m}\{(\frac{(fps)F_y(1-2g^2)}{2\,fpsg})\}$$

$$V_{2y} = \frac{(fps)F_y + g}{m}\{(\frac{F_y(1-2g^2)}{2g})\}$$

$$V_{2y} = \frac{\{(fps)F_y + g\}\{F_y(1-2g^2)\}}{2gm} \tag{25}$$

Eq (24) and Eq (25) respectively represent the relation which represent velocity of projectile with respect to displayed frame in game and illustrate frame animation.

VII. Impulse Force and Final Velocity

The equation for the final velocity due to the impulse force will be considered as follows

$$V_2 = V_1 + at - gt$$

$$V_2 = V_1 + (a-g)t \tag{26}$$

Where V_2 is the final velocity which is the consequence of the initial impulse, and V_1 is the initial velocity, a is the acceleration provided to the body whereas g represent the force of gravity applied on the motion of the body. This equation only tells us about the instantaneous velocity of a body at any given time. Indeed, apart from calculating the velocity, we are more interested in getting the current displacement at y-axis by the body with a given value of velocity. For e.g. What will be the distance covered by the body when $t = 4$ and $V_1 = 8\ m/s$? In this scenario, it can be said that every displacement is dependent upon the previous displacement that is why the following equation is given to fulfill our requirement.

$$S_1 = V_1\ ; \qquad S_2 = V_2 + V_1$$

$$S_3 = V_3 + V_2 + V_1$$

$$S_4 = V_4 + V_3 + V_2 + V_1$$

$$S_4 = V_1 + (a-g)(3) + V_1 + (a-g)(2) + V_1 + (a-g)(1) + V_1 + (a-g)(0)$$

$$S_n = (n+1)V_1 + (a-g)\sum_{i=1}^{n} i \tag{27}$$

Converting $\sum_{i=1}^{n} i$ notation to the general formula of triangular series $\frac{n(n+1)}{2}$

$$S_n = (n+1)V_1 + (a-g)\frac{n(n+1)}{2} \tag{28}$$

The total time taken by the body to complete its trajectory where the velocity of the body becomes equal to zero. The equation for total time taken for the flight is as follows:

$$S_n = (n+1)V_i + (a-g)\frac{n(n+1)}{2}$$

$$0 = (n+1)V_i + (a-g)\frac{n(n+1)}{2}$$

$$0 = (n+1)\{V_i + (g-a)\frac{n}{2}\}$$

$$0 = \{V_i + (g-a)\frac{n}{2}\} \qquad V_i = (g-a)\frac{n}{2}$$

$$2V_i = (g-a)n\ ; \quad \frac{2V_i}{(g-a)} = n\ ; \quad n = \frac{2V_i}{(g-a)} \qquad 29$$

$$T = \frac{2 * V_i}{g - a}$$

The equation for calculating the maximum height achieved by the body during the flight is derived by substituting EQ (29) into EQ (27) as follows

$$S_{\max} = (\frac{V_i}{(g-a)} + 1)V_i + (a-g)(\frac{\frac{V_i}{(g-a)}(\frac{V_i}{(g-a)} + 1)}{2})$$

$$S_{\max} = (\frac{V_i(V_i + g - a)}{(g-a)}) - (\frac{V_i(V_i + g - a)}{2(g-a)})$$

$$S_{\max} = \frac{2\{V_i(V_i + g - a)\} - V_i(V_i + g - a)}{2(g-a)}$$

$$S_{\max} = \frac{V_i(V_i + g - a)}{2(g-a)} \qquad (30)$$

VIII. Experimentation

A. Spread Sheet Simulaiton of Pojectile Trajectory

The purpose of spread sheet simulation is to visualize and test our proposed linear impulse based recursive projectile model. We already discussed Linear Drag Model in Section III. The force along x-axis is assumed to be constant according to our simulation assumption. The required initial parameters are set as follows for demonstration purpose

$F_x = 1 \qquad Initial\ Height\ H_0 = 0$

$F_y = 420 \qquad mass = 11.5\ unit$

$acceleration\ due\ to\ gravity = -9.8$

$V_{1x} = 0\ ;\ V_{1y} = 0;$

$FPS = 60 \ \rightarrow\ t_i = i * \frac{1}{60}\ seconds \quad i = 1,2,3$

With above initial conditions we use Eq (11) and Eq (13) to update velocity along each axis respectively at each frame or display of projectile. For simplicity, the change in time between frame displays is taken to be unit which means that time instances are taken as a fraction of 60. **Table 1** and **Table 2** Presents Numerical Results of progression in velocity along each axis. From Table 1, it has been observed, that the projectile is moving with uniform acceleration along x-axis as the difference between two adjacent velocities remained a constant.

It has been observed from **Table 1** and **Table 2** that the generated impulsive force has enough large magnitude over a very small interval of time (fraction of a second) that it causes a significant change in the momentum which in turn lift the projectile up in the air in response to initial drag during its flight. To build simulation start with $v_1 = 0$; compute $v_2 = \frac{F_x}{m} + v_1$ as described in Eq (11)

TABLE I. Simulation Sample EQ (11)

i	V_{1x}	$V_{1.1}$	V_{2x}
1	0	0.086957	0.086957
2	0.08695652	0.173913	0.173913
3	0.17391304	0.26087	0.26087
4	0.26086957	0.347826	0.347826
5	0.34782609	0.434783	0.434783
6	0.43478261	0.521739	0.521739
7	0.52173913	0.608696	0.608696
8	0.60869565	0.695652	0.695652
9	0.69565217	0.782609	0.782609
10	0.7826087	0.869565	0.869565
11	0.86956522	0.956522	0.956522
12	0.95652174	1.043478	1.043478
13	1.04347826	1.130435	1.130435
14	1.13043478	1.217391	1.217391
15	1.2173913	1.304348	1.304348
16	1.30434783	1.391304	1.391304

TABLE II. Simulation Sample EQ (13), EQ (17)

i	$t_i = i * \frac{1}{60}$	V_{1y}	V_{2y}	H_i
1	0.016	0	36.5217391	0
2	0.033	36.52174	72.1913043	36.52174
3	0.05	72.1913	107.008696	72.1913
4	0.066	107.0087	140.973913	107.0087
5	0.083	140.9739	174.086957	140.9739
6	0.1	174.087	206.347826	174.087
7	0.116	206.3478	237.756522	206.3478
8	0.133	237.7565	268.313043	237.7565
9	0.150	268.313	298.017391	268.313
10	0.166	298.0174	326.869565	298.0174
11	0.183	326.8696	354.869565	326.8696
12	0.200	354.8696	382.017391	354.8696
13	0.216667	382.0174	408.313043	382.0174
14	0.233333	408.313	433.756522	408.313
15	0.25	433.7565	458.347826	433.7565
16	0.266667	458.3478	482.086957	458.3478

B. Spread Sheet Simulation of Time as Triangular Series

It has been observed that current displacement of projectile can be seen as series of instantaneous velocities sum together where each instant can take the form of triangular number in series. The simulation sample of model described in Section VII is illustrated in Table 3 below. Some initial parameters for demonstration purpose are given following values

$$m = 11.5;\ \ g = -9.8;\ \ n = \{0,1,2,3 \ldots\ldots\};\ t = \frac{n(n+1)}{2}$$

$$\frac{F_y}{m} = 0.608696 \quad\ ; \quad \frac{g}{F_y} * FPS = -0.02333$$

TABLE III. SIMULATION SAMPLE EQ (22)

t	g/(Fy*FPS)t	n + g/(Fy*FPS)t	Fy/m* (g/(Fy*FPS)t
0	0	0	0
1	-0.023333	0.976666667	0.594492754
3	-0.07	1.93	1.174782609
6	-0.14	2.86	1.740869565
10	-0.233333	3.766666667	2.292753623
15	-0.35	4.65	2.830434783
21	-0.49	5.51	3.353913043
28	-0.653333	6.346666667	3.863188406
36	-0.84	7.16	4.35826087
45	-1.05	7.95	4.839130435
55	-1.283333	8.716666667	5.305797101
66	-1.54	9.46	5.75826087
78	-1.82	10.18	6.196521739
91	-2.123333	10.87666667	6.62057971
105	-2.45	11.55	7.030434783
120	-2.8	12.2	7.426086957

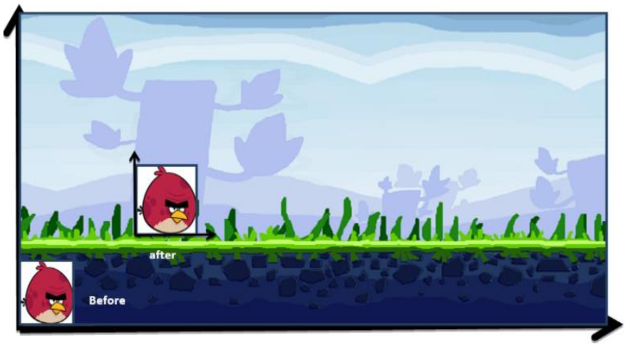


Fig. 1. Background Modeling and Translation Concept.

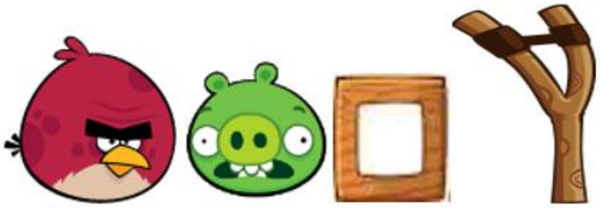


Fig. 2. Foreground Entities of Game as Textured Polygon.

C. Projectile Simulation on Box 2D and OpenGL

Game Programming is an important and impressive tool to practice simulation. Box2D is used to simulate core projectile functionality whereas OpenGL 'freeglut' library on windows is used for rendering and to interact with user through mouse and keyboard. Box2D modeled projectile as linear impulse force [13] [14]. The 2D game pipeline consist of following minimal basic steps to experiment with projectile. (1) Modeling (2) Animation (3) Collision Detection. Modeling of game world means setting up background and foreground. Background is indeed a flat polygon with height of 480 and width of 1920 pixels. The wide polygon is split into smaller polygon to be used as textured sheet. Fig. 1 demonstrate simple UV mapping on a rectangle to model background comprising of clouds, grassy land, and sun. So textured polygon made it convenient for a programmer to have nice background using glossy images. Fore ground entities Comprises of main characters such as Birds, Pigs, walls (Boxes) and slingshot. Game Characters are designed with simple polygons and we employ uv mapping to animate characters. These characters are represented by Bitmap images as shown in Fig. 2. As a matter of fact, all foreground objects are modelled at origin; later they are translated to specific position by using coordinate system transformation as illustrated in the for the Bird.

The land type and background in the game need to change as projectile launces and trying to hit the target at distance ahead of its current position. This effect is achieved by setting predefined multiple polygons shown in Fig. 3 as Polygon A, B and C. When the game starts, user will able to see the initial background but when it uses sling shot to launch projectile, the coordinates get changed and user will experience change of background at the output screen. These multiple lands or polygons will be visible to the user at its screen and the view will be updated at runtime. Foreground modeling is another important part of the game as they will change with their state on different triggers and events. All Game character (birds), pigs (enemy), boxes, hurdles and other bitmaps are counted as foreground elements. The foreground is divided into two section of the screen. The left portion of the foreground is dedicated for modeling the slingshot and the birds which are in the control of the user. This portion will enable the user to adjust the angle and the speed for launching the birds at a trajectory of projectile motion through the slingshot in order to hit the walls and to eliminate the pig. The other portion of the foreground consists of the defensive walls for the pigs. All the Birds in game are modeled by drawing circles and then applying bird texture on them. While Boxes and slingshots are modeled by applying texture on to the rectangles. When the game is initialized, all the characters and the objects in the game are drawn at the initial position of the screen, at (0, 0). The positions of these various models in this game are translated with respect to their functionality and behavior, as the initial bird is set near the slingshot elevated by default ready to be shot by the user, on the other hand other birds wait for their turn in a row wise manner. The pigs are modeled in between the walls or the boxes on the other side of the screen.

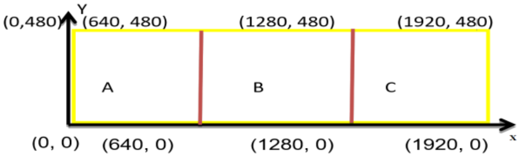


Fig. 3. Wide Polygon Design to Map Levels in Game.

Physics played a vital role in animation of all characters. Following aspects need to be considered in order to model the desired animation. (1) In which direction it should be launched to meet the desired result? (2) How high it will go? (3) How long it will take to land on the floor? (4) Where it will hit the ground? (5) How far away from the ground? (6) Collision Detection. Slingshot is used in this game in order to set the launching direction and to make accurate shot at the pigs. The slingshot uses the birds for attacking pigs confined in a defensive wall. The slingshot is animated with the help of mouse events and buttons. Slingshot is stretched with the left button of the mouse for adjusting the force or the speed and the angle at which the bird will be shot. The right button is used to launch the bird at the pigs with the angles and the force adjusted earlier. Birds being shot through the slingshot are translated at an angle with an initial velocity and become projectile under the influence of gravity. The animation of the bird is handled according to trajectory model already described in detail from Eq (1) till Eq (13). Observe the arrangement of Woods in Fig 4. Falling of wood Boxes is established under the acceleration due to gravity and is achieved by using Newton's second equation of motion. $h = \mathrm{v_i}\,\mathrm{t} + \frac{1}{2}gt^2$ But, the initial velocity $\mathrm{v_i} = 0$

Therefore, the rate at which body is falling is: $g = \frac{2\mathrm{h}}{\mathrm{t}^2}$

Another Noticeable construct is once again the application of parametric equation to generate points on the curve as already described in Eq (7) and Eq (8). After modeling and animation, the last aspect of a basic 2D game is to detect collision between entities. Principal of Separating Axis Theorem (SAT) states that: "**If two convex objects are not penetrating, there exists an axis for which the projection of the objects will not overlap.**" To detect whether two objects are being collided, projection of their axis are calculated. If the distance between all the projections is equals to or less than zero, then it is said that the object has collided with each other. If projection of bird overlaps with projection of box on both horizontal and vertical, animation event is triggered. This concept has been extended and applied on every polygon as shown in Fig. 5.

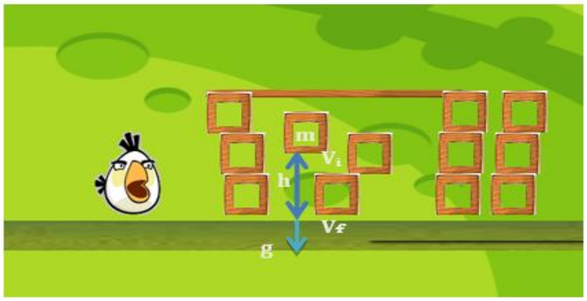


Fig. 4. Animation to Show Fall of Woods or Boxes.

One can note that how a 2D collision detection problem can be transformed into 1D problem. Orange and Blue Lines in Fig. 5. are Projection of Polygons A and B on axes respectively. Black Line represents the Portion on the line where a plane can be inserted separating A and B. Two steps are needed to perform SAT: (1) Finding axis (2) Projection of shapes (3) Collision detection. Axes are simply normal of each shape's edges. This can be calculated by subtracting the vertices of the respective edges and then taking perpendicular of it as show in the Pseudocode of Table 4. Projecting shapes onto Axes and then comparing overlapped portion of the projection would result in Collision Detection. Pseudo code in Table 5 models this idea

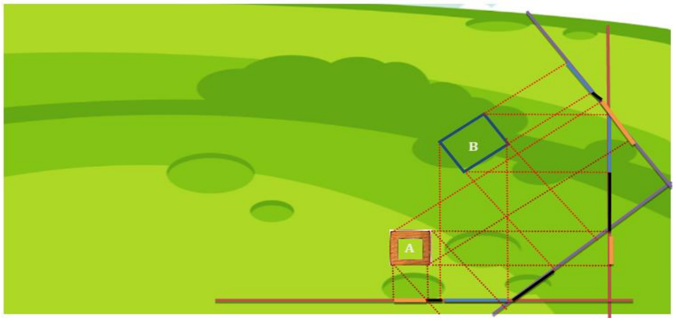


Fig. 5. Overlapping Projections for Collision Detection.

TABLE IV. FINDING AXIS BY VECTOR SUBTRACTION

```
for all Edges in A
        P1 = First Vertex of A
        P2 = Second Vertex of A
        Vector edge = P1.subtract (P2);
        Vector normal = edge.perp ();   // Axis
        arrayOfAxis.add (normal);
end for
```

TABLE V. COLLISION DETECTION BY OVERLAPPING TECHNIQUE

```
for all Shapes in S
  for all Axes of S in A
    Projection p1 = s.project (A);
    Projection p2 = s2.project (A); // s2 is the second shape.
if (!p1.overlap(p2))    return false;
    end for Axes
end for S
```

IX. Result and Discussion

Successful Simulation of extensive mathematical models has been demonstrated using both Spread Sheet and C++ programming by integrating Box2D and OpenGL. Fig. 6 shows the result of applying impulse force as agent at initial time of the motion, after which the velocity produced by this impulse and force of gravitation contributes for further movement of the projectile. In our scenario the impulse is being applied at the center of the body. The y-component of the force mimics the angle of launching the projectile towards the target while x-component is the linear force or speed with which the bird is supposed to launch. FPS is 60 frames per second and is the rate at which box2D updates or refreshes its View Window. Fig. 7 shows that it is possible that flight time will take form of triangular series and the projectile trajectory is successfully followed if triangular increment is used for simulation. The simulation of linear impulse in box2D and OpenGL requires two arguments: (1) An initial vector $\mathbf{V}_0$ containing x and y component of velocity (2) A Point from where impulse has to be triggered. Y-component of final velocity is denoted with $V_{2y(t)}$ i.e. velocity at time instance t in future; g is 9.8 m/s, 'F' represents y-component of force applied on mass 'm'. FPS is defined as the default frame per second value of box2D which is 60. Reciprocal of 'FPS' provides us with time to render single frame which in turn plot the position of body. To mimic original game, projectile is fired and controlled through mouse pointer. To aim pigs at a certain angle; birds are controlled by click buttons of the mouse. Left button is used for adjusting speed value of the bird whereas the right button controls the angle of the shot being made. All the variable setup is shown in Fig. 8 for reader convenience. By using Mouse interactions with the system we developed multiple techniques to deal with the angle, at which the bird would fly, and magnitude of force, with which the bird would be fired, from the slingshot. The movement of mouse on the horizontal axis (xFactor) would modify the magnitude of force by some fraction. Similarly, the vertical displacement (yFactor) would intervened the angel of flight by some value as shown in the Fig. 9. In the end we have presented a comparison of OpenGL and Box2D simulation loop in Table 6.

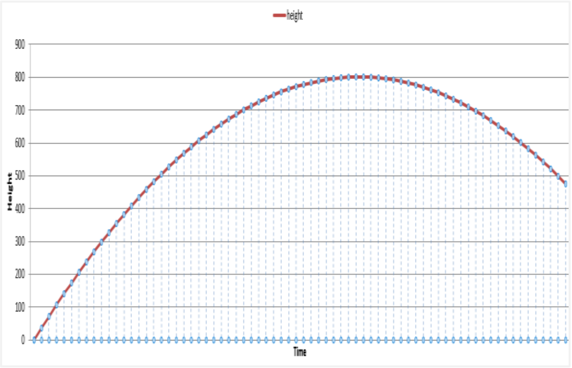


Fig. 6. Simulation Result of EQ (17) Recurrence Relation of Projectile Height as a Function of Projectile Final Velocity and Frame Per Second.

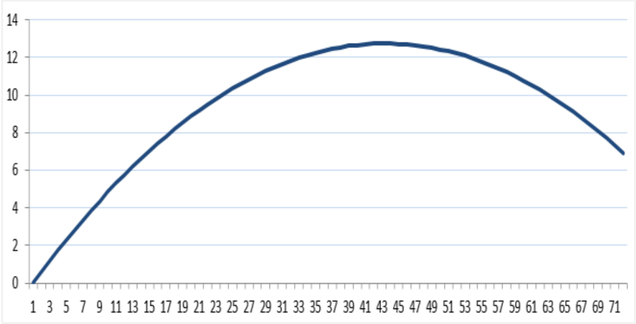


Fig. 7. Simulation Result of EQ (22) Where Flight Time Takes the form of Triangular Series and Mimics Projectile.

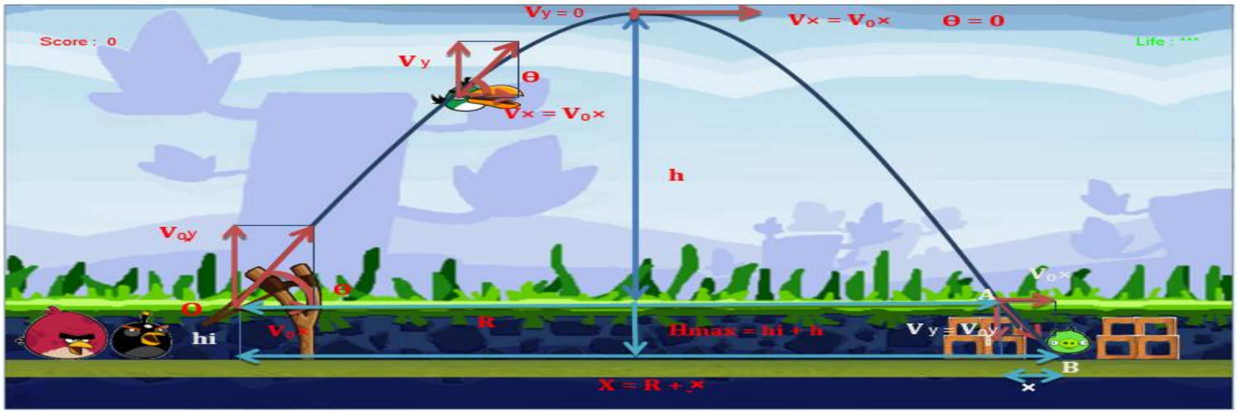


Fig. 8. Simulation Settings of Projectile Linear Drag Model using Box2D and GLUT API.

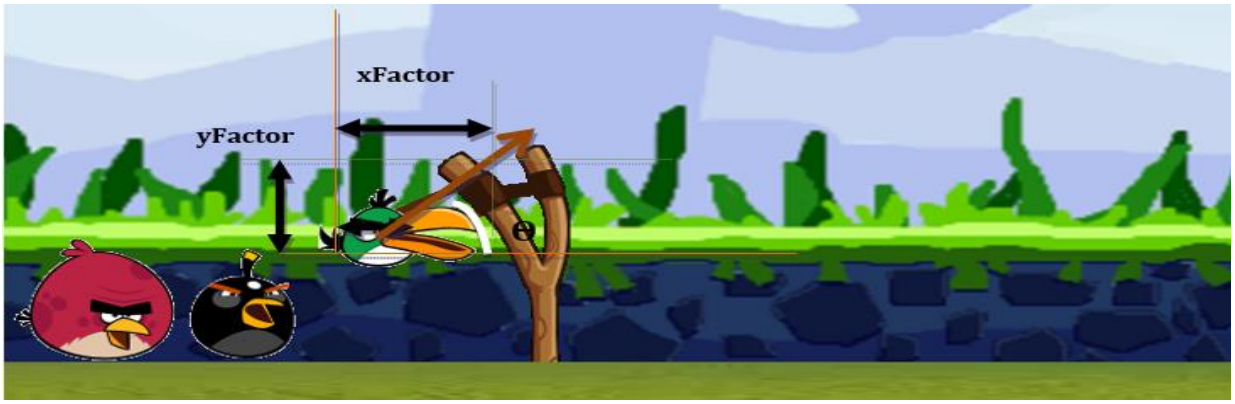


Fig. 9. Mapping of Mouse Drag Action in Game onto Impulse Force (xFactor) and Angle of Projectile (yFactor).

TABLE VI. COMPARISON OF GLUT AND BOX2D SIMULATION LOOP

OpenGL (glut) Simulation Loop	Box2D Simulation Loop
Step 1 System-dependent initialization • setup a window on the screen • bind OpenGL to this window Step 2 OGL and CS initialization • setup coordinate system • setup initial values Step 3 Begin eternal loop • wait for system to deliver event (e.g., mouse moved) • If redraw event, execute OpenGL commands to draw scene by calling Draw ()	Step 1 Creation of World • Create a gravity vector • Create Global b2world object with gravity vector Step 2 Creation of Bodies • Create body object • Insert it into the world • Create its fixture • Create a shape for the fixture • Assign the fixture to the body Step 3 Rendering of world's Bodies For all bodies n in the world { Draw n }

X. CONCLUSION AND FUTURE WORK

Replica of popular angry bird game has been successfully implemented during BSCS-Computer Graphics class project at Department of computer Science, University of Karachi. Integration of physics engine Box2D and OpenGL on Visual Studio and windows platform has been successfully achieved. Linear Drag Simulation on Spread Sheet doesn't exactly match with functionality available in Box 2D and hence It has been concluded that mathematical modeling and simulation is an essential ingredient of learning which cannot be excluded from class environment due to easy access of commercial engines and open source libraries. Instead it is very important to exercise elementary pipeline of engines as a class project. This will enhance one's programming skills and also inculcate bottom up approach for building innovative software's either on the top of existing one or by developing it from scratch. Future work will comprise of dimension analysis of drag model presented in this paper. The work is also extended by demonstration of projectile using Vulkun API instead of glut or free glut with intention of handling physics and rendering in separate threads for effective GPU utilization.

REFERENCES

[1] Semih M.Olcmen, Gray C.Cheng, Richard Branam and Stanley E Jones. "Minimum drag and heating 0.3 caliber projectile nose geometry." doi: 10.1177/0954406218779094

[2] G.P.de Carpentier, "Analytical ballistic trajectories with approximately linear drag", Int. J. Comp. gam. Tech.2014.

[3] O.A. Lasode , O.T.Popolla, Olaleya. "Modeling the Projectile Motion of a soccer ball under linear drag influence" J. Research Info. Civil Engg, Vol 6, No.2 2009.

[4] P.Coutis, "Modelling the prpjection motion of a cricket ball", Int. J. Math. Edu. Sci. Tech., vol. 29, pp.789-798, 2006.

[5] N.Azarnia, "A progression of projectiles: examples from sports", The Coll. Math. J., vol.25, no.05, pp.436-442,1994.

[6] J.Yang, G.K.W.Wong and C.Dawes, "An exploratory study on learning attitude in computer programming for the twenty-first century", N. Med. Edu. Change, pp.59-70, 2018.

[7] F.J.G.Penalvo and A.J.Mendes, "Exploring the computational thinking effects in pre-university education", Comp. Hu.Behav., vol.80, pp.407-411, 2018.

[8] M.J.Nathan, M.Wolfgram, R.Srisurichan, C.Walkington and M.W.Alibali, "Threading mathematics through symbols,sketches,software,silicon and wood: Teachers produce and maintain cohesion to support STEM integration", J. Edu. Res., vol.110, no.3, pp.272-293, 2017.

[9] M.J.Rodrigues and P.S.Carvalho, "Teaching physics with angry birds: Exploring the kinematics and dynamics of the game", Phy. Edu., pp.431-437, 2013.

[10] R.Bidarra, "Interdisciplinary game project: Opening the graphics (Back) door with the soft skills key", Edu. Paper, p.9-16, 2011.

[11] M.Shaker,N.Shaker and J.Togelius, "Evolving playable content for cut the rope through a simulation-based approach",proc.9th AAAI conf. on AI and interactive digital entertainment, pp.72-78, 2013.

[12] S.Leutenegger and J.Edgington, "A game first approach to teaching introductory programming", proc. 38th SIGCSE Technical Symposium on Comp. Sci. Edu., vol.39, pp.115-118, 2007.

[13] A.R. Shankar, "Physics Engine Basics". In: Pro HTML5 Games. Apress, Berkeley, CA, 2012.

[14] I. Parbarry, "Introduction to Game Physics with Box2D" CRC Press, 2013.

[15] D. Shreiner and OpenGL, OPENGL PROGRAMMING GUIDE, 7th ed., Addison-Wesley Professional, Aug. 2009.

[16] A. Changjan and W. Mueanploy." Projectile motion in real-life situation: Kinematics of basketball shooting" J. Phys.: Conf. Ser. 622, 2015.

[17] N.Henelsmith. "Projectile Motion: Finding the Optimal Launch Angle", Whitman college, 2016.

Embedded Feature Selection Method for a Network-Level Behavioural Analysis Detection Model

Mohammad Hafiz Mohd Yusof [1,2], Mohd Rosmadi Mokhtar[2], Abdullah Mohd. Zain[2], Carsten Maple[3]
[1] Faculty of Information Technology & Sciences, INTI International University, Nilai, Negeri Sembilan, 71800, Malaysia
[2] Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia
[3]Cyber Security Center, WMG, University of Warwick Coventry, CV4 7AL, U.K

***Abstract*—Feature selection in network-level behavioural analysis studies is used to represent the network datasets of a monitored space. However, recent studies have shown that current behavioural analysis methods at the network-level have several issues. The reduction of millions of instances, disregarded parameters, removed similarities of most of the traffic flows to reduce information noise, insufficient number of optimised features and ignore instances which are not an entity are amongst the other issue that have been identified as the main issues contributing to the inability to predict zero-day attacks. Therefore, this paper aims to select the optimal features that will improve the prediction and behavioural analysis. The training dataset will be trained to use the embedded feature selection method which incorporates both the filter and wrapper method. Correlation coefficient, *r* and weighted score, w_j will be used. The accepted or selected features will be optimised uses Beta distribution functions, *β*, to find its maximum likelihood, l_{max}. The final selected features will be trained by the Bayesian Network classifier and tested through several testing datasets. Finally, this method was compared to several other feature selection methods. Final results show the proposed selection method's performance against other datasets consistently outperform other methods.**

***Keywords*—Feature selection; intrusion detection; behavioural analysis**

I. Introduction

Behavioural analysis has become a trending research area compared to signature-based studies [1]. Computer networks in general are less studied due to the lack of leveraging behaviour of malware attacks in the network environment [2]. An article by [3] stated that behavioural-based detection methods are effective in malware detection and prediction. Meanwhile, the author of [4] describes the behavioural analysis model as being used to discover malware adaptation tactics that are difficult to understand through static signatures. These statements have led to the discussion in this paper on the existing studies in relation to behavioural-based analysis methods, specifically in the network environment.

Features selected could be different between research field like in image authentication [19] steganography [20] and wireless sensor networks [18]. Feature selection in network-level behavioural analysis studies is used to represent the network dataset of a monitored space [4]. The research of [4] used Internet Protocol addresses as a feature to represent the monitored space. The author in [5], on the other hand, used application protocol HTTP to represent his selection feature. However, recent studies have shown that current behavioural analysis methods at the network-level have several issues, such as the inability to predict zero-day attacks, high-level assumptions, non-inferential analysis, a lack of ground truth datasets, a lack of distribution modelling refinement processes and performance issues [6]. Feature selection methods give a better understanding of the dataset, prepare a framework or technique to improve prediction performance, reduce computational time, reduce the effect of dimensionality and improve prediction performance in machine learning or in pattern recognition applications [7].

However, network features are different, whereby the packets are too discrete and robust, and might therefore not be sufficiently modelled through the time of propagation. To improve the accuracy of the dataset, a certain elimination algorithm has to be applied. Removing information or instances from the network dataset will lead to inaccurate results [6]. Since suitable algorithms for extracting portions of the feature from the packets automatically is an open question [8] in research, this paper aims to select for the **optimal features** that will improve the prediction and behavioural analysis.

II. Preliminaries

Based on the above-mentioned issues, three problems are the inability to predict, high-level assumptions and non-inferential analysis. These could be further grouped into their mutually shared common criteria, summarised in the following points.

A. *Reduced Parameters (Instances), θ*

The numerical characteristics of a population are often denoted by parameter θ and the numerical description of a subset is denoted by y which is uncertain before a dataset is obtained. The level of uncertainty decreases once the dataset has been identified. Given space, Θ is set of potential parameters θ, thus $\theta \in \Theta$ (2) so that the product of all possible outcomes of parameter Θ and unknown parameters X becomes Ω denotes the universal, $\Omega{=}X.\ \Theta$ [9], thus it is important to obtain as much information about the parameters as possible to derive informative results.

As illustrated in Figure 1, conceptually, these are the building blocks of a universal set Ω which is the outcome of all possible parameters and the unknown parameters as well. Given Θ is the space of all possible parameter values θ where $\theta \in \Theta$. In the diagram, there are two sets of parameters θ_n and

θ_{an}. These sets are the element of the parameter space Θ. If a method is used to reduce or discard each of the parameter sets, it will limit the parameter or instance information which could be used to drawn further conclusions or the inability to predict unknown (zero days) attacks. This problem could be solved through the prior information.

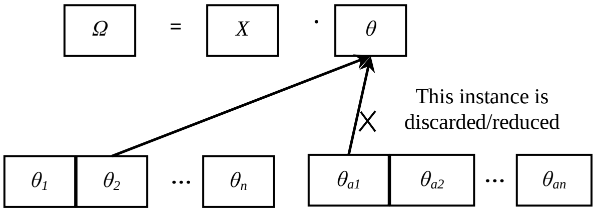


Fig. 1. Reduced Parameters or Instances Explained in Diagram.

B. Lack of Priori, $p(\theta)$

Prior distribution $p(\theta)$ explains the certainty that θ signifies the accurate population characteristics explained in [9]. As shown in Figure 2 above, it is a derivation of the previous problem. As far as the previous problem is concerned, the parameters are reduced by some reduction process or totally ignored, which could affect the results or conclusions. It was also stated earlier that the problem could be resolved by establishing prior information. The prior information, or simply priori, is done by the probabilistic method. For instance, parameters θ_n and $\theta_{an,}$ instead of being reduced or limited, have been represented by $p(\theta_n)$ and $p(\theta_{an})$ which is the notation for prior information. However, this doesn't happen in the previous method. Instead, they choose the method which ignored prior information like an in state-transition or another method that represents the collection of information of the main features in the data collection without determining inferential or in-depth analysis. It can also involve simply assuming the probability of the parameter occurrences. This leads to high-level assumptions and non-inferential analysis problems.

In a volatile or in a critical infrastructure network environment such as in the energy industry, the lack of prior information could cause a catastrophic false alarm as happened in the history of Iranian nuclear plant, in a Saudi Aramco oil and gas plant and in the healthcare industry. A lack of information capabilities could lead to the breach of patient information and malware attacks as happened during the WannaCry malware that attacked hospitals, mostly in Europe.

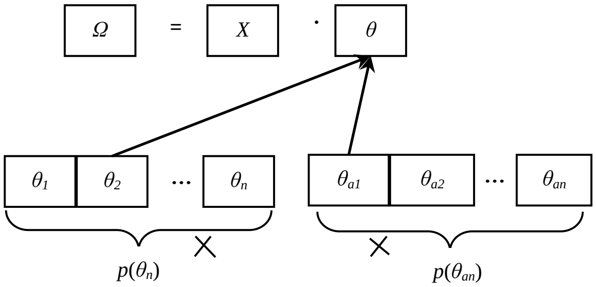


Fig. 2. Lack of Priori Information Explained in Diagram.

Due to the **reduced instances issue and data normalisation practices,** and since the malware analysis in computer networks in general are less studied due to the lack of leveraging behaviour of the malware attacks in the network environment as mentioned by the author in [2], this paper has proposed a feature selection method and process flow that suits the complexity and traffic in networks which use the embedded feature selection method and optimised beta distribution function.

III. Feature Selection

Normally, feature selection starts with a pre-processing phase like for instance in the work by [10]. They had their dataset processed early for it to be later represented it in a vector of real numbers. Then the data was normalised, selected and classified. The data acquired through the data collection stage was firstly analysed to produce the elementary instances or features. The whole processes contained three main stages which were pre-processing, feature selection and classification. This payload was loaded for pre-processing. The data was presented as a real number vector. Thus, every symbolic feature or nominal feature was converted into numerals. In the NSL dataset, the nominal feature included a protocol of type UDP, TCP and ICMP and the service protocol of type FTP, HTTP and telnet.

The dataset was then analysed using feature selection and finally classified by the chosen classification method. The feature selection method gives us a better understanding of the dataset, prepares a framework or technique to improve **prediction performance**, reduces computational time, reduces the effect of **dimensionality** and improves the prediction performance in pattern recognition or machine learning [7].

As defined in [11], feature selection methods were classified into wrapper and filter methods. Finally, the embedded methods combine both the filter and wrapper method, and include feature extraction by means of integral phase of the training procedure deprived of disjointing the dataset into testing datasets or training datasets. On this research, we will apply the proposed feature selection method on a supervised dataset.

A. Filter Method

The filter method uses variable ranking methods as the basis for its variable selection criteria. The ranking method is used mainly because of its uncomplicatedness, simplistic approach and its success history in recording certain pragmatic applications. Unique features contain useful and relevant information about the property of the dataset or instances [7]. This relevant and useful property is used to measure [11] the usefulness of the feature compared to other feature, and finally to discriminate it from that other feature's label. For instance, one of the criteria of the simplest principle is the correlation coefficient, also known as Pearson correlation. Correlation coefficient ranking is able to identify linear dependencies between the target and the variable. The Pearson **correlation coefficient**, *r* is defined as below.

$$r=\frac{1}{n-1}\Sigma\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_x}{\sigma_y}\right) \quad (1)$$

Where *n* signifies the full training set number whereby set, x_i indicates the *i*th variable of feature *x*. Meanwhile σ_x and μ_x are the standard deviation and mean aimed at feature *x* respectively. Whilst y_i could be the label or other correlated features or to test dependency features in the dataset whereas σ_y and μ_y are the standard deviation and mean aimed at feature *y* respectively.

The above equation is to determine the product of *z*-value of both features. *Z*-value is to determine how a single instance of a set of features is positioned from its mean and its standard deviation. So any -1 ≤ *r* ≤ 1 value that is drawn towards positive 1, we could conclude that there is positive correlation.

Network traffic are usually **linearly dependent** on each other as discussed by [8], such that some novel cyber or network attacks are variations of the previous identified attacks and its signature could be sufficient to detect and prevent some other novel variants. Coefficient correlation is suitable for processing network flow. They also explained that **some probing attack scans are correlated** with a much larger time scanning interval compared to normal traffic. However, the correlation ranking is only able to detect linear dependencies between the target and the variation.

B. Wrapper Method

Unlike filter methods (FM) which use the ranking method as the criterion for its relevance feature, the wrapper method (WM) on the other hand relies on the classifier or **classification method** for obtaining a feature or instance subset.

Therefore, the simplified version of algorithms for instance, sequential searching algorithm or evolutionary algorithm such as Genetic Algorithm (GA) or Particle Swarm Optimization (PSO) that will harvest local optimum outcomes. They are applied as they can generate good computationally feasible results.

Wrapper methods can be divided into Heuristic Search Algorithms and Sequential Selection Algorithms. Sequential Selection Algorithm are named as it is because its algorithm is designed as iterative in process. It starts with a full dataset and in the process, the features are removed till the maximum objection function has been gained. On the other hand, it begins with an empty set and throughout the process, the features are added until they reach the maximum.

On the other hand, heuristic search algorithm is about reaching the local optimum results by applying an evolutionary algorithm such as a Genetic Algorithm (GA). GA (31) is used to select the features whereby the chromosome bits are used to denote the selected features. It is based on the natural selection theory by Darwin. Searching the GA provides both data exploration and data exploitation.

$$f = \{\sum_{\mathrm{i}=1}^{n} c_i v_i \,, \mathrm{if} \sum_{\mathrm{i}=1}^{n} c_i\, w_i \leq k\ .0\ otherwise \tag{2}$$

Particle Swarm Optimisation (PSO) assumes a “swarm” of *N* particles (32). General Particle Swarm Optimization algorithm is simple. PSO is initialised with a group of random solutions or particles, which is then searched for by learning of the next generations. Particles will swarm throughout the space, and are tested or evaluated across the fitness criterion. In each iteration, the particles will be updated by following two “best” values. The below equation represents the PSO algorithm.

$$v_{n+1} = v_n + c_1 rand1(\,) * \left(p_{best,n} - currentPosition_n\right) + c_2 rand2(\,) * g_{best,n} - \ currentPosition_n \tag{3}$$

Where, v_{n+1} is the velocity of particle at the *n+1*th iteration and v_n is the velocity of particle at *n*th iteration. C_1 is acceleration factor related to the *gbest* and C_2 is the acceleration factor related to *lbest*. *rand1*( ) and *rand2*( ) is the random number between 0 and 1.

The main disadvantage of a wrapper method is that it requires a number of computational processes in order to obtain the final feature. Having said that, for instance, if the dataset sample is large, then most of the execution of the algorithm will be allocated to train the predictor. Note that our research will reduce this by calculating the optimised value in the ranking process. In the next section, this paper will elaborate on the embedded methods to then try to leverage the drawbacks or disadvantages found in the Wrapper or Filter methods.

C. Embedded Method

The main purpose of the embedded methods [12,13,14] is to lessen the computational time that is used to reclassify the dissimilar subsets which completed in WM. This is done by incorporating the feature selection process in the FM as part of the training process [7]. The main approach is to incorporate FM and WM.

For instance, a method was to use the weights of a classifier to remove the feature based on the rank [12,15]. For example, let w_j be denoted as

$$w_j = \frac{\mu_j(+) - \mu_j(-)}{\sigma_j(+) + \sigma_j(-)} \tag{4}$$

Where μ_j (-) and μ_j (+) and are the mean of samples in class + and class – and σ_j is the variance of the respective classes and *j*=1 to D. Equation 13 can be used as a ranking criterion to sort the features. The rank vector *w* can be used to classify since the features rank proportionally. This contributes to the correlation. Another weighted score is the true normal score, whereby in order to create a normal profile, it is necessary to index each attributes’ instances as *i*=1,2…*n*. The model was build based on the ratio of the normal number of training data, R_i against the total number of packets associated with each attribute, N_i. The probability of the normal score, P_i=R_i/N_i is represented by

$$Pi = \sum_{i=1}^{n} \frac{R_i}{N_i}, i = 1,2,3 \dots, n \tag{5}$$

IV. Dataset

Table 1 shows the basic features of Network Socket Layer (NSL) dataset which is an updated version of the KDD Cup 1999 data set. The KDD Cup 1999 dataset was used for a data mining completion which was organised in conjunction with the Fifth International Conference on Knowledge Discovery and Data Mining. During the competition, the challenge was to design a **predictive model** or network **intrusion detector** or that was able of differentiating between attack connections or intrusions, and baseline connections. This dataset contained standard data to be analysed, which also included varieties of computer-generated intrusion scenarios in a military network environment, specifically simulating LAN connectivity of U.S Air Force [8].

However, from a network practitioner's point of view, the KDN or NSL datasets are **not realistic** and do not reflect modern attacks, and not even attacks back in 1998 [16]. Today's attacks are primarily SQL injections. The KDN dataset was also focused around attacks with some background noise, while the actual traffic was largely data. Furthermore, it was a simulated dataset within a large virtual network.

To apply objectivity, in this research, final classification method using Bayesian Network will be applied over a ground-truth dataset. That ground-truth dataset or simply raw dataset was obtained from the local asset, which was a host tagged to among the largest healthcare provider in Malaysia. This is more adequate to strategize the scan rate of one-to-one modelling. The traffic is more resemblance to one-to-one connection. One to one model is to mimic a connection of a single infected machine that is transacted throughout the network.

TABLE I. BASIC FEATURES OF INDIVIDUAL TCP CONNECTIONS

Feature Name	Description	Type
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of ``wrong" fragments	continuous
urgent	number of urgent packets	continuous
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete

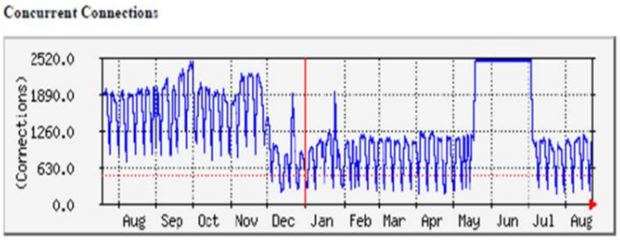


Fig. 3. Ground-Truth Dataset from Asset (Internet Load Balancer) that is Tagged to the Healthcare Provider in Malaysia.

Traffic captured and monitor from August 2016 until August 2017 and traffic information was extracted from the Cisco FMC, Steal Head Riverbed WAN optimizer and Hgiga internet load balancer network appliances as [4] stated that network activity profiles of infection network environment is dependent on both distribution of activity across internet and malware propagation techniques or targets that might differ from each network population profiles.

Figure 3 below manifests definition laid by [4] which indicates traffic distribution activity across the internet could potentially highlight the malware propagation activity. It shows some scanning activity happened from early May 2017 until end of June 2017 and after that period the counts appeared to decline to a baseline level towards the end of data collection period.

IT personnel from the healthcare provider confirmed that during that period Trend Micro DDAN (Virtual Analyzer) has sent a lot of suspicious object (SO) information to the Trend Micro OS indicated some malicious activity and Trend Micro CM has pushed latest signature to all the endpoints to disinfect the malware attacks.

Evaluation to demonstrate that malware really propagated during this period from May 2017 to July 2017 is presented in the following subsection report. The healthcare equipped with Trend Micro DDAN report which trapped all suspicious object in the environment and analyze it in their Virtual Analyzer (VA).

V. Methodology

Figure 5 depicted the whole proposed method process flow was based on work by [10], whereby the training set was first processed in the pre-processing phase. In [10], the pre-processing was conducted to differentiate between normal and attack traffic. The data was then weighted by normal function whereby the nominal data was converted into numeric data, followed by calculating its normal function. The distribution of the numerical data was determined by its normal function. The whole process is known as data normalisation. Instances will be greatly reduced during this process and it affects the classification process (detection or prediction), as proven in the results and discussion section. In the proposed method, pre-processing was done by applying equation whereby, each features' instances were converted into its normal score and normal traffic was compared to produce a **baseline** value and attack traffic. This was compared to produce a **threshold** value.

The next phase is followed by the feature selection process. In [10], the feature selection used flexible MI, which was suggested to select the feature by argmax. In the proposed feature selection method, equation 3 (***Eq.*** 3) which is for numerical instances and equation 13 (***Eq.*** 13) which is for nominal instances, was applied and the instances that can be manageably optimised by the modelled optimisation function and distributed by the generic Beta function, β of which its maximum likelihood, l value will be selected. The results show that by processing the instances optimisation value and its beta distribution function, the features selected are significantly minimum yet the detection accuracy is very high compared to the previous method. The false alarm rate has also been reduced.

Phase 1: Pre-processing modelling

B, packet capture or space for baseline traffic and

B=b= $\{b_1, b_2, \dots b_n\}$ (6)

C, packet capture or space of attack traffic

C=c= $\{c_1, c_2, \dots c_n\}$ (7)

The above is the representation of the raw dataset for both the attack and baseline traffic. Equation (6) and equation (7) were applied in equation (8), which produced a new equation.

$$Pb, c = \sum_{b,c=1}^{n} \frac{R_{b,c}}{N_{b,c}}, b, c = 1,2,3 \dots, n \quad (8)bc$$

Where $P_{b,c}$ is the normal score for dataset b and c. The output from this process is the dataset that will be labelled as numerical or nominal and a change of notation for instance f_1 to indicate feature number 1. Equation 8b,c will be used to represent the nominal data for future feature selection processes.

Phase 2: Embedded feature selection modelling

The embedded method incorporated both the Filter (FM) and Wrapper method (WM). As in the work done by [10], the selected FM method used a correlation coefficient, r which is good to process multidimensional data with multi-array instances. In this work, it was used to process the numerical types of dataset. The formula given is:

$$r_{b,c} = \frac{1}{n-1} \Sigma \left(\frac{x_{bi}-\mu_{xb}}{\sigma_{xb}}\right)\left(\frac{y_{ci}-\mu_{xb}}{\sigma_{yc}}\right) \quad (1)bc$$

Whereby r_{bc} is the coefficient value for both dataset b and c. Equation (1)b,c is formulated by applying equation (4) and (5) into equation (1). In [10] Any r<0 will be rejected. However, in the proposed method, any **r>0** will be rejected. Then for nominal dataset, weighted score, w_j will be used for feature selection processing. Any $w_j<0$ will be rejected. Then the selected features will be optimised to avoid optimVal (optimal value) errors in the Beta distribution function in order to find the likelihood value. The optimisation formula has been given below.

Baseline_f_i_beta =

$10^{-x} \sum i=1$ to n , (-) $\log_{10} Baseline_{fibeta}$ (9)

Where $-x$ is the power value for absolute value 10, which is to avoid the instances exceeding 1.0 which could produce optimVal error and m must be between $0<m<1$. Beta function, β is given by the formula below.

Beta ~ (λ_{fi} ; α, β)=

$$\frac{1}{Beta\ (\alpha,\beta)} . \lambda_{fi}{}^{\alpha-1}\left(1-\lambda_{fi}\right).^{\beta-1}\ , where\ 0 < \lambda_{fi} < 1 \quad (10)$$

Whereas maximum likelihood, l function in the form of log function is given by the formula below.

Maximum likelihood, ℓ_{fi_beta}

$= \ln [\sum_{fi=0}^{n} \frac{1}{Beta(\alpha,\beta)} . \lambda_{\mathrm{fi}}{}^{\alpha\text{-}1} (1\text{-} \lambda_{\mathrm{fi}})^{\beta\text{-}1}$,

where

$0 < \lambda_{fi} <1]$=

$(\alpha\text{-}1) \sum_{fi=0}^{n} \ln(fi) + (\beta\ \text{-}1) \sum_{fi=0}^{n} \ln(1\text{-}fi)\text{-}N \ln\mathrm{Beta}(\alpha, \beta)$ (11)

Where N, is the total number of i.i.d observations

Phase 3: Classification modelling

This classification was based on the work by [4] on Naïve Bayes. However, we modified it to **incorporate the Bayesian Network approach**. Bayesian Network as depicted in Figure 4, is approach that were classified under probabilistic theorem and being highly chosen is mainly due to their flexibility and ability to model uncertain events such as the Bayes theorem which has been considered as the state-of-the-art technology [6]. Because of the intuitive ability to model uncertainty and complex chronological relationships amongst variables, Bayesian network is successfully applied in several research areas and domains [17]. State-of-the-art predictive analytics method of uncertainty and the detection of the unknown, using the Bayesian Network method, have been proven in other research areas, especially in the domain of Clinical Expert System studies, Artificial Intelligence (AI) and Pattern Recognition. Modelling the classification based on Bayesian Network has been given below.

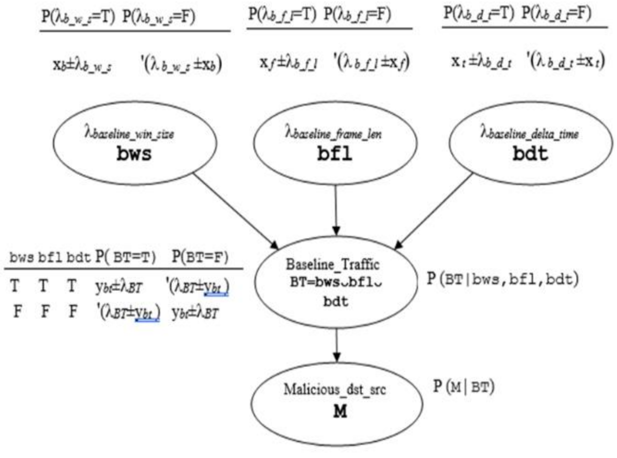


Fig. 4. Bayesian Network Classification Model.

The model is then written in its conditional probability as derived in the following forms. This model is supplied to the classifier.

$$Pr(f_i) = \lambda_{fi} \text{ where } i{=}1\ldots.j\ ,\ 1\ldots.k\ ,\ 1\ldots.l \qquad (12)$$

$$Pr(f_i \ldots j,k,l) = Pr(\lambda_{fi}\ldots j).\ Pr(\lambda_{fi}\ldots k).\ Pr(\lambda_{fi}\ldots l) \qquad (13)$$

$$Pr(Malicious, M) = Pr(\lambda \mathrm{M}).\ Pr(`\lambda \mathrm{M}) \qquad (14)$$

$$Pr(f_i\ldots j,k,l|\ f_i) = Pr(\lambda_{fi}\ldots j.\ \lambda_{fi}).\ Pr(`\lambda_{fi}\ldots j.\ \lambda_{fi}) \qquad (15)$$

$$Pr(M \mid f_i\ldots j,k,l|\ f_i\ ) = \lambda_M\ .\ \lambda_{fi}\ldots j\ .\ \lambda_{fi}$$

Equation. (13) in *Equation.* (14)

VI. Results and Discussions

In this section, we have discussed the results of applying the proposed Embedded Feature Selection model to the KDD dataset using the Bayesian Network classifier. We implemented two feature selection methods as shown in phase 2 in the methodology section, whereby the features training dataset and variations of NSL and also the KDD dataset in the later operation was used in the testing dataset as well. Comparisons between the selection algorithm could only be done using a single dataset and the selection techniques indicated that more feature or instance information is not always good in the context of machine learning applications [7]. Table 2 below shows the training dataset descriptions.

TABLE II. Training Dataset Descriptions

Type of packet	Number of packet	Total packet
Normal	9711	22544
Attack	12834	22544

A. Feature Selection Results

The features selected were still on 1-dimensional with 2 arrays of data. This means that the feature is the same attribute but constructed in 2 arrays of information.

For duration, *f1* correlation coefficient, the r_{f1} score was -0.009742335. This indicates a negative correlation. Correlation ranking can only detect linear dependencies between the variable and the target. Hence, in the case any $-1 \leq r \leq 1$ value that is drawn towards positive 1, we can conclude that there is a positive correlation. Negative correlation means that there is no linear dependency between the two datasets. Thus, the duration of traffic transactions between normal and attack traffic has no correlation and no dependency. In this case, $r<0$ will be rejected. Note, however, that in some models, it will be accepted as no correlation, which means that a threshold could be constructed.

The results of the entire feature selection process have been summarised in the following table 3.

TABLE III. Feature Selection Process Summary

Features	Correlation score, r_{fi}	Weighted score, w_{fi}
f_1	-0.009742335	-
f_2	0.976564	-
f_3	0.026048497	-
f_4	0.781732335	-
f_5	-0.000713582	0.019534
f_6	0.002013838	0.098281
f_7	-	0.023361
f_8	-0.00419	0.028317
f_9	-	0.025794

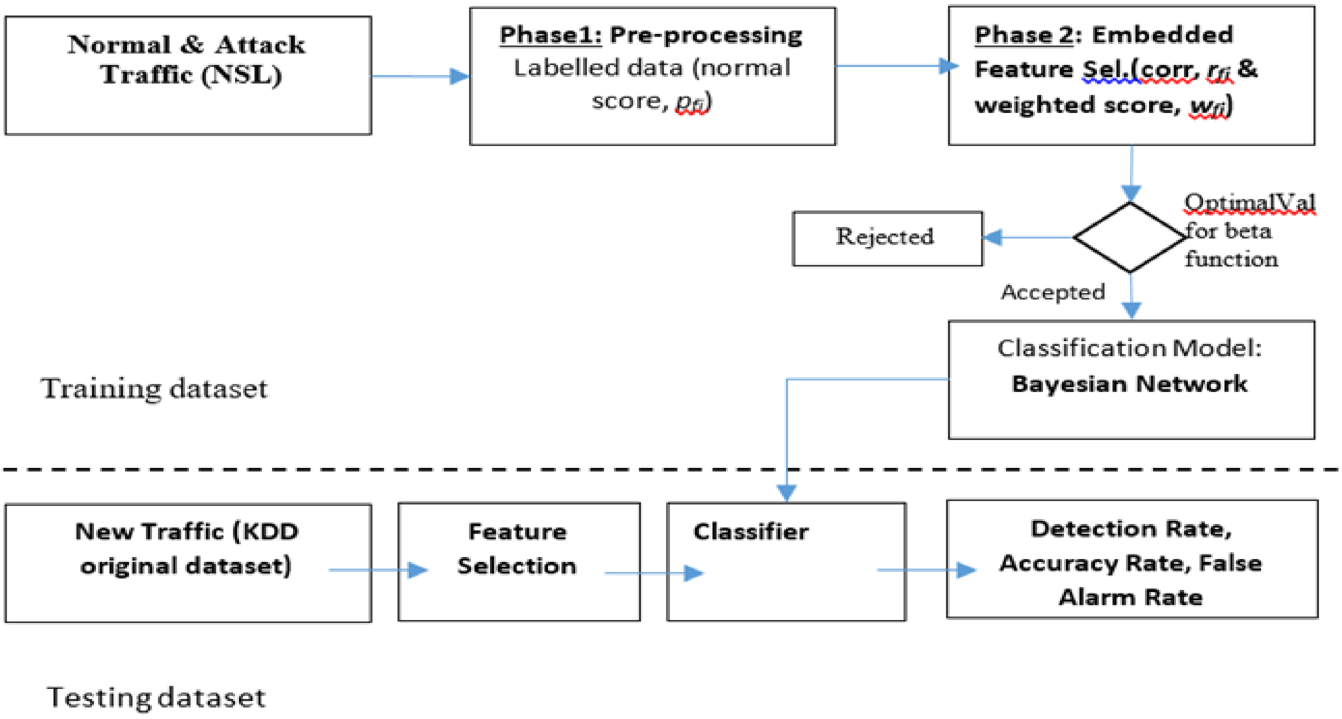

Normal & Attack Traffic (NSL)
Phase1: Pre-processing Labelled data (normal score, pfi)
Phase 2: Embedded Feature Sel.(corr, rfi & weighted score, wfi)
Rejected
OptimalVal for beta function
Accepted
Classification Model: Bayesian Network
Training dataset
New Traffic (KDD original dataset)
Feature Selection
Classifier
Detection Rate, Accuracy Rate, False Alarm Rate
Testing dataset

Fig. 5. Proposed Feature Selection Workflow.

TABLE IV. FEATURES ACCEPTED OR REJECTED DURING FEATURE SELECTION PROCESS

Feature selection approach	Number of features	Features Selected
Original Features	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	-
Weighted score *w*>0	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	f_5, f_6, f_7, f_8, f_9
Weighted score *w*<0	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	-
Correlation coeff. *r*>0,	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	f_2, f_3, f_4, f_6
Correlation coeff. *r*<0	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	f_1, f_5, f_8
Maximum likelihood	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$	f_1, f_5, f_6

Finally, based on the feature selection model, table 4 below shows the list of features that were selected or rejected during the process.

B. Classification Rresults

Only the following features in table 5 below were selected after optimisation, distributed using the beta function. The maximum likelihood features will be selected.

The performance of this selected feature over its classification model was based on the true positive (TP value), true negative (TN value), false positive (FP), false negative (FN), detection rate (DR), accuracy (ACCR) and false alarm rate (FAR).

Detection rate, on the other hand, is used to measure true positive traffic over the sum of true positive and false negative (positive traffic wrongly classified as negative). The formula is the following

$$Detection\ Rate, DR = \frac{TP}{TP+FN} \tag{16}$$

Accuracy is used to measure all true traffic which consists of the sum of the true positive and true negative over the sum of all traffic of a true positive, true negative, false positive and false negative nature. The formula is denoted as the following.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

TABLE V. OPTIMISED FEATURES

λduration , f1	Shape 1: 0.3601539 Shape 2: 533.8212008 Loglikelihood:66584.47	Mean (optimized prior), $\lambda_{duration} = \frac{\alpha}{\alpha+\beta}$ = 0.000674
λsrc_bytes , f5	Shape 1: 0.5537134 Shape 2: 1347.6133442 Loglikelihood: 67446.38	Mean (optimized prior), $\lambda_{src\_bytes} = \frac{\alpha}{\alpha+\beta}$ = 0.000411
λdst_bytes, f6	Shape 1: 0.7658164 Shape 2: 1466.8806705 Loglikelihood:63927.07	Mean (optimized prior), $\lambda_{dst\_bytes} = \frac{\alpha}{\alpha+\beta}$ = 0.000522

TABLE VI. KDDTRAIN+_20PERCENT CLASSIFICATION TABLE

Lamda information	Baseline	KDDTest+_20Percent	Ratio (dataset over attack)	Threshold (spike/attack)	Threshold (ratio normal over attack)	Difference
Pr(λduration)	47.07	4507	0.351	12833	0.00366	0.347
Pr(λsrc_bytes)	2530.77	10973.087	0.671259475	16347.0134	0.15481562	0.516443855
Pr(λ dst_bytes)	4165.553553	957.895274	2.082316478	460.0142601	9.055270487	-6.972954009
Pr(*BT*)	496249242.3	47373387043	0.490903953	96502354064	0.005142354	0.485761599
Pr(*BT*|λ duration)	23360051486	2.13512E+14	0.172407396	1.23841E+15	0.0000188629	0.1723885332
Pr(*BT*|λ src_bytes)	496251773.1	5.19832E+14	0.32952393	1.57753E+15	0.0000003146	0.3295236151
Pr(*BT*|λ dst_bytes)	2.06715E+12	4.53787E+13	1.02221739	4.43925E+13	0.0465654041	-0.9756519855

Finally, the false alarm rate (FAR) is used to measure the false positive alarm, which means the negative traffic that was wrongly classified as positive. This is a very serious issue because it may cause an attack vector. The formula is denoted as the following.

$$False\ Alarm\ Rate, FAR = \frac{FP}{FP+TN} \tag{18}$$

The above table 6 shows the differences between the two dimensional information of each features' **lambda, λ information**. It is generated from the differences between the ratio value of the testing dataset, in this case KDDTest+_20Percent, and attack traffic over the ratio of the normal dataset over attack traffic.

For NSL-40% dataset, *Pr*(λ duration) training ratio exceeded the baseline threshold by 0.699, which indicates that this is attack traffic. *Pr*(λduration) differences, this time, had increased almost 50% from the previous dataset. This may be due to a 20% increase in the traffic. Out of that, only 0.04% of this traffic was flagged as normal. Thus, the entire dataset was still attack traffic and was flagged as negative tuple or *TN*, the same as in the previous dataset. It was then a true alarm or *TP* because the alarm truly reflected the tuple condition.

For the train dataset, the *Pr*(λ duration) training ratio, this time, never exceeded the baseline threshold. It scored below the threshold by -0.000305238, which indicates that this is a normal traffic. Thus, the entire tuple will be flagged as positive tuple or *TP*. It will then be alarmed as a true alarm or *TP* because the alarm truly reflected the tuple condition, as a true positive.

TABLE VII. PROPOSED METHOD PERFORMANCE AGAINST OTHER DATASET

Dataset	Desc.	Detection Rate	Accuracy	False Alarm Rate (FAR)
NSL-40%	Attack flow (0.04% normal)	**100%** (1.0)	**86%** (0.857143)	**14%** (0.142857)
KDDTest+_20Percent	Attack flow	**100%** (1.0)	**86%** (0.857143)	**14%** (0.142857)
KDD-Train+	Normal flow	**86%** (0.857143)	**86%** (0.857143)	**0%**

Table 7 shows above the proposed method performance against other dataset. For instance, NSL-40%, the detection rate was 100% because the classification model successfully classified all alarms as attack traffic even though 0.04% of the flow was normal traffic. Only one tuple *Pr*(λdst_bytes) was flagged as positive which is correct, however the intersection probability *Pr*(*BT*|λdst_bytes) was actually negative. This is a false positive alarm, whereby negative traffic was alarmed positive. Hence, it affects the *FAR* and accuracy as well, which scored 14% and 86% respectively. This is a serious failure, however, due the intersection probability that is included in this model, this flag could be re-examined and re-flagged to the correct alarm.

Finally, table 8 above shows proposed method performance against other feature selection method. Example, for the method that uses correlation coefficient, r whereby the accepted feature selection was when *r*>0, most features were nominal features. The features need to be changed into a numerical dataset and afterwards, distributed using data normalisation. When this happens, most of the instances of the features will be altered and reduced in dimension or volume. Thus, can be seen the poor result obtained, especially in relation to the false alarm, whereby 89% of the detections were false alarms. The feature selection process was validated 5 folds.

Then we apply the classification model to predict zero - day attack in the ground truth dataset mentioned before as depicted in Figure 6 below. Two months' traffic prior the attack was sampled to determine the detection rate of the model.

From Figure 7 below, it is obvious, that the proposed Predictive analytics model has accurately detected a zero-day attack a few months' prior the actual attack. In October 2016, the model was already able to detect almost 60% of the traffic was prepared to the zero attack with 75% accuracy. In January 2017, 5 months before the attack, the model has detected 86% of the traffic was directed towards the attack and this time with 100% accuracy.

TABLE VIII. PROPOSED METHOD PERFORMANCE AGAINST OTHER FEATURE SELECTION METHOD

	KDDTest+_20Percent dataset		
Feature Selection methods	Detection Rate	Accuracy	False Alarm Rate (FAR)
Weighted score *w*>0	**27%** (0.272727)	**27%** (0.272727)	**72%** (0.727273)
Weighted score *w*<0	N/A	N/A	N/A
Correlation coeff. *r*>0	**11%** (0.111111)	**11%** (0.111111)	**89%** (0.888889)
Correlation coeff. *r*<0	N/A	N/A	N/A
Proposed Embedded method (Optimized)	**100%** (1.0)	**86%** (0.857143)	**14%** (0.142857)

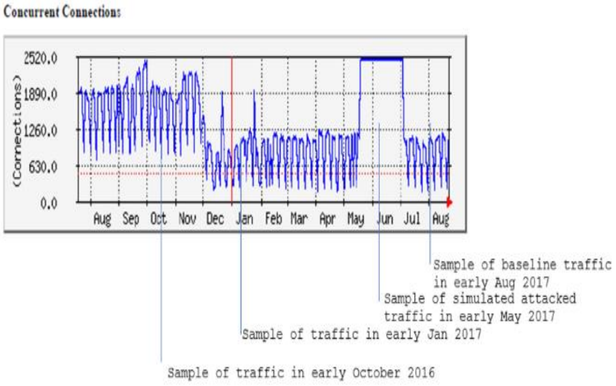


Fig. 6. Sampled Traffic.

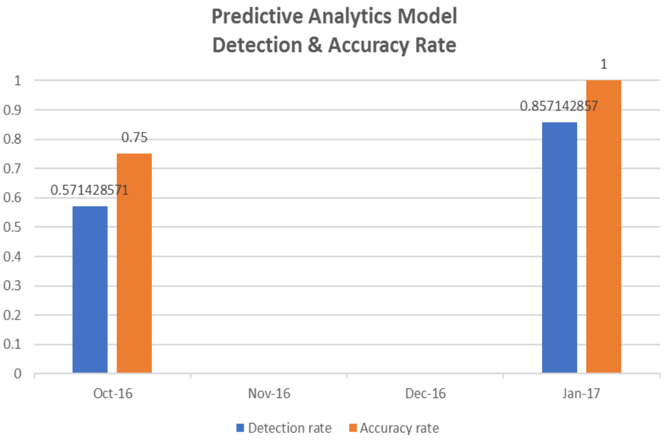


Fig. 7. Zero-day Prediction Shows the Detection Reaches 86% Detection with 100% Accuracy 5 months' Prior the Attack.

VII. CONCLUSION

Improve prediction and behavioural analysis. The training dataset will be trained to use the embedded feature selection method which incorporates both the filter and wrapper method. The correlation coefficient, *r* and weighted score, w_j will be incorporated. The accepted or selected features will be optimised using the Beta distribution function, *β*, to find its maximum likelihood, l_{max}. Finally, the selected features will be trained by the Bayesian Network classifier and will be tested through the inclusion of several testing datasets. Finally, this method will be compared to other feature selection methods. The results show that the proposed method's performance against other methods consistently outperforms other feature selection method. The detection rate for both NSL and KDDTest20% datasets was 100%, while KDD-Train+ scored 86%. This is because one of the tuple *Pr*(λdst_bytes) was flagged as positive which is correct. However, the intersection probability *Pr* (*BT*|λdst_bytes), or the baseline traffic given the lamda information, λdst_bytes, was actually negative. There was some reduction in the rate, otherwise it would have scored 100% as well. The False Alarm Rate was 14%, however, due to the **intersection** probability that was included in the **model**, this flag could be re-examined and re-flagged to the correct alarm. On the other hand, the detection rate and accuracy rate for the proposed optimised feature selection method scored 100% and 86%, which outperformed the other models.

Results applied onto ground-truth dataset also indicated that the prediction reaches 86% detection with 100% accuracy 5 months' prior the attack.

REFERENCES

[1] M. H. M Yusof and Mokhtar M. R, "Review on Taxonomy of Malware Analysis Studies". Advanced Science Letters. 2018. Vol. 23 Issue 12.

[2] S. G. Nari. "Automated Malware Classification based on Network Behavior." 2013 International Conference on Computing, Networking and Communications, Communications and Information Security Symposium

[3] L. Xue and G. Sun. Design and Implementation of Malware Detection System based on Network Behavior. SECURITY AND COMMUNICATION NETWORKS. 2015. doi: 10.1002/sec.993

[4] R. Weaver. Visualizing and Modeling the Scanning Behavior of the Conficker Botnet in the Presence of User and Network Activity. IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. doi: 10.1109/TIFS.2015.2396478

[5] M. Zaman, T. Siddiqui, M. R Amin and M. S Hossain. Malware Detection in Android by Network Traffic Analysis. Next Generation Mobile Apps, Services and Technologies (NGMAST), 66- 71. doi: 10.1109/NGMAST.2014.57

[6] M. H. M Yusof, "A Review of Predictive Analytic Applications of Bayesian Network," International Journal on Advanced Science, Engineering and Information Technology., 2016. 6(6) ISSN: 2088-5334.

[7] G. Chandrashekar and F. Sahin. A Survey on Feature Selection Methods. Journal of Computers and Electrical Engineering 40 (2014)16-28.

[8] S.J. Stolfo ; Wei Fan ; Wenke Lee ; A. Prodromidis ; P.K. Chan.Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project. DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings. DOI: 10.1109/DISCEX.2000.821515

[9] T. Koski and J. M Noble. Bayesian Networks. United Kingdom: John Wiley & Sons, Ltd. 2009.

[10] M. A. Ambusaidi, H. Xiangjian, N. Priyadarsi and T. Zhiyuan. Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm. IEEE Transactions on Computers (Volume: 65, Issue: 10, Oct. 1 2016).

[11] R. Kohavi and G. H. John. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157– 82.

[13] P. Langley. Selection of relevant features in machine learning. In: AAAI fall symp relevance; 1994.

[14] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–70.

[15] P. A. Mundra and J. C Rajapakse. Svm-rfe with mrmr filter for gene selection. IEEE Trans Nanobiosci 2010;9.

[16] A. Mousse. "KDD1999 dataset Features explanations". April 28, 2018. [Online].Available https://stackoverflow.com/questions/17024961/kdd1999-dataset-featuresexolaination?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa

[17] A. Feizollah, N. B. Anuar, R. Salleh, F. Amalina. Comparative study of k-means and mini batch kmeans clustering algorithms in android malware detection using network traffic analysis. 2014 International Symposium on Biometrics and Security Technologies (ISBAST),2014; 193 - 197. doi: 10.1109/ISBAST.2014.7013120

[18] A. Z. Ariffin and H. Song, "Secure Knowledge and Cluster-Based Intrusion Detection Mechanism for Smart Wireless Sensor Networks," in IEEE Access, vol. 6, pp. 5688-5694, 2018. doi: 10.1109/ACCESS.2017.2770020

[19] A. Nadeem, Shukur. Z, Sulaiman. R, Current techniques in JPEG image authentication and forgery detection, Journal of Engineering and Applied Sciences.104-112. 2017.

[20] Ahmed Ali, Mohd Rosmadi Mokhtar, Loay Edwar George, 2017. Enhancing The Hiding Capacity of Audio Steganography Based On Block Mapping. Journal of Theoretical and Applied Information Technology, 1141-1148.

Solving Dynamic Programming Problem by Pipeline Implementation on GPU

Susumu Matsumae[1]
Department of Information Science
Graduate School of Science and Engineering
Saga University
Saga, Japan

Makoto Miyazaki[2]
Department of Information Science
Graduate School of Science and Engineering
Saga University
Saga, Japan

***Abstract*—In this paper, we show the effectiveness of a pipeline implementation of Dynamic Programming (DP) on GPU. As an example, we explain how to solve a matrix-chain multiplication (MCM) problem by DP on GPU. This problem can be sequentially solved in $O(n^3)$ steps by DP where n is the number of matrices, because its solution table is of size $n \times n$ and each element of the table can be computed in $O(n)$ steps. A typical speedup strategy for this is to parallelize the $O(n)$ step computation of each element, which can be easily achieved by parallel prefix computation, i.e., an $O(\log n)$ step computation with n threads in a tournament fashion. By such a standard parallelizing method, we can solve the MCM problem in $O(n^2 \log n)$ steps with n threads. In our approach, we solve the MCM problem on GPU in a pipeline fashion, i.e., we use GPU cores for supporting pipeline-stages so that many elements of the solution table are partially computed in parallel at one time. Our implementation determines one output value per one computational step with n threads in a pipeline fashion and constructs the solution table totally in $O(n^2)$ steps with n threads.**

***Keywords*—Dynamic programming; pipeline implementation; GPGPU**

I. Introduction

In this paper, we show the effectiveness of a pipeline implementation of Dynamic Programming (DP) on GPU. As an example, we explain how to solve a *matrix-chain multiplication (MCM)* problem [1] by DP on GPU. This problem can be sequentially solved in $O(n^3)$ steps by DP where n is the number of matrices, because its solution table is of size $n \times n$ and each element of the table can be computed in $O(n)$ steps. A typical speedup strategy for this is to parallelize the $O(n)$ step computation of each element, which can be easily achieved by parallel prefix computation, i.e., an $O(\log n)$ step computation with n threads in a tournament fashion. By such a standard parallelizing method, we can solve the MCM problem in $O(n^2 \log n)$ steps with n threads.

It has been studied well to speed up DP programs using GPU (e.g. [2], [3]), where they mainly focus on optimizing the order of accessing data by proposing novel techniques avoiding memory access conflicts. In this study, we consider adopting a pipeline technique and implementing the DP program on GPU in a pipeline fashion. The pipeline computation technique [4] can be used in situations in which we perform several operations $\{OP_1, OP_2, \ldots, OP_n\}$ in a sequence, where some steps of each OP_{i+1} can be carried out before operation OP_i is finished. In parallel algorithms, it is often possible to overlap those steps and improve total execution time.

In our approach, we solve the MCM problem on GPU in a pipeline fashion, i.e., we use GPU cores for supporting pipeline-stages so that many elements of the solution table are partially computed in parallel at one time. Our implementation determines one output value per one computational step with n threads in a pipeline fashion and constructs the solution table totally in $O(n^2)$ steps with n threads. This paper is an extended version of our conference paper [5].

The rest of this paper is organized as follows. Section II introduces problem definitions and base algorithms. Section III explains our pipeline implementations for DP on GPU and offers some experimental results. Section IV explains how to apply the pipeline implementation technique to the MCM problem, and finally Section V offers concluding remarks.

II. Preliminaries

In this section, we introduce some preliminary definitions and base algorithms. We first define a simplified DP problem to be solved on GPU, and then explain our GPU implementations of programs.

A. Simplified DP Problem

In this study, we implement a typical DP program on GPU. To simplify the exposition, we focus on programs that solve such a simplified DP problem defined as follows:

Definition 1: (Simplified DP Problem) A one-dimensional array $\texttt{ST}[0, \ldots, n-1]$ of size n as a solution table, a set $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ of k integers representing offset numbers, and a semi-group binary operator $\otimes$ over integers are given. Every element of set $\mathcal{A}$ satisfies the following inequality:

$$a_1 > a_2 > \cdots > a_k > 0.$$

Then, a *simplified DP problem* (S-DP problem) is to fill all the elements of array ST in such a way that each $\texttt{ST}[i]$ is computed by the following equation:

$$\texttt{ST}[i] = \otimes_{1 \le j \le k} \texttt{ST}[i - a_j] \quad (1)$$

where

$\texttt{ST}[0]$, $\texttt{ST}[1]$,. . . ,$\texttt{ST}[a_1 - 1]$ are preset with initial values. ■

For example, Fibonacci number problem is equal to the S-DP problem where $k = 2, a_1 = 2, a_2 = 1$, $\otimes = +$, and ST[0]=ST[1]=1.

B. Conventional Approach to S-DP Problem

To begin with, we show a straightforward sequential algorithm that solves the S-DP problem. Fig. 1 shows the algorithm.

A Sequential Algorithm for S-DP Problem

```
for i = $a_1$ to $n-1$ do
  ST[i] = ST[i−$a_1$];
  for j = 2 to $k$ do
    ST[i] = ST[i] $\otimes$ ST[i−$a_{\mathtt{j}}$];
```

Fig. 1. A sequential algorithm for S-DP problem

The outer loop computes values from ST$[a_1]$ to ST$[n-1]$ in order, and the inner loop computes ST[i] for each i by equation (1). Since the outer loop takes $n - a_1 + 1 = O(n)$ steps and the inner loop requires $O(k-1)$ steps, this algorithm takes $O(nk)$ steps in total.

Next, we consider parallelizing the algorithm for S-DP problem. The straightforward approach is to parallelize the inner loop by using GPU cores. We can easily write a multi-thread program that executes the inner loop-body, ST[i] = ST[i] $\otimes$ ST[i$-a_{\mathtt{j}}$], for each j in parallel using $k-1$ threads at one time. Such an implementation, however, does not improve the time cost, because every thread has access to the same ST[i] and thus memory access conflicts occur. As a result, those memory conflicts should be automatically solved at run-time by the serializing mechanism of GPU, and consequently the whole time cost stays in $O(nk)$ steps, which is the same time cost as that of the sequential implementation.

To avoid the memory access conflicts, we can use a well-known standard parallel prefix computation algorithm [6], [7], in which the computations of $\otimes$ over k values are executed in a tournament fashion. Since the parallel prefix computation runs in $O(\log k)$ steps for k values, obviously the entire time cost can be improved to $O(n \log k)$ steps by using k threads.

Although we can successfully reduce the time cost from $O(nk)$ to $O(n \log k)$ by using the parallel prefix computation, it is not work-time optimal because there are many idle threads during the computations in a tournament fashion. In the next section we propose other parallel implementation strategy and show that we can improve the time cost further.

III. Pipeline Implementation on GPU

In this section, we explain our proposed parallel implementations for S-DP problem on GPU. Our program runs in a pipeline fashion.

A. Pipeline Implementation for S-DP Problem

In our implementation, we use a group of k threads to establish k-stage pipeline, and this thread group treats k consecutive elements at one time in parallel. Fig. 2 describes our pipeline algorithm for the S-DP problem. The index variable i of the outer loop stands for the head position of the k-thread group. The inner loop controls each thread's behaviour in such a way that the j-th thread executes computation for ST[i−j+1] using the value stored in ST[i−j+1−$a_{\mathtt{j}}$].

A Pipeline Algorithm for S-DP Problem

```
for i = $a_1$ to $n+k-2$ do
  for j = 1 to $k$ do in parallel
    Thread j executes the following operation if
    $a_1 \leq i_{\mathtt{j}} < n$ where $i_{\mathtt{j}} = \mathtt{i} - \mathtt{j} + 1$:
      $\mathtt{ST}[i_{\mathtt{j}}] = \begin{cases} \mathtt{ST}[i_{\mathtt{j}} - a_{\mathtt{j}}]; & (\mathtt{j} = 1) \\ \mathtt{ST}[i_{\mathtt{j}}] \otimes \mathtt{ST}[i_{\mathtt{j}} - a_{\mathtt{j}}]; & (\mathtt{j} > 1) \end{cases}$
```

Fig. 2. A pipeline algorithm for S-DP problem

An execution example is shown in Fig. 3, where $k = 3$, $a_1 = 5$, $a_2 = 3$, and $a_3 = 1$ hold and the initial values are already stored in ST[0], ST[1],. . . , ST[4]. In Step 1, the head position i of the thread group is 5. In this step the only one thread is activated and executes ST[5] ← ST[0]. In Step 2, the head position is incremented to 6, and two threads are activated. The first thread treats ST[6] and the second thread works on ST[5]. In Step 3, the head position becomes 7, and now all $k = 3$ threads actively execute operations for ST[7], ST[6], and ST[5] respectively. It should be noted that finally in Step 3 the content of ST[5] is completely determined while those of ST[7] and ST[6] are partially computed and not yet determined. From Step 3, all the $k = 3$ threads are active until Step $n - a_1$ when the head position i of thread group reaches $n-1$, and after that step the number of active threads decreases one by each step. As you can see there is no memory access conflict in this example.

As for the time-complexity of our pipeline implementation, from a theoretical viewpoint, it takes only $O(n)$ steps, because the outer loop takes $n + k - a_1 - 1 = O(n)$ cycles and the inner loop requires $O(1)$ time if there is no adjacent offset pair (a_m, a_{m+1}) such that $a_m = a_{m+1} + 1$.

However, from a practical viewpoint, because of the memory access conflicts, the inner loop may take more time steps. Actually, in the worst case when consecutive offset numbers are given, those ST$[i_{\mathtt{j}} - a_{\mathtt{j}}]$, in the right hand side of the assignment statement, coincidentally become the same element of array ST and hence the worst memory access conflicts occur. In such a case, all threads in the inner loop are serialized and it takes time proportional to k. See Fig. 4 for such a worst case example. In this example, all four threads try to have access to ST[$\mathtt{i} - 4$] at the same time in the inner loop.

Let $seq = (a_p, a_{p+1}, \ldots, a_q)$ be one of the longest subsequences of given offset numbers $(a_1, a_2, \ldots, a_k)$ satisfying $a_r = a_{r+1} + 1$ for all $p \leq r < q$. Then, it is easy to check that in the inner loop every thread r $(p \leq r < q)$ has access to the

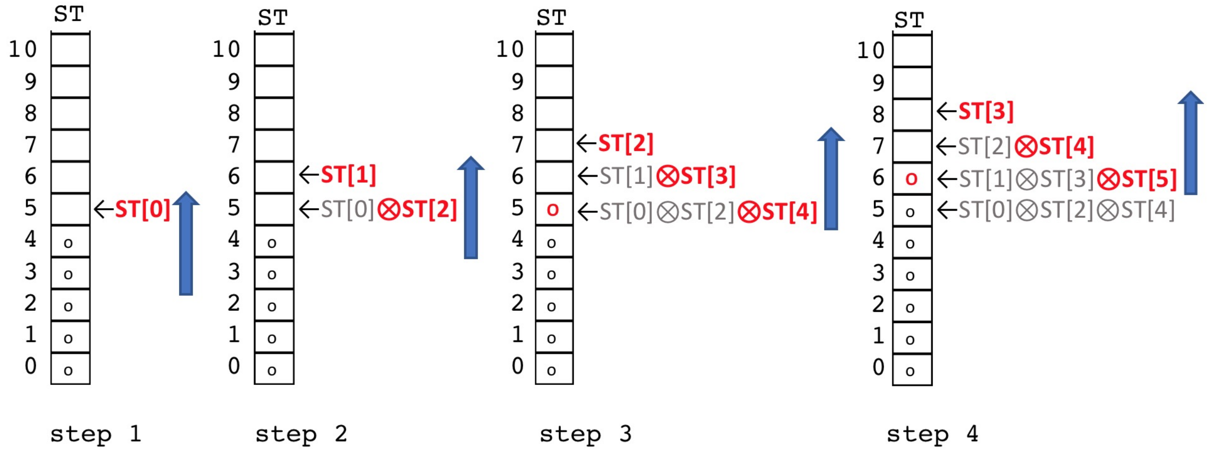


Fig. 3. An execution example for the case where $k = 3$, $a_1 = 5$, $a_2 = 3$, and $a_3 = 1$ hold and initial values are preset to ST[0], ST[1], . . ., and ST[4].

same element of array ST, and as a result those conflicts are serialized at run time by GPU's serializing mechanism and hence the memory access time becomes $(q - p + 1)$ times slower than that of conflict-free case. For such a case, in [5], we proposed a *2-by-2 pipeline implementation* technique where each thread invoked in the inner loop executes two computations for each element of array ST. The details can be found in [5].

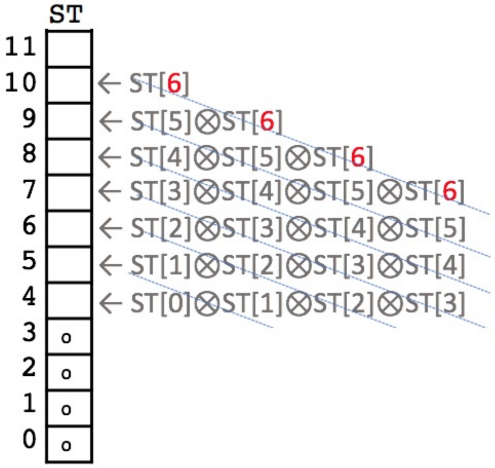

ST
11
10
9
8
7
6
5
4
3
2
1
0
o
← ST[6]
← ST[5]⊗ST[6]
← ST[4]⊗ST[5]⊗ST[6]
← ST[3]⊗ST[4]⊗ST[5]⊗ST[6]
← ST[2]⊗ST[3]⊗ST[4]⊗ST[5]
← ST[1]⊗ST[2]⊗ST[3]⊗ST[4]
← ST[0]⊗ST[1]⊗ST[2]⊗ST[3]

Fig. 4. An example for the worst case where the offset numbers are consecutively given. In this example, we have $k = 4$, $a_1 = 4$, $a_2 = 3$, $a_3 = 2$, and $a_4 = 1$.

B. Experimental Results

Before going to the next section, we show the performance of our pipeline implementation on GPU. We use a computer with 3.40 [GHz] Intel Xeon CPU E3-1245 v3 and an NVIDIA GeForce GTX TITAN Black. The OS is Ubuntu 17.10 and we use g++ 7.2.0 for compiling cpp programs and CUDA 9.2 [8]. The experimental results are shown in TABLE I.

TABLE I shows the average execution time for each implementation on GPU. In the table, SEQUENTIAL, NAIVE-PARALLEL, and PIPELINE respectively stand for the sequential implementation, the naive multi-thread implementation, and our pipeline implementation (for a general case). Here, we use min operation for ⊗. The average is computed among 100 executions for each setting.

TABLE I. EXECUTION TIME OF SEQUENTIAL, NAIVE PARALLEL, AND PIPELINE IMPLEMENTATIONS (MSEC)

	SEQUENTIAL	NAIVE-PARALLEL	PIPELINE
$2^{14} \leq n \leq 2^{15}$, $2^{12} \leq k \leq 2^{13}$	274	64	78
$2^{16} \leq n \leq 2^{17}$, $2^{14} \leq k \leq 2^{15}$	4,288	368	386
$2^{18} \leq n \leq 2^{19}$, $2^{16} \leq k \leq 2^{17}$	68,453	3,018	2,408

As for the comparison between the sequential implementation and the parallel ones, parallel implementations are much faster even though it is NAIVE-PARALLEL. Although there is no difference in time between NAIVE-PARALLEL and PIPELINE until $n \leq 2^{17}$, PIPELINE is faster than NAIVE-PARALLEL when $n \geq 2^{18}$.

IV. SOLVING MCM PROBLEM

In this section, we explain how to apply our pipeline implementation technique to more general DP problems. As an example, we deal with the matrix chain multiplication problem (MCM problem) [1].

A. Outline of Pipeline Implementation for MCM Problem

It is well-known that the MCM problem can be efficiently solved by DP with a two-dimensional solution table of triangular shape, and that each element is computed along the diagonal direction. See Fig. 5 for an example. In the figure, each number represents the order of elements to be computed.

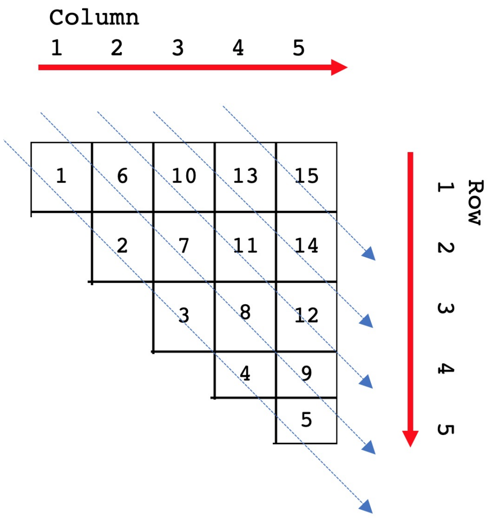


Fig. 5. A solution table example for a matrix chain multiplication problem.

The content of each element is computed by a binary operator $\downarrow$ returning a smaller operand and a binary function $f(*,*)$ as in Fig. 6. The detailed definition of the MCM problem can be found in [1]. In the figure, the element marked 13 is computed by the elements marked by 1, 6, 10, 11, 8, and 4. If we write $\texttt{ST}[x]$ for the element marked x here, the computation is expressed as

$$\begin{aligned}&\texttt{ST}[13]\\&= f(\texttt{ST}[1], \texttt{ST}[11]) \downarrow f(\texttt{ST}[6], \texttt{ST}[8]) \downarrow f(\texttt{ST}[10], \texttt{ST}[4]).\end{aligned}$$

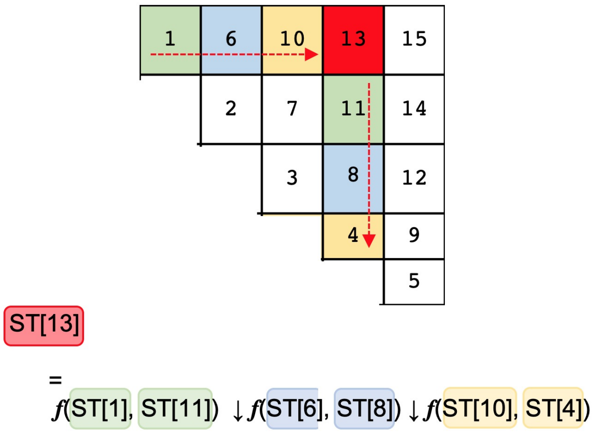


Fig. 6. An example of computing an element of solution table. Here, we write $\texttt{ST}[x]$ for the element marked x.

Since the elements of two-dimensional solution table are computed in a total order (linear order), we can line up them into a linear array according to that order. Once the solution table is transformed into a linear array, we can apply the pipeline technique to the MCM problem as well. An execution example is shown in Fig. 7.

B. Detailed Pipeline Implementation for MCM Problem

Here, we explain how to solve the MCM problem by pipeline implementation in details.

Firstly, we assume that the original two-dimensional solution table of diagonal shape is already mapped to a one dimensional solution table $\texttt{ST}$ appropriately. That is, we line up all the elements of the two-dimensional table into a linear array $\texttt{ST}$ according to the total (linear) order in which each element is computed along the diagonal direction as in Fig. 5 and 7. Since the number of input matrices is n, the number of elements in the solution table is $n(n+1)/2$. First n elements of array $\texttt{ST}$ are preset with initial values, those correspond to the elements located in the diagonal line of the original two-dimensional table.

Let each computation for $\texttt{ST}[i]$ be represented as

$$\texttt{ST}[i] = \downarrow_{1\leq j\leq k_i} \; f(l_{(i,j)}, r_{(i,j)}). \tag{2}$$

It should be noted that here each k_i may differ from others and that it is $n-1$ at largest. See Fig. 5 and 7 for an example. Here, we have $n = 5$, and $\texttt{ST}[13]$ and $\texttt{ST}[12]$ can be represented as follows:

$$\begin{aligned}&\texttt{ST}[13]\\&= f(l_{(13,1)}, r_{(13,1)}) \downarrow f(l_{(13,2)}, r_{(13,2)}) \downarrow f(l_{(13,3)}, r_{(13,3)})\\&= f(\texttt{ST}[1], \texttt{ST}[11]) \downarrow f(\texttt{ST}[6], \texttt{ST}[8]) \downarrow f(\texttt{ST}[10], \texttt{ST}[4])\end{aligned}$$

and

$$\begin{aligned}&\texttt{ST}[12]\\&= f(l_{(12,1)}, r_{(12,1)}) \downarrow f(l_{(12,2)}, r_{(12,2)})\\&= f(\texttt{ST}[3], \texttt{ST}[9]) \downarrow f(\texttt{ST}[8], \texttt{ST}[5]).\end{aligned}$$

To solve such defined MCM problem by the pipeline implementation on GPU, we have to modify the pipeline algorithm designed for the S-DP problem, because the offset numbers for each $\texttt{ST}[i]$ may differ from others. Thus, for the MCM problem, we propose a modified pipeline algorithm in Fig. 8. In the MCM-pipeline algorithm, each element of $\texttt{ST}$ is computed by equation (2), and $\texttt{ST}[1]$, $\texttt{ST}[2]$,… ,$\texttt{ST}[n]$ are preset with initial values. In substep 1, 2, and 3, the computation of $f(l_{(i,j)}, r_{(i,j)})$ is executed, and in substep 4, that obtained value is used for the computation by $\downarrow$ and the result is stored to $\texttt{ST}[i]$.

C. Conflict-Free Memory Access of MCM Algorithm

In this subsection, we prove that no memory access conflict occurs during the execution of the MCM-pipeline algorithm described in Fig. 8.

To begin with, we prove the following lemma.

Lemma 1: In substep 1 of an execution of the inner loop-body of the MCM algorithm, each thread has access to a distinct element of the array $\texttt{ST}$.

Proof: Assume that threads p and q try to read the same element of $\texttt{ST}$ in substep 1 and that $p < q$ holds. Let (row_p, col_p) (resp. (row_q, col_q)) be the pair of row and column indexes of the elements in the original two-dimensional solution table of

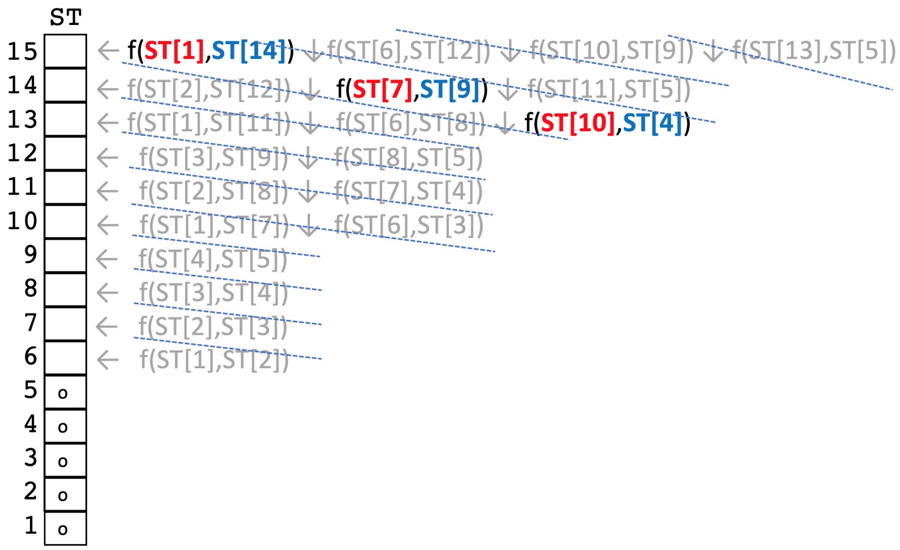

ST
15 ← f(ST[1],ST[14]) ↓ f(ST[6],ST[12]) ↓ f(ST[10],ST[9]) ↓ f(ST[13],ST[5])
14 ← f(ST[2],ST[12]) ↓ f(ST[7],ST[9]) ↓ f(ST[11],ST[5])
13 ← f(ST[1],ST[11]) ↓ f(ST[6],ST[8]) ↓ f(ST[10],ST[4])
12 ← f(ST[3],ST[9]) ↓ f(ST[8],ST[5])
11 ← f(ST[2],ST[8]) ↓ f(ST[7],ST[4])
10 ← f(ST[1],ST[7]) ↓ f(ST[6],ST[3])
9 ← f(ST[4],ST[5])
8 ← f(ST[3],ST[4])
7 ← f(ST[2],ST[3])
6 ← f(ST[1],ST[2])
5 o
4 o
3 o
2 o
1 o

Fig. 7. An execution example of pipeline implementation for the matrix chain multiplication problem.

A Pipeline Algorithm for MCM Problem

```
for i = $n+1$ to $n(n+1)/2+n-2$ do
  for j = 1 to $(n-1)$ do in parallel
    Thread j executes the following operation if
    $i_j \leq k_{i_j}$ where $i_j = i - j + 1$:
      (substep 1)
        $v_l = l_{(i_j,j)}$;
      (substep 2)
        $v_r = r_{(i_j,j)}$;
      (substep 3)
        $v_s = f(v_l, v_r)$;
      (substep 4)
        $ST[i_j] = \begin{cases} v_s; & (j = 1) \\ ST[i_j] \downarrow v_s; & (j > 1) \end{cases}$
    where $v_l, v_r$, and $v_s$ are local variables in a
    thread.
```

Fig. 8. A pipeline algorithm for MCM problem

triangle shape of the MCM problem for which thread p (resp. q) is now computing. Since threads p and q try to read the same element of ST and in substep 1 they read the value for the left argument of the function f, the relation $row_p = row_q$ must hold. Then, since threads p and q respectively read the p-th and q-th elements from the left in the same row of the original two-dimensional solution table, the relation $p = q$ must hold if the two threads read the same element of ST, which contradicts the assumption $p < q$. ■

Next, we prove the following lemma, which can be proved in a similar way to the proof of Lemma 1.

Lemma 2: In substep 2 of an execution of the inner loop-body of the MCM algorithm, each thread has access to a distinct element of the array ST.

Proof: Assume that threads p and q try to read the same element of ST in substep 2 and that $p < q$ holds. Let (row_p, col_p) (resp. (row_q, col_q)) be the pair of row and column indexes of the elements in the original two-dimensional solution table of triangle shape of the MCM problem for which thread p (resp. q) is now computing. Since threads p and q try to read the same element of ST and in substep 2 they read the value for the right argument of the function f, the relation $col_p = col_q$ must hold. On the other hand, thread p reads the p-th elements below of the row row_p of the original two-dimensional solution table, and thread q does the q-th elements below of the row row_q of the table. Since threads p and q read row $(row_p + p)$ and row $(row_q + q)$ respectively, the relation $row_p + p = row_q + q$ must hold if the two threads read the same element of ST. Since $p < q$ and $col_p = col_q$ hold, the relation $row_p < row_q$ must hold from the way of mapping from the original two-dimensional solution table to the linear array ST. This leads to the relation $row_p + p \neq row_q + q$, which contradicts the assumption that threads p and q read the same element of array ST. ■

In substep 3, each thread simply executes computation using local variables. In substep 4, it is obvious that each thread has access to a distinct element of ST. Therefore, by Lemma 1 and Lemma 2, we obtain the following theorem.

Theorem 1: No memory access conflict occurs during the execution of the MCM-pipeline algorithm described in Fig. 8. ■

As for the time-complexity of the MCM-pipeline implementation, from a theoretical viewpoint, it takes only $O(n^2)$ steps with $(n-1)$ threads, because the outer loop takes $O(n^2)$ cycles and the inner loop requires $O(1)$ time (Theorem 1).

V. Concluding Remarks

In this study, we examined the effectiveness of pipeline implementations of Dynamic Programming (DP) on GPU. As an example, we explained how to solve a matrix-chain multiplication (MCM) problem by DP on GPU. This problem can be sequentially solved in $O(n^3)$ steps by DP where n is the number of matrices. In our approach, we solve the MCM problem on GPU in a pipeline fashion, i.e., we use GPU cores for supporting pipeline-stages so that many elements of the solution table are partially computed in parallel at one time. Since our implementation determines one output value per one computational step with $O(n)$ threads, we can solve the MCM problem in $O(n^2)$ steps using $O(n)$ threads, which is an ideal speedup from the $O(n^3)$-step sequential DP algorithm.

For future work, we plan to evaluate the performance of our pipeline implementations. From the experimental results shown in Section III, it is obvious that the ideal speed up is not attained here. This is mainly due to the limitations on the bandwidth of memory on GPU. That is, as the problem size becomes large, all threads cannot always have access to the target memory at one time, because unavoidable access conflicts occur. We also plan to study the relation between the memory bandwidth and the performance of our pipeline implementation on some theoretical GPU models (e.g., [9]).

Acknowledgment

This paper is an extended version of our conference paper [5] where we proposed a pipeline-implementation for the S-DP problem and discussed the memory access conflict issues. In this paper, we provide the MCM algorithm in details and formally prove lemmas and theorem in Section IV.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[2] Y. Ito and K. Nakano. A gpu implementation of dynamic programming for the optimal polygon triangulation. *IEICE Transactions on Information and Systems*, E96.D(12):2596–2603, 2013.

[3] K. Nakano. A time optimal parallel algorithm for the dynamic programming on the hierarchical memory machine. In *2014 Second International Symposium on Computing and Networking (CANDAR)*, pages 86–95, 2014.

[4] S. H. Roosta. *Parallel Processing and Parallel Algorithms: Theory and Computation*. Springer, 1999.

[5] M. Miyazaki and S. Matsumae. A pipeline implementation for dynamic programming on gpu. In *International Workshop on Parallel and Distributed Algorithms and Applications (in conjunction with CANDAR'18)*, Nov. 2018.

[6] F. T. Leighton. *Introduction to Parallel Algorithms and Architectures: Array, Trees, Hypercubes*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992.

[7] V. Kumar, A. Grama, A. Gupta, and G. Karypis. *Introduction to Parallel Computing: Design and Analysis of Algorithms*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1994.

[8] Cuda toolkit 9.2 download. `https://developer.nvidia.com/cuda-downloads`.

[9] K. Nakano and S. Matsumae. The super warp architecture with random address shift. In *2013 20th International Conference on High Performance Computing (HiPC)(HIPC)*, volume 00, pages 256–265, Dec. 2013.

FPGA based Hardware-in-the-Loop Simulation for Digital Control of Power Converters using VHDL-AMS

Abdelouahab Djoubair Benhamadouche[1]
Department of Electronic, Faculty of Technology,
University Ferhat Abbas Sétif 1,
Sétif 19000, Algeria

Adel Ballouti[2]
Department of Electronic, Faculty of Technology,
University Mohamed Boudiaf,
M'Sila 28000, Algeria

Farid Djahli[3]
LIS Laboratory, Department of Electronic,
University Ferhat Abbas Sétif 1,
Sétif 19000, Algeria

Abdeslem Sahli[4]
LEPCI Laboratory, Electronics Department,
University Ferhat Abbas Sétif 1,
Sétif 19000, Algeria

***Abstract*—This paper presents a new approach for complex system design, allowing rapid, efficient and low-cost prototyping. Using this approach can simplify designing tasks and go faster from system modeling to effective hardware implementation. Designing multi-domain systems require different engineering competences and several tools, our approach gives a unique design environment, based on the use of VHDL-AMS modeling language and FPGA device within a single design tool. This approach is intended to enhance hardware-in-the-loop (HIL) practices with a more realistic simulation which improve the verification process in the system design flow. This paper describes the implementation of a software/hardware platform as effective support for our methodology. The feasibility and the benefits of the presented approach are demonstrated through a practical case study of a power converter control. The obtained results show that the developed method achieves significant speed-up compared with conventional simulation methods, using minimum resources and minimum latency.**

***Keywords*—Hardware-in-the-Loop (HIL) simulation; Field-Programmable Gate Array (FPGA); VHDL-AMS; power converter; digital controller**

I. Introduction

In a top-down development process, design tasks are critical keys in system implementation success. In technologic research or industry fields, such as electronic, automation or mechatronic, designing system details go through different steps of analysis and modeling [1], where modeling consists in developing abstract descriptions of some physical realities in such a way that they are useful for the design process [2]. Conceived models are then used as an input to a specific simulator to study systems behavior or to explore unconsidered functionalities.

Modeling at different levels of abstraction of multi-domain systems needs the use of different methodologies and tools that should manage mixed-signal design and hardware/software challenges. For that reason, mixed-signal hardware description languages such as VHDL-AMS are indispensable, it is intended to provide a unifying trend that will link the various tasks of analog and mixed-signal design in a coherent framework to support different design methodologies and different design tools [2], [3].

On another side, Verification and validation (V&V) procedures are the solutions for a system design success, they go along with each step of the development cycle. Different techniques can be used to perform those procedures. Usually, it depends on system architecture and hardware implementation. In different engineering systems such automotive, mechatronic or control systems, best practice for V&V requests the construction of a real prototype to test the whole system or one element of the system such as a new control algorithm [4]. This prototype is subjected to several cycles of testing and re-design in an expensive facility [5].

Therefore, hardware-in-the-loop (HIL) methods emerge in system design to be an effective way to resolve that issue. HIL methods seem to be an effective alternative to accelerate verification through a specific and relatively low-cost equipment, that permit to finally build a real and an operational prototype [6], [7].

In this paper, an original method for HIL simulation is presented, this method is proposed to simplify and accelerate systems design. This method is based on analog and mixed-signal (AMS) languages and programmable devices. The idea is to use the same hardware description language to model the digital part and the analog part of an AMS system, the digital component (essentially control procedure) is then implemented in FPGA, and the analog part is simulated through the use of VHDL-AMS and dedicated simulation software.

This paper is organized as follows. Section 2 introduces the essential of our approach that includes a short description of the context of each contribution of this work. Section 3 presents the implementation aspect of FPGA prototyping. Then, section 4 describes an application case to test and to show the benefits of the effective implementation of our method, where DC/DC converter is modeled in VHDL-AMS and controlled via a digital Proportional-Integral-Derivative

(PID) regulator implemented in FPGA. The last section states the conclusion and summarizes the most salient features of our contribution.

II. Design Approach

Modeling and simulating a multi-domain system need the use of dedicated tools that can perform multi-abstraction simulation, functional prototyping, verifications, and validations. In the present work, those different design phases would be done through the use of FPGA device and VHDL-AMS modeling language.

In this work, VHDL-AMS is used as the mainstay for the system design; it will be used as a support language for modeling and simulation in the different phases of the system design flow. VHDL-AMS is an extension of the IEEE standard VHDL language, this language allows designers to model any mixed-signal plant that can be described by a system of differential and algebraic equations (DAE's), it supports the hierarchical description and the simulation of continuous and mixed continuous/discrete systems with conservative and non-conservative semantics [2], [8].

VHDL-AMS design requests a simulation process that involves a combined simulation between the analog part and the digital part of an AMS plants model. This is done through data communication and synchronization between a discrete-time and continuous-time simulator engines. In most applications which use VHDL-AMS, the digital side which is VHDL is usually excluded from the design or used only for behavioral simulation, and the analog facet of VHDL-AMS is the only one highlighted part [9]–[12]. In the present work, we will take advantages of VHDL to enhance VHDL-AMS design and vice versa, this is done through the validation of the following two points:

- Using VHDL-AMS for mixed systems modeling for pure simulation and verification.
- Using VHDL-AMS to develop FPGA application, in order to make a functional simulation, hardware implementation, and verification.

The second part of our design approach is using FPGA based hardware-in-the-loop techniques for test and verification. Usually, FPGAs are used when we need fast processing and parallel computing. Currently, they are also used for system rapid prototyping and system verification, where they take an important part on hardware-in-the-loop platforms. Several works have been recently proposed for FPGA based hardware-in-the-loop [13]–[18], where FPGA benefits are highlighted to simplify and accelerate considerably hardware-in-the-loop simulation. This technique is increasingly used in the development of system control in several fields of application.

Therefore, in this paper, we introduce a new method to use hardware-in-the-loop techniques, where FPGA device is associated with a simulation software to make an online simulation with VHDL-AMS model as resumed in Figure 1.

In order to reach our goal, we have to find the best way to bring together FPGA prototyping and VHDL-AMS modeling in an advanced software/hardware simulation platform.

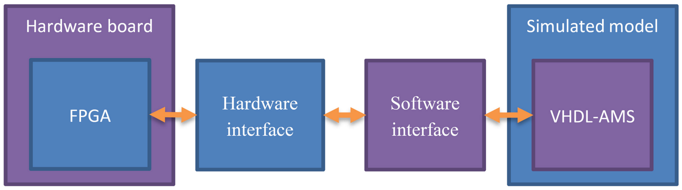


Fig. 1. AMS design for FPGA based HIL verification

III. Design Implementation

The aim of the proposed platform is to set up a data communication between the FPGA device and VHDL-AMS simulator. Figure 2 presents a simplified representation of what the HIL platform must enclose, three parts are essential; VHDL-AMS simulator, an FPGA development board and a software application to manage the simulation.

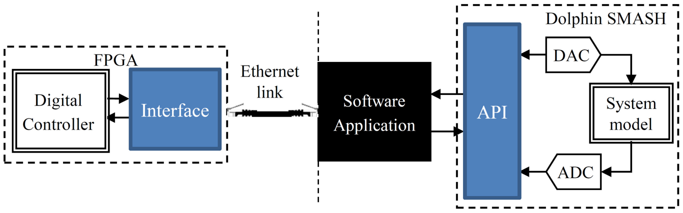


Fig. 2. Simplified representation for the proposed platform

A. VHDL-AMS Simulator

As mentioned in the last section, carrying out our approach need the use of a software which can perform VHDL-AMS simulation and have the ability to communicate with other applications. SMASH software from Dolphin integration is chosen to execute our approach, this software can support the full IEEE VHDL-AMS standard and a set of other modeling languages [19], [20]. Moreover, SMASH integrates an application programming interface (API) which gives functions and services to customize and to control simulation running.

B. The Software Application

To perform a hardware-in-the-loop simulation using VHDL-AMS, we have to control the simulation process through the SMASH API. According to Figure 3, a C++ application is built to allow data exchange between simulated models and the FPGA development board, this application performs the following tasks:

- The first task that should be accomplished is controlling the simulation process, so the application will be able to collect data from the simulated models and have to force signals values during simulation runs.
- The second task is to transfer data from and to the FPGA Board. Data collected from the simulator are assembled in one datagram then transferred in an UDP/IP packet to the FPGA through the Ethernet link. Afterward, the application is placed in listening mode and awaiting a response from the FPGA. When received, the datagram is unwrapped to extract the

needed data to force new signals values, and to allow SMASH to continue the simulation running.

- The last performed task is to synchronize and to manage data exchange between the simulator and the FPGA board.

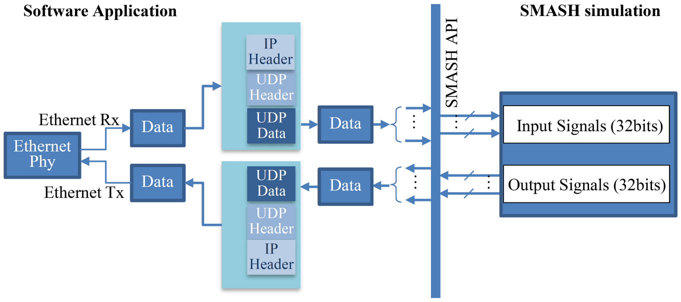


Fig. 3. Software application global view

C. FPGA Implementation

In the present approach, the HIL simulation does not require a sophisticated FPGA board, it simply needs a low-performance device with an Ethernet interface. To achieve the needed functionalities, a minimum configuration has to be implemented in the FPGA device as depicted in Figure 4.

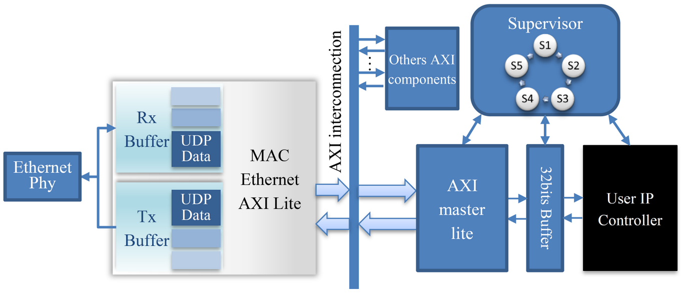


Fig. 4. Implementation of the hardware controller in the FPGA

To build a reconfigurable and an evolvable FPGA implementation, the AXI interconnection is used as the basement of our design, with a large number of compatible AXI IP core, we can extend our design to execute other useful functions. Moreover, it can integrate a processor core which is able to execute different control algorithms.

As described in Figure 4, the minimum configuration includes four essential blocks, that are:

1) Ethernet MAC (Media Access Controller): Many Ethernet controller cores can be used, it depends on several parameters like the targeted FPGA, connection speed or channel protocol. The proposed platform uses AXI Ethernet lite MAC from Xilinx [21], it provides an efficient Ethernet interface with least resources and can be configured to operate at the rate of 100Mb/s.
2) AXI Master: This module is written in VHDL, it follows the AXI protocol and provides signals to perform read/write operations in the AXI bus, this is done to control the Ethernet module. On the other side, this module exchange data with the user controller module.
3) User IP Controller: This is the implementation of the digital controller under test.
4) Supervisor: This module manages all the operations executed by the overmentioned modules. The supervisor works under the control of a finite state machine (FSM). The FSM provides control signals to manage and synchronize data transaction between the different modules.

IV. Results and Performances

A. Case Study

To illustrate our concept and bring out its benefits on system design, a typical example of an electronic power system is used. As shown in Figure 5, the used system consists of a buck power converter with a closed loop controller; this power converter is a typical example of AMS systems, it involves analog and digital behavior and can include other comportments from other disciplines like a thermal behavior or mechanical drive.

As depicted in Figure 5, a voltage sensor gets the digital value of the output voltage, this value is then compared to the reference voltage Vref which generates an error signal e(k) ready at the input of the Proportional-Integral-Derivative (PID) controller. The PID controller makes the feedback loop to regulate the output voltage across the load resistor (R). In this case, the PID runs continuously to generate a duty-cycle u(k) that control the PWM generator to drive the transistor switch (SW).

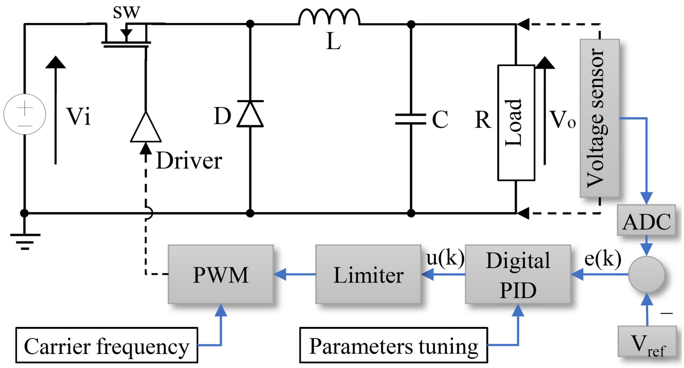


Fig. 5. Buck power converter structure with PID control loop

The simplest digital form of the PID controller in the discrete-time domain algorithm is given by the equation below:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt} \tag{1}$$

Where the controller input $e(t)$ is the difference between the reference and the load voltage, and $u(t)$ is the controller output. The three parameters K_p, K_i, and K_d represent proportional, integrator and derivative gain coefficients.

TABLE I. System Parameters

Parameters	Values
PWM frequency	10.0 kHz
Input voltage	24.0 V
Filter inductor	1.0 mH
Filter capacitor	1.0 mF
Load resistor	1.0 Ω
Step simulation	Variable 2.0us - 50.0us

For a digital implementation, we are more interested in the discrete-time domain formula, this is given by the equation (2).

$$u[k] = u[k-1] + K_1 e[k] + K_2 e[k-1] + K_3 e[k-2] \quad (2)$$

Where: $K_1 = T_s K_i + \frac{K_d}{T_s}$, $K_2 = -K_p - 2\frac{K_d}{Ts}$, $K_3 = \frac{K_d}{Ts}$, and T_s is the sampling period.

In this work, a structural description is used to make a simulation model, each part of the system is designed in different levels of abstraction, depending on the needed accuracy and the complexity in each model. The analog part of the system is modeled in VHDL-AMS and the controller is initially described in VHDL-AMS, then a VHDL code is built to be finally implemented in FPGA.

To build a valid VHDL code intended for an FPGA implementation, first, we have to rearrange the code to make a synthesizable VHDL description according to the available FPGA resources [22]. when verified, the PID blocks are assembled with the others needed blocks for the HIL process, then timing constraints are checked for the final implementation. At this stage, an important issue is to be considered, a binary format representation of digital data has to be set, we choose a 32-bit fixed point format for input/output signals of the digital controller.

B. VHDL-AMS Simulation

First of all, the whole system is modeled in VHDL-AMS including system parameters as shown in Table I, the system is then simulated in a mixed signal mode with a variable time step, where each part of the system is verified with different scenarios especially the digital controller model.

The simulation waveforms depicted in the Figure 6.a shows the load voltage value for variable voltage reference; the result allows us to verify that the digital controller draws the output voltage of the analog buck converter according to the input reference. Also, in Figure 6.b we can see the different signals involved in the functioning of the digital PID such as the clock, Error, voltage signal...etc.

C. FPGA Simulation and Verification

In order to validate our approach, we perform numerous tests using the developed platform based on FPGA and VHDL-AMS simulator. The hardware part can be implemented in any FPGA development board that includes an Ethernet PHY. To be consistent with our objectives, we consider a low-cost development board; an Avnet LX9 MicroBoard embedding a Xilinx Spartan-6 FPGA, which is suitable for low-budget design (Figure 7).

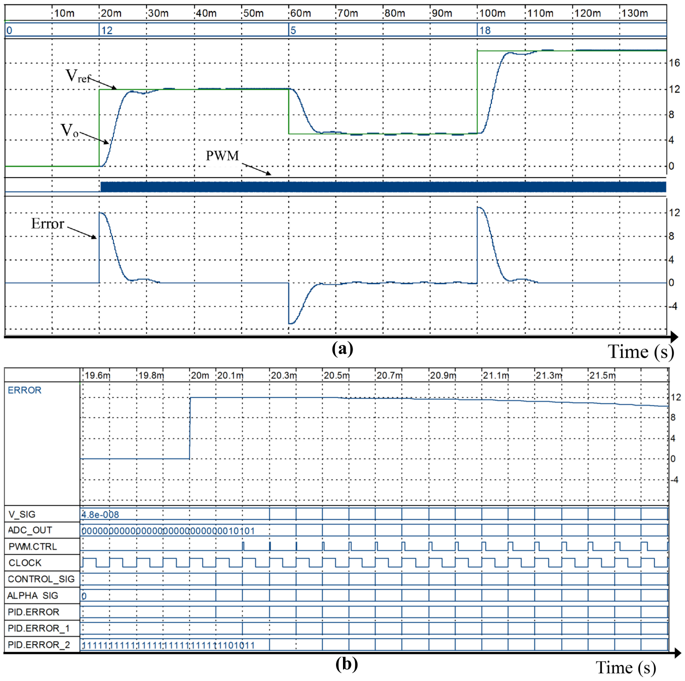


Fig. 6. VHDL-AMS simulation waveforms, (a) Voltage load and voltage error for variable voltage reference, (b) PID digital signals.

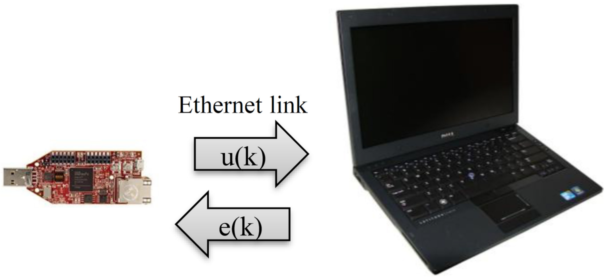


Fig. 7. FPGA based Hardware-in-the-loop plan.

After programming the FPGA, the developed software application is executed, then SMASH software loads the plant model. After initializing the software parameters and the UDP/IP Ethernet communication channel, the user can begin the HIL simulation using a VHDL-AMS testbench, this last generates the necessary signals for the different test scenarios.

To compare efficiently software simulation and FPGA based HIL simulation, the same simulator parameters as for pure VHDL-AMS simulation (presented in the last section) are applied. Experimental results are obtained with a 100Mb/s Ethernet channel and data are transferred to, and from the FPGA board as vectors of 32 bits, and the FPGA is running at 100MHz.

For the test validation, the same scenarios are considered as used for the previous simulation. The Figure 8.a shows the result of FIL simulation, where the load voltage waveform is identical to the software simulation. In the zoomed view

TABLE II. FPGA USED RESOURCES FOR HARDWARE IMPLEMENTATION

Component	Slice Registers	Slice LUT	DSP blocks
Available	11440	5720	16
AXI_Ethernet_lite	492 [4,3%]	494 [8,6]	0
Supervisor + AXIMaster	197 [1,7%]	220 [3.8]	0
PID_dig	127 [1,1%]	217 [3,8]	12[75%]
Global	816 [7,1%]	931 [16,2]	12[75%]

at 20ms (Figure 8.b), the waveforms show a delayed response compared with the software simulation, this is due to latency in the transmission channel and the timing constraints associated with the FPGA implementation

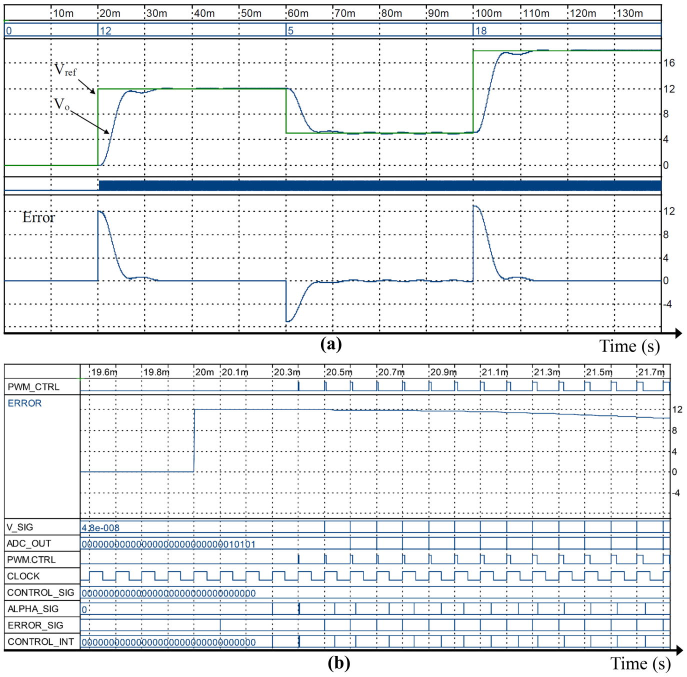


Fig. 8. HIL simulation waveforms (a) Voltage load and voltage error for variable voltage reference (b) PID digital signals.

D. FPGA Resources

Table II lists the FPGA resources used to implement the needed components, consumed resources depend on the used FPGA technology, and proportions are referred to the available resources of a Spartan-6 XC6SLX9.

In this table, the AXI Ethernet lite module is a Xilinx generated core which provides the Ethernet interface, the PID_dig represents the digital controller part, and the third component refers to the supervisor part that is necessary to connect and control all needed components via an AXI bus.

The global utilization of LUT represents only 16,2% of the available LUT, and the integration of the PID controller represent only 3,8%, those results show that the proposed HIL platform don't need large FPGA resources, and with the actual FPGA device there are enough resources to replace the digital PID with a more complex controller or to add other modules for additional functionalities.

TABLE III. TIME SIMULATION FOR DIFFERENT SCENARIOS

	Simulation Mode	CPU global time	Logic kernel time	Improvement of simulation time
Scenario 1 140ms & 10kHz	Software	7,83 s	5,38 s	83,66%
	HIL_AMS	1,28 s	0,109 s	6,55 s
Scenario 2 1000ms & 10kHz	Software	39,14 s	30,735s	75%
	HIL_AMS	9,79 s	0,733 s	29,35 s
Scenario 3 140ms & 10kHz	Software	29,22 s	21,62s	80,10%
	HIL_AMS	5,70 s	0,608 s	23,52 s
Scenario 4 1000ms & 50kHz	Software	224,7 s	182,39 s	82,45%
	HIL_AMS	39,42 s	3,76 s	184,48 s

E. Timing Performance

To bring out the benefits of the implementation of our approach in terms of time processing, several simulation scenarios are performed in software mode and in FPGA mixed mode with a variable simulation time step. The simulation times have been measured using a 2.5 GHz Intel Core i5 with 4 GB of RAM memory.

Table III summarises the CPU time and the logic kernel time required to perform for the different scenarios. Those scenarios simulate the same design for different simulation times and for different switching frequencies, this implies a different volume of analog and discrete-time simulation points which demand various hardware resources.

The needed simulation time is significantly reduced (about 80%). This is due to the fact that software mixed simulation uses big CPU resources for digital simulation, while in our approach; the digital simulation is exclusively achieved by the FPGA device.

On the other hand, we measured the time latency for different simulation cases. We note that the latency for network transmission and processing within the FPGA is about 55μs and the latency for software processing is about 2 5ms. Improving hardware and software latency for a specific application case can lead to real-time hardware-in-the-loop simulation.

V. CONCLUSION

The work exposed in this paper focuses on the presentation of a new approach for simulation and verification of complex systems. With this goal in mind, this work introduces a unique design tool for hardware-in-loop simulation based on FPGA and VHDL-AMS. Thereby, benefits of VHDL-AMS mixed design and FPGA rapid prototyping are combined together to improve system design and to allow fast design development time and a short time to market.

In order to validate our approach and to demonstrate its effectiveness, results of the implementation and simulation of a regulated power converter are presented, where the digital controller is coded using VHDL language then implemented in FPGA and the rest of the system is simulated on a host PC using VHDL-AMS models. The studied application gives an overview of the capabilities of this method to accomplish multiple design benefits, where several experiences have been achieved shown a significant reduction of simulation processing time and a low latency for data transfer. On another hand, hardware resources are minimized, to match any low-cost FPGA development board for a high-reliability simulation.

REFERENCES

[1] D. Gianni, A. D'Ambrogio, and A. Tolk, Modeling and Simulation-Based Systems Engineering Handbook, CRC Press, 2014.

[2] F. Pêcheux, C. Lallement, and A. Vachoux, "VHDL-AMS and Verilog-AMS as alternative hardware description languages for efficient modeling of multidiscipline systems", IEEE Trans. Comput.-Aided Design Integr. Circuits Syst, vol. 24, pp. 204–225, 2005.

[3] G. G. Gielen and R. Rutenbar, "Computer-aided design of analog and mixed-signal integrated circuits", Proc. IEEE , vol. 88, pp. 1825–1854, 2000

[4] P. G. Maropoulos and D. Ceglarek, "Design verification and validation in product lifecycle", CIRP Ann-Manuf. Techn, vol. 59, pp. 740–759, 2010.

[5] G. G. Parma and V. Dinavahi, "Real-time digital hardware simulation of power electronics and drives", IEEE Trans. Power Del, vol. 22, pp. 1235–1246, 2007.

[6] M. Bacic, "On hardware-in-the-loop simulation", Proceeding of the 44th IEEE Conference on Decision and Control, Seville, pp.3194—-3198, 2005.

[7] A. Bouscayrol, "Different types of Hardware-In-the-Loop simulation for electric drives", Proceeding of the International Symposium on Industrial Electronics, Cambridge, pp. 2146–2151, 2008.

[8] E. Christen and K. Bakalar, "VHDL-AMS-a hardware description language for analog and mixed-signal applications", IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process, vol. 46, pp. 1263–1272, 1999.

[9] A. Laouamri, F. Djahli, A. Rabhi, A. Benhamadouche, A. Ballouti, and A. Bendjadou ,"A Virtual Prototype of Proton Exchange Membrane Fuel Cell Using VHDL-AMS Language", J FUEL CELL SCI TECH, vol. 6, pp. 024502, 2009.

[10] Y.-A. Chapuis, L. Zhou, H. Fujita, and Y. Hervé, "Multi-domain simulation using VHDL-AMS for distributed MEMS in functional environment: Case of a 2D air-jet micromanipulator", Sensors and Actuators A: Physical, vol. 48, pp. 224–238, 2008.

[11] L. Zerioul, M. Ariaudo, and E. Bourdel, "RF transceiver and transmission line behavioral modeling in VHDL-AMS for wired RFNoC", ANALOG INTEGR CIRC S, vol. 92, pp. 103–114, 2017.

[12] A. Rezgui, L. Gerbaud, and B. Delinchant, "VHDL-AMS electromagnetic automatic modeling for system simulation and design", IEEE Trans. Magn, vol. 50, pp. 1013–1016, 2014.

[13] Ó. Lucía, I. Urriza, L. Barragan, D. Navarro, Ó. Jiménez, and J. M. Burdío, "Real-time FPGA-based hardware-in-the-loop simulation test bench applied to multiple-output power converters", IEEE Trans. Ind. Appl, vol. 47, pp. 853–860, 2011.

[14] B. Lu, X. Wu, H. Figueroa, and A. Monti, "A low-cost real-time hardware-in-the-loop testing approach of power electronics controls", IEEE Trans. Ind. Electron, vol. 54, pp. 919–931, 2007.

[15] M. Matar, D. Paradis, and R. Iravani, "Real-time simulation of modular multilevel converters for controller hardware-in-the-loop testing", IET Power Electron, vol. 9, pp. 42–50, 2016.

[16] Ó. Jiménez, Ó. Lucía, I. Urriza, L. A. Barragan, D. Navarro and V. Dinavahi, "Implementation of an FPGA-Based Online Hardware-in-the-Loop Emulator Using High-Level Synthesis Tools for Resonant Power Converters Applied to Induction Heating Appliances", IEEE Trans. Ind. Electron, vol. 62, pp. 2206–2214, 2015.

[17] A. Penczek, R. Stala, L. Stawiarski, and M. Szarek, "Hardware-in-the-Loop FPGA-based simulations of switch-mode converters for research and educational purposes", Przeglad Elektrotechniczny, vol. 87, pp. 194–200, 2011.

[18] C. Foucher and A. Nketsa, "A Dynamic FPGA-based Hardware-in-the-Loop Co-simulation and Prototype Testing Platform", Proceeding of The Tenth International Conference on Systems, Barcelona, pp. 68–73, 2015.

[19] P. V. Nikitin, C.-J. R. Shi, "VHDL-AMS based modeling and simulation of mixed-technology microsystems: a tutorial", The VLSI Journal, vol. 40, pp. 261–273, 2007.

[20] Dolphin Integration. SMASH Simulator. http://www.dolphin.fr/index.php/eda_solutions/products/smash/overview.

[21] Product Guide, "LogiCORE IP AXI Ethernet Lite MAC (v1.01.b). Xilinx®. Retrieved from https://www.xilinx.com/support/documentation/ip_documentation/axi_ethernetlite/v1_01_b/ds787_axi_ethernetlite.pdf

[22] V. Sklyarov, I. Skliarova, A. Barkalov, and L. Titarenko, Synthesis and Optimization of FPGA-Based Systems, Springer Science & Business Media, 2014

Partial Greedy Algorithm to Extract a Minimum Phonetically-and-Prosodically Rich Sentence Set

Fahmi Alfiansyah[1]
School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu
Bandung, West Java, Indonesia 40257

Suyanto[2]
School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu
Bandung, West Java, Indonesia 40257

***Abstract*—A phonetically-and-prosodically rich sentence set is so important in collecting a read-speech corpus for developing phoneme-based speech recognition. The sentence set is usually searched from a huge text corpus of million sentences using the optimization methods. One of the commonly used optimization methods for this case is a Least-to-Most Greedy (LTMG) algorithm. It is effective in minimizing the number of phoneme-units. Unfortunately, it does not distribute their frequencies. In this paper, a new method called Partial LTMG algorithm (PLTMG) is proposed to search an optimum set containing triphones and prosodies those are distributed in a near-uniform fashion. Testing on an Indonesian text corpus of ten million sentences crawled from some websites of newspapers and novels shows that the proposed method is not only capable of minimizing both phoneme-units and prosodies but also effective in distributing their frequencies.**

***Keywords*—Automatic speech recognition; minimum sentence set; prosody; speech corpus; triphone**

I. Introduction

Before 2014, an Automatic Speech Recognition (ASR) or Computer Speech Recognition generally has three components, i.e. acoustic model, pronunciation or word lexical model, and language model. Most ASR systems use the statistical approaches with supervised learning. The Hidden Markov Model is a commonly used method to train both acoustic and language models.

In 2014, many researchers start to develop an End-to-End Automatic Speech Recognition (E2EASR), which trains those three components of ASR in a single model [1], [2], [3], and [4]. The E2EASR does not need both pronunciation and language models commonly used in the previous conventional ASR. Hence, it can be embedded into a microdevice since it does not consume a high memory.

The first effort to build an E2EASR system is conducted by some researchers in 2014 using a classification-based approach called Connectionist Temporal Classification (CTC) (Graves 2014). This system consists of a layer of CTC and a Recurrent Neural Networks (RNN), which is abbreviated CTC-RNN. This system learns two components of ASR, pronunciation and acoustic models. However, this system gives many spelling mistakes so that it needs an external language model separately. In [2], some researchers at Baidu Research build the Deep Speech 2, an E2EASR that is successfully applied to English and Mandarin in 2015.

In 2016, some researchers from CMU, Google Brain, and University of Montreal propose an attention-based ASR model. The model is called ”Listen, Attend, and Spell” (LAS) [5], [6], and [7]. Unlike the CTC-based ASR, this LAS model is capable of learning all ASR components (acoustic models, pronunciation models, and language models) simultaneously. Hence, this LAS is the first fully E2EASR model with no external language model. In [8], the researchers consider that the LAS system is a more successful model than the CTC-based systems.

In 2017, the researchers from CMU, MIT, and Google Brain develop Latent Sequence Decompositions (LSD) that directly outputs the sub-word units, which are not only wider but also more natural than characters [9]. In early 2018, the researchers from Johns Hopkins University, Baltimore, USA, develop a new architecture called multi-modal data augmentation network (MMDA) that supports multi-modal inputs (acoustic and symbolic). The MMDA seeks to avoid the use of external language models with a much smaller combined text corpus and speech corpus to train the E2EASR [10]. Hence, a well-designed speech corpus is very important to train a high-performance E2EASR.

Many methods have been proposed to design a speech corpus, such as described in [11], [12], [13], [14]. The methods generally use a sub-word unit called triphone, i.e. a sequence of three contextual phonemes. A triphone is commonly written using a format L-X+R, where X is a target, L and R are a prefix and a postfix of the target respectively. Three samples of converting different types of sentences in Bahasa Indonesia (declarative, interrogative, and imperative/exclamatory) into cross-word triphone forms are listed in Table I: ”Aku pergi.” (”I go.”) pronounced as /ku p@rgi./, ”Apa kabar?” (”How are you?”) pronounced as /p kbr?/, and ”Ambil itu!” (”Take it!”) pronounced as /mbi itu!/, where /sil/ is a silence. All Indonesian phonetic symbols described in [15], which are based on the International Phonetic Alphabet (IPA), are adopted in this paper.

TABLE I. Conversion of Three Types of Sentences into Triphone Forms, where /sil/ is a Silence

Sentences	Triphone Forms
Aku pergi.	sil-+k -k+u k-u+p u-p+@ p-@+r e-r+g r-g+i g-i+. i-.+sil
Apa kabar?	sil-+p -p+ p-+k -k+ k-+b -b+ b-+r -r+? r-?+sil
Ambil itu!	sil-+m -m+b m-b+i b-i+ i-+i -i+t i-t+u t-u+! u-!+sil

Developing a read-speech corpus needs a well-designed text of transcription to be read by hundreds or even thousands of varying speakers based on their ages, accents, dialects, and genders [14], [16], [17]. The text of transcription is commonly a minimum phonetically-and-prosodically rich sentence set searched from a huge text corpus. Why prosody? The prosody affects how a speech sentence is being interpreted [18], [13]. For example, two sentences ”This is mine.” and ”This is mine?” have different prosodies (intonations) and consequently have different interpretations.

One of the effective optimization methods to find an optimum set is a Modified LTMG (MLTMG) that is proposed in [19]. Unfortunately, this algorithm just minimizes the phoneme-units but does not care to balance their frequencies. Hence, in this paper, a new method called Partial LTM Greedy algorithm (PLTMG) is proposed to search a phonetically-and-prosodically rich sentence set with balanced frequencies from an Indonesian text corpus.

II. Related Work

A speech corpus can be generally developed using either a phonetically-balanced or a phonetically-rich text corpus [20]. A phonetically-balanced corpus is a sentence set that follows Zipfian’s law, where each triphone is represented proportionally to its frequency. This corpus is not good enough to build an ASR, a speech synthesizer, or a pronunciation quality assessment. In contrast, a phonetically-rich text corpus that is a uniform triphone representation gives more accurate results for those tasks. It has a high variety of triphones in a sentence that uniformly distributed regardless their appearances in a language.

Many optimization methods have been proposed to develop a phonetically-rich text corpus. They are commonly based on either a greedy approach, such as described in [21] and [22], or an evolutionary computation as described in [23]. However, the greedy-based approach is more widely used in practice since it provides a much faster processing time as well as a higher scalability.

In [21], the researchers show that an LTMG is capable of extracting smaller sentence sets and fewer computation costs than the other standard greedy algorithms. But, this algorithm has two problems. Firstly, it just selects a to-be-covered unit randomly when there are some units have the same frequencies. Secondly, it may produce redundant sentences as the covering score is computed based on a set of to-be-covered units updated by the previous selection. In [19], the researcher proposes an MLTMG to solve both problems by 1) collecting all sentences those contain the same frequencies to-be-covered units into a subset, then select the best one from it and 2) evaluating each sentence in the extracted minimum-so-far set to check its redundancy.

Unfortunately, the MLTMG also has two drawbacks. When some sentences containing a to-be-covered unit with the same scores, it cannot choose the best one. When some sentences have the same scores but different to-be-covered units or frequencies, it just randomly selects a sentence without other calculation nor consideration. The MLTMG extracts a optimum set by sequentially selecting the best sentences in a greedy way based on a ratio-based scoring formula. Besides, as explained in [19], it is just evaluated using a relatively small motherset of 500 k sentences without considering any prosody. Hence, in this research, the MLTMG is improved by proposing some new procedures to handle a much bigger motherset of 10 M sentences with considering the prosodies.

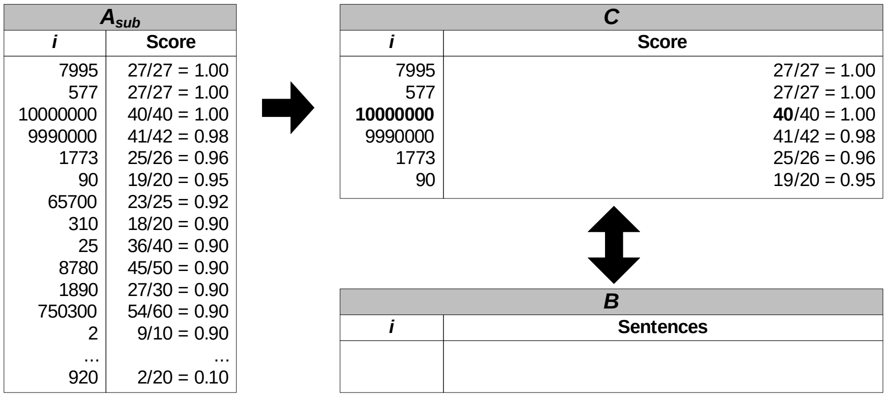
III. Proposed Partial LTM Greedy

Here, the MLTMG is improved by taking into account the number of to-be-covered triphones as well as their frequencies before selecting a sentence so that this method is called a PLTMG. This new method is simply implemented by replacing the step 5 in the MLTMG in [19] with four new steps to become:

1) Let A be a mother sentence set, B be an empty set, and U be a list of all to-be-covered unique triphone tokens sorted by their frequency in ascending order;
2) Select all infrequent triphones (those have the least frequency) from U and then store them in a subset U_{sub};
3) Select all sentences from A those contain at least one triphone in U_{sub} and put them in a subset A_{sub};
4) For each sentence in A_{sub} calculate its score using a formula in Eq. (1)
$$S_i = \frac{V_i}{T_i}, \tag{1}$$
where V_i and T_i are the number of to-be-covered triphones and the total triphone tokens in the ith sentence respectively;
5) Sort the scores of sentences in ascending order;
6) Define P, a small number between 0 and 1, that states a percentage of sentence scores selected to compete;
7) Take the top P percent of sentences having scores bigger than $(1 - P) \times thebestscore$ and then store them in a subset C;
8) From C select a sentence with the highest score. If there are two or more sentences with the same highest scores then select one containing the most to-be-covered triphones. If there are two or more sentences having the most to-be-covered triphones then choose a sentence containing the least frequent triphones in B. Delete all triphones appear in the selected sentence from both U and U_{sub}. Remove all sentences from C.
9) Do step 3 to 8 until U_{sub} is empty;
10) Do step 2 to 9 until U is empty.

Step 8 in the proposed PLTMG can be easily explained using two illustrations in Fig. 1 and Fig. 2. The Fig. 1 illustrates a case where there are three sentences with the same highest scores of 1.00, i.e. the sentence index of 7995, 577, and 10000000. Since the 10000000th sentence has the maximum number of to-be-covered triphones of 40 and B is initially empty, the sentence is selected as the best one.

Meanwhile, Fig. 2 illustrates a case where there are two sentences with the same highest scores of 1.00 as well as the highest number of to-be-covered triphones of 27. In this case, let the 7995th sentence contains ”***Pergi** jauh dariku, katanya*.” (Go away from me, he said) and the 577th sentence consists of ”*Kami mendapatkan ijazahnya*!” (We get the certificate!). Since B contains a sentence ”*Sudah lama ia tidak **pergi** ke rumah mertua di desa*.” (For a long time he does not go to

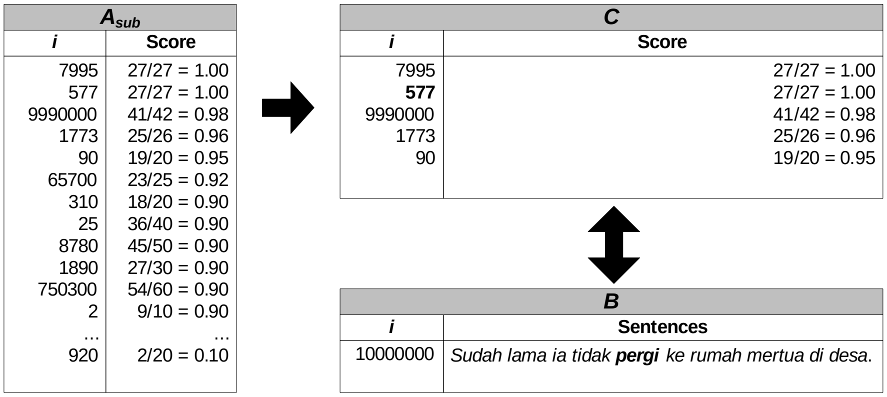


A_{sub}	
i	**Score**
7995	27/27 = 1.00
577	27/27 = 1.00
10000000	40/40 = 1.00
9990000	41/42 = 0.98
1773	25/26 = 0.96
90	19/20 = 0.95
65700	23/25 = 0.92
310	18/20 = 0.90
25	36/40 = 0.90
8780	45/50 = 0.90
1890	27/30 = 0.90
750300	54/60 = 0.90
2	9/10 = 0.90
...	...
920	2/20 = 0.10

C	
i	**Score**
7995	27/27 = 1.00
577	27/27 = 1.00
10000000	**40**/40 = 1.00
9990000	41/42 = 0.98
1773	25/26 = 0.96
90	19/20 = 0.95

B	
i	**Sentences**

Fig. 1. PLTMG for a case where there are two or more sentences with the same highest scores but there is only one sentence has the highest number of to-be-covered triphones



A_{sub}	
i	**Score**
7995	27/27 = 1.00
577	27/27 = 1.00
9990000	41/42 = 0.98
1773	25/26 = 0.96
90	19/20 = 0.95
65700	23/25 = 0.92
310	18/20 = 0.90
25	36/40 = 0.90
8780	45/50 = 0.90
1890	27/30 = 0.90
750300	54/60 = 0.90
2	9/10 = 0.90
...	...
920	2/20 = 0.10

C	
i	**Score**
7995	27/27 = 1.00
577	27/27 = 1.00
9990000	41/42 = 0.98
1773	25/26 = 0.96
90	19/20 = 0.95

B	
i	**Sentences**
10000000	*Sudah lama ia tidak* ***pergi*** *ke rumah mertua di desa.*

Fig. 2. PLTMG for a case where there are two or more sentences with the same highest scores as well as highest number of to-be-covered triphones

the home of his father in law at the village.) with a word ”*pergi*”, then the 7995th sentence is not selected. Instead, the PLTMG chooses the 577th sentence as it has the least frequent triphones covered in B.

Using those steps, the PLTMG should produce a sentence set containing slightly more to-be-covered triphones but lower frequencies than the MLTMG. Besides, this algorithm is also expected to be capable of avoiding some sentences with the same scores in a competition so that the random selection used in the MLTMG can be removed.

IV. Result and Discussion

The text corpus used here is a set of 10 M sentences that is collected by crawling some Indonesian websites of news and novels as describes in [24]. The corpus covers three types of sentence: declarative (ended by ”.”), interrogative (ended by ”?”), and imperative/exclamatory (ended by ”!”). Based on the corpus, a phonemic dictionary of 128,779 words is generated by an automatic Indonesian G2P system described in [15]. Phonetizing each sentence in the mother set using the dictionary, and then converting the phonemic sequences into triphones, produce 289,096,873 triphone tokens and 18,909 unique triphones as listed in Table II. It means the ratio of unique triphone and the tokens is very low, only 0.000065.

Some experiments are performed using a personal computer of an i7 processor and 4 GB RAM to get the runtime of 5 hours per experiment. In the PLTMG, the variable P functions to select some sentences to compete. For example, if the best to-be-covered unit score on the iteration is 1 and P = 0.05 then the minimum score to compete will be 0.95. The PLTMG is tested using P = 0.05, 0.1, and 0.2 to see its behavior in extracting the mother set.

TABLE II. Statistics of the Mother Sentence Set

Total number of sentences appear	10,000,643
Number of declarative sentences	9,938,093
Number of interrogative sentences	50,314
Number of imperative/exclamatory sentences	12,236
Total number of words appear	47,590,317
Number of distinct words	128,779
Number of triphone tokens	289,096,873
Number of unique triphones	18,909
Average number of triphones per sentence	28.91

The experimental results in Table III proves that the proposed PLTMG is effective to decrease the standard deviation of triphone frequencies, where the standard deviation decreases by around 0.34 on each specified P. However, the number of triphones are higher than those produced by the Modified LTM Greedy. The PLTMG with P = 0.20 produces much more triphone tokens (up to 170,108) than the MLTMG. The PLTMG with P = 0.10 reaches an optimum sentence set. It produces a lower standard deviation than the PLTMG with P = 0.05 and fewer triphone tokens than the PLTMG with P = 0.20. In addition, the triphone frequencies on the PLTMG decrease

TABLE III. STATISTICS OF THE OPTIMUM SENTENCE SETS EXTRACTED BY MLTMG AND PLTMG ALGORITHMS

Algorithm	#triphones	#sentences	Avg. triph. freq.	Std. triph. freq.
MLTMG	165,673	7,334	8.76	30.42
PLTMG, $P = 0.05$	166,527	7,286	8.80	30.08
PLTMG, $P = 0.10$	167,604	7,263	8.86	29.74
PLTMG, $P = 0.20$	170,108	7,206	8.99	29.39

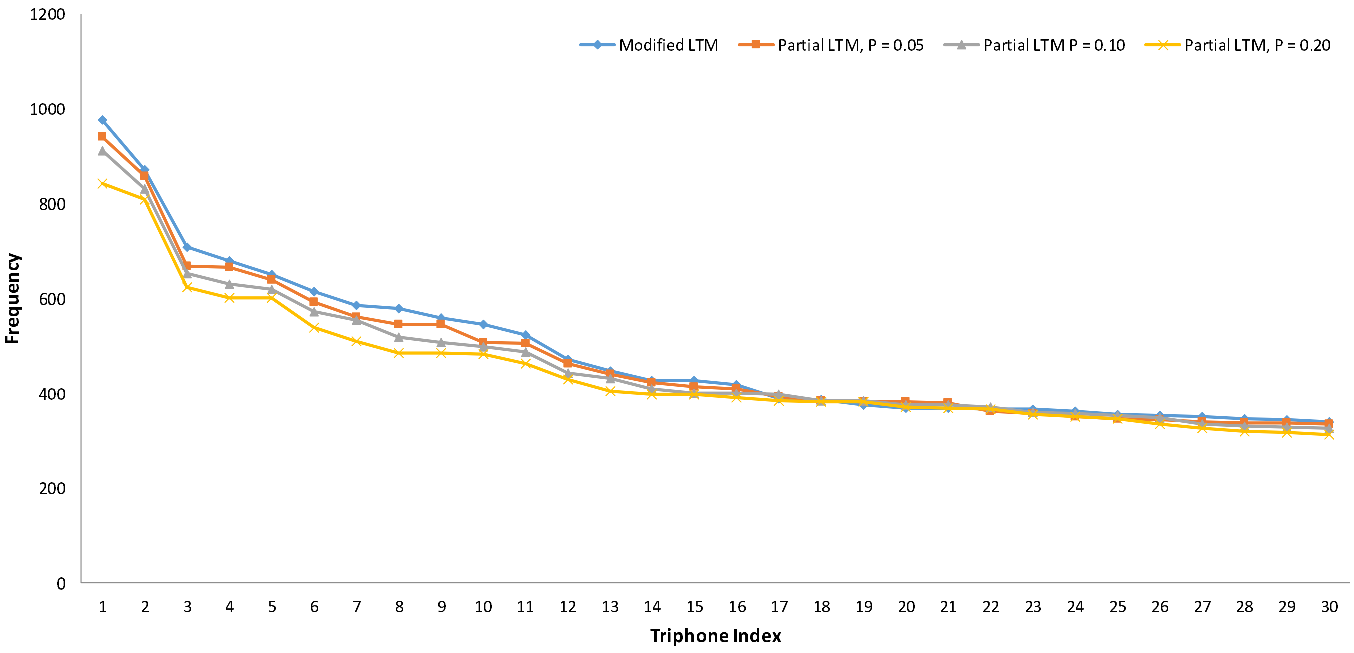


Fig. 3. The first thirty most-frequent triphones

as illustrated in Fig. 3. It shows that there are differences in frequencies at the beginning of the largest triphones since the PLTMG takes into account the number of triphone frequencies in a sentence to find the lowest frequency in the sentence.

V. CONCLUSION

The proposed PLTMG is effective to produce a sentence set that contains more uniformly distributed triphones than the previous MLTMG. The value of P affects the number of triphones as well as their standard deviations. The greater P the lower standard deviation. Unfortunately, the bigger P the more triphones selected. However, the PLTMG enables a user to make any adjustment to get the optimum extracted sentence set. In the future, the user can also apply the PLTMG to a much bigger motherset of hundreds of millions or even billions of sentences to get much more unique triphones.

ACKNOWLEDGMENT

We would like to thank all colleagues in the School of Computing, Telkom University, for the great support and suggestions.

REFERENCES

[1] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep Speech: Scaling up end-to-end speech recognition," pp. 1–12, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[2] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," pp. 1–28, 2015. [Online]. Available: http://arxiv.org/abs/1512.02595

[3] A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng, "Lexicon-Free Conversational Speech Recognition with Neural Networks," *Proceedings the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015. [Online]. Available: http://deeplearning.stanford.edu/lexfree/

[4] A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng, "Building DNN acoustic models for large vocabulary speech recognition," *Computer Speech & Language*, vol. 41, pp. 195–213, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230816301930

[5] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," pp. 1–16, 2015. [Online]. Available: http://arxiv.org/abs/1508.01211

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[8] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," 2016.

[9] W. Chan and Y. Zhang, "Latent Sequence Decompositions," pp. 1–12, 2017.

[10] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, “Multi-Modal Data Augmentation for End-to-end ASR,” in *Interspeech*, 2018. [Online]. Available: http://arxiv.org/abs/1803.10299

[11] M. Pinnis, A. Salimbajevs, and I. Auziņa, “Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian,” in *The Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016, pp. 775–780.

[12] D. Koržinek, K. Marasek, Ł. Brocki, and K. Wołk, “Polish Read Speech Corpus for Speech Tools and Services,” in *CLARIN*, 2017, pp. 54–62.

[13] S. M. Hosseini and H. Sameti, “Creating a corpus for automatic punctuation prediction in Persian texts,” in *2017 Iranian Conference on Electrical Engineering (ICEE)*, may 2017, pp. 1537–1542.

[14] H. Abera and S. H/Mariam, “Design of a Tigrinya Language Speech Corpus for Speech Recognition,” in *Workshop on Linguistic Resources for Natural Language Processing*, vol. 9, 2018, pp. 78–82.

[15] S. Suyanto, S. Hartati, and A. Harjoko, “Modified Grapheme Encoding and Phonemic Rule to Improve PNNR-Based Indonesian G2P,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 3, pp. 430–435, 2016.

[16] C. Kurian, “Development of Speech corpora for different Speech Recognition tasks in Malayalam language,” in *International Conference on Natural Language Processing*, no. December, 2015, pp. 229–236.

[17] D. Arnold, F. Tomaschek, K. Sering, F. Lopez, and R. H. Baayen, “Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit,” *PLOS ONE*, vol. 12, no. 4, pp. 1–16, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0174623

[18] N. Moore, “What ’ s the point ? The role of punctuation in realising information structure in written English,” *Functional Linguistics*, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40554-016-0029-x

[19] Suyanto, “Modified Least-to-Most Greedy Algorithm to Search a Minimum Sentence Set,” in *IEEE TENCON*, 2006.

[20] G. Mendonça, S. Candeias, F. Perdigão, C. Shulby, R. Toniazzo, A. Klautau, and S. Aluísio, “A method for the extraction of phonetically-rich triphone sentences,” in *2014 International Telecommunications Symposium (ITS)*, 2014, pp. 1–5.

[21] J.-s. Zhang and S. Nakamura, “An Efficient Algorithm to Search For A Minimum Sentence Set For Collecting Speech Database,” in *ICPhS*, 2003, pp. 3145–3148.

[22] K. Arora, S. Arora, K. Verma, and S. S. Agrawal, “Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages,” in *INTERSPEECH*, 2004.

[23] M. Nicodem, I. Seara, R. Seara, D. Anjos, and R. Seara-Jr, “Selecao automatica de corpus de texto para sistemas de sıntese de fala,” in *XXV Simposio Brasileiro de Telecomunicacoes (SBrT)*, 2007.

[24] B. Nugroho and B. Nurtomo, “Greedy Algorithms to Optimize a Sentence Set Near-Uniformly Distributed on Syllable Units and Punctuation Marks,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 10, pp. 291–296, 2018.

NB-IoT Pervasive Communications for Renewable Energy Source Monitoring

Farooque Hassan Kumbhar
Department of Computer Science
National University of Computer and Emerging Science
Pakistan

***Abstract*—Renewable sources like solar and wind energy have seen a drastic increase in the market, especially in developing countries where electricity prices are high and QoS and QoE, both are at their lowest. In this paper, we innovate by proposing a paradigm of smart off-grid from sensing using an Internet of Things (IoT) based smart meter for continuous monitoring, to reporting a daily user on their smart devices using IoT middleware. Our proposed smart off-grid system keeps track of the performance and faults of the off-grid equipment. Under communication technology scrutiny, we model 3GPP Narrow Band IoT (NB-IoT) collision and success probability of grouping smart meter communications to avoid random access channel (RACH) congestion. The proposed smart off-grid communications outperform existing systems and achieve 1.3 to 20 times higher SINR, more than 30 Mbps data rate in 4G, three times higher data rate in NB-IoT,** 25% **fewer collisions and** 25% **higher success rate.**

Keywords*—*NB-IoT; smart off-grid; RACH; 4G LTE

I. Introduction

Low cost and reliable energy sources have always been and will always be a major part of human interest. Users and investors are moving towards off-grid solutions like nuclear, wind and solar, powering 12-volt appliances from a bulb to an air-conditioner. Currently (in 2017), the worldwide installed RES capacity led by China accounts for more than 1500 gigawatt[1], as shown in Figure 1. Cheaper and easily manageable energy solutions attract more and more small-scale investors, suppliers and distributors, challenging the monopoly of traditional government electrical grids [1]. These off-grid systems not only produce energy but also reduce transmission losses and cost of production, distribution, and maintenance [2]. Work in [3] discusses a relay selection scheme for cellular networks, powered by green energy sources. However, the major goal is to reduce power consumption and dependency of cellular relay stations from traditional grids, the work is related to our proposed aggregation scheme for cellular network relays.

We propose smart meter based smart off-grid monitoring of RES using a number of sensors like moisture, light, motion, humidity, production, etc. and IoT middleware services/ servers. Installation of a number of smart meters, especially in urban areas requires communications network providers. Wi-Fi, Bluetooth, ZigBee, and other short-range communication technologies lack in range, internet, and prevalent service. On the other hand, the cellular networks have a widespread network with tons of base stations (BSs) having the range in kilometers and can be useful assets for the IoT communications. Moreover, the 3GPP standardized Narrow Band IoT (NB-IoT) with 4G Long Term Evolution (LTE) coverage characteristic with lower power consumption. However, the inherent random access channel (RACH) challenge persists. In NB-IoT communications, the device initializes the process after getting PRACH information from the SIB-2 message and transmits continuous repeated RACH. The BS responds by sending RAR message and receives message 3 from the device. The BS sends contention resolution message to the device and initiates subsequent communications and resource assignments. In case of any message failure, the device resends preamble after 12 ms. Unlike LTE, the RACH process messages are repeatedly sent between device and BS, illustrated in Figure 2. However, the collision or contention occurs similar to the LTE, if two or more than two devices send a request on the same randomly selected RACH. The contention burden is increased in the NB-IoT with repeated transmission occupying resources.

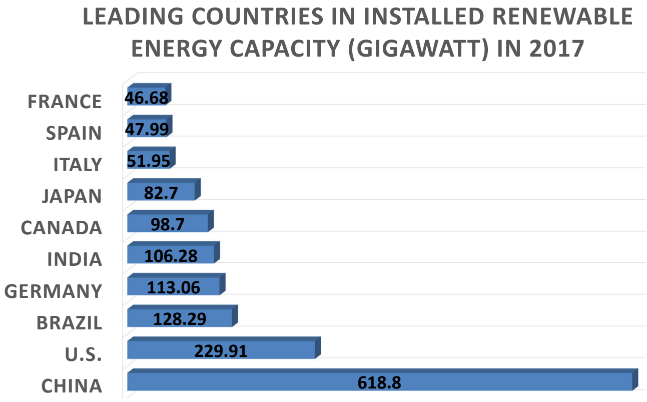


Fig. 1. Renewable energy capacity statistics

Authors in [4] describe that a device can withdraw its next scheduled transmission message if it gets unmatched Time Advance (TA) information of RAR and avoids possible request collision. However, there is still a lot of room to improve and reduce additional delays for the devices. Exhaustive study and observation of existing literature deduce that there is a need for an adequate and suitable architecture for smart RES management and control with continuous communications, which is not yet proposed. However, 5G network is embarking with a plethora of data rate to mobile devices but the delay intolerant IoT networks like smart grids and smart meters can make use of the 4G and NB IoT networks[5, 6]. On the

[1] IRENA - Renewable Energy Capacity Statistics 2018, page 2-5

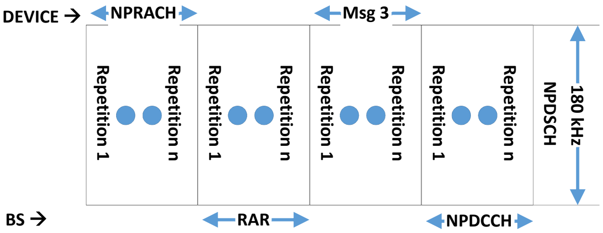


Fig. 2. NB IoT RACH procedure and structure

other hand, timely network connectivity becomes a critical need in a disaster situation [7]. We are the first to design and propose an outright architecture which not only discusses smart meter design and monitoring for RES but also tackles the communication challenges. Our major contribution and innovation over existing systems are:

1) design a smart meter and an innovative architecture for renewable sources
2) NB-IoT contention model with continuous monitoring of RES
3) higher signal-to-interference and noise (SINR) value by 20 times and reduced collisions and increase success by 25% over existing systems.

The rest of the article is organized as follows: Section 2 exhaustively discusses existing literature work on smart grids/ off-grid renewable systems and existing literature for the 4G LTE RACH and NB-IoT RACH issues.

Section 3 presents the proposed solution with smart meter design and proposed aggregation scheme. We analytically model SINR, collision and success probabilities with proof of concept. Performance evaluation and results are analyzed in Section 4. Section 5 concludes the paper.

II. Literature Review

In this section, study of exhaustive literature is discussed, providing useful insights on what has been achieved and things that are lacking in earlier smart off-grid systems and 4G LTE RACH solutions.

A. Smart Meter and Off-Grid

There are numerous aspects in our take on the smart off-grid system like designing a smart meter for continuous monitoring and contributing with an intelligent and distributed architecture for control of RES. Existing literature is sporadically diverse in the areas like the design of micro-grids, distribution planning of RES, decentralize control, energy scheduling, etc.

Research work on small-scale RES in [1] focus on contractual trading of stored energy for conflicts avoidance using energy informatics. The authors model a coalitional game for direct trading among suppliers with fair revenue division. However, the trading requires a centralized entity as an aggregator in the main micro-grid station. Authors in [2] propose a micro-grid design for RES and autonomous control overcharging or discharging of energy storage units. The proposed automation envisions to extract maximum power from the RES and provide quality power to the user. However, we plan to extend and provide deeper insights using a number of sensors and IoT middleware. Work in [3] discusses a relay selection scheme for cellular networks, powered by green energy sources. However, the major goal is to reduce power consumption and dependency of cellular relay stations from traditional grids, the work is related to our proposed aggregation scheme for cellular network relays. [8] introduce a multi-objective and multi-level model for the distribution system with RES and energy storage. The authors model RES and energy storage planning as an optimization problem using modified Pareto-based particle swarm optimization (PSO). Our proposed scheme can benefit [8] by providing pervasive and continuous monitoring and control. Another study in [9] model penetration of RES as multi-generator interconnected power network. The authors utilize a decentralized adaptive neural network feedback controller to stabilize dc link voltage oscillations during grid turbulence. The simulation and performance evaluation use IEEE 14-bus power system for damping oscillations after disturbances. However, the study lacks in automation and in continuous control by the users.

Flexible supply-demand management in [10] focus on optimal energy scheduling for residential smart grids. Authors provide a solution for cost-effective energy scheduling with a centralized RES. The study highlights the trade-off between RES energy and associated cost, and volatility of RES for optimal exploitation. However, a major drawback of the study is to consider centralized renewable sources which question the practicality of the work. [11] utilizes a quality of experience (QoE) based approach for RES management in the residential environment. Authors propose profile based QoE aware appliance scheduling and RES power allocation. A central controller communicates with individual smart meters to change operational state on or off of the appliances. Authors in [12] assume a scenario of a number of RES in a centralized network and propose to balance RES in a micro-grid using meteorological forecast for next 24 hour and plant location. The solution depends on the weather forecast to enable most productive RES for energy generation. However, unlike [12], we consider communication challenges of continuous monitoring and insights of the RES using middleware [13].

B. Random Access in LTE

Inherently, NB-IoT utilizes Orthogonal Frequency Division Multiple Access (OFDMA) for resource sharing over RACH. In an OFDMA slot, each device randomly selects a RACH and in a case of multiple devices selecting the same RACH, a collision or contention happens [14]. Figure 2 outlines the repetitive access communication flow between device and BS. Initially, a device chooses a RACH and requests for resources. The BS responds with a random access response (RAR) message, containing a unique identifier. In the subsequent scheduled transmission, the BS realizes that two or more than two devices have been assigned the same RACH. The BS drops the requests and the devices are required to request again after a random back-off period. This adds delays in the communications resulting in user frustration and could cause catastrophic results in delay-intolerant applications like vehicular or medical. With the increase in the access intensity the chances of collision increase.

TABLE I. LITERATURE REVIEW OF SMART RES AND RACH SOLUTIONS

Literature	**Research Focus**	**Major Details**
Smart Off-Grid Solutions		
Zhiyong Li et.al [1]	Small scale RES trading	• Fair revenue and division model using coalitional game formulation • Provides optimal consumption, prices and profit results
S.K. Tiwari et.al [2]	Design and control of micro-grid	• Autonomous RES control system • Automatic charging and discharging
Hui-Ju Hung et.al [3]	Cellular relay selection with green energy	• Relay selection • Reduced power consumption of relay stations. • Use of green energy for relays.
Rui Li et.al [8]	Cooperative distribution system	• RREs and energy storage planning and distribution • Optimal solution using Pareto-based PSO
Shaghayegh Kazemlou et.al [9]	Decentralized RES control	• Decentralized controller for RESs and energy storage units • Neural network controller
Yuan Wu et.al [10]	Optimal energy scheduling	• Cost effective and optimal exploited RES system • Centralized RESs and energy storage units
Virginia Pilloni et.al [11]	Smart home energy management	• Smart meter based appliance energy controller • Profile based approach for energy management
Mattia Marinelli et.al [12]	Predictive control strategy for RES	• Weather based forecast and RES control • Day ahead energy planning
Existing RACH solutions		
Kab Seok Ko et.al [4]	Time alignment Matching	• Collision avoidance. • Applicable only for overlapping area.
Farhadi et.al[15]	Group based signaling (aggregation)	• Device aggregation to reduce RACH competition • Frequency reuse utilization. • Possesses grouping overhead.
Chang et.al [16]	Machine-to-Machine data gathering	• A novel data perspective of Machine-to-Machine Communication. • Incompatibility with large number of Machine-to-Machine.
Zheng et.al [17]	Prioritized Human and Machine Type Communications	• Extreme prioritization techniques • Each techniques focuses only one type of communications • Causes longer delays.
Huasen et.al [18]	Interrupted Poisson Distribution	• Active devices estimation in cellular system • Device barring and congestion reduction. • Causes longer delays.
Shao-Yu Lien et.al [19]	Cooperative access barring	• Barring parameter selection using Multiple eNB information • Possesses less overhead. • Applicable only for overlapping area.
Tzu-Ming Lin et.al [20]	Dynamic ACB with device classification	• Five categories for incoming traffic • Dynamic access class barring utilization. • Strategic, static approach.
Hasan et.al [21]	Q learning at device end	• Utilization of Q Learning in device • Collision avoidance.

Authors in [4] describe that a device can withdraw its next scheduled transmission message if it gets unmatched Time Advance (TA) information of RAR and avoids possible request collision. Work in [15] propose to reduce uplink requests by aggregating device requests. Initially, each device communicates to one another and selects a group delegate which corresponds to BS. Frequency reuse in the different group of devices increases the spectral efficiency within the same BS. However, the solution requires a great deal of frequency reuse and grouping management. Another similar work in [16] groups the devices and aggregate the uplink requests to reduce contention. The article models the size of the group as the NP-Hard problem and suggests two approaches, Cross Entropy-based randomized approach, and Tabu Search. However, the solution is more focused on the group size problem than reducing RACH contentions. Another research in [17] evaluates two scenarios of Prioritized Human-Type-Communications and Prioritized Machine-Type-Communications to highlight the extreme communication procedures.

Authors of [18] discuss an Interrupted Poisson Distribution estimation approach for active user's calculation in the network. The number of active users, help in reducing RACH competition by device barring in the subsequent slot. On the other hand, the barred or delayed devices suffer from additional delays. [19] proposes a barring approach, where Access Class Barring (ACB) is improved to a cooperative design. Cooperative ACB gains 30% higher success over ACB, but a number of devices suffer from additional delays. Work in [20] outline five classes of incoming traffic and applies Dynamic Access Barring (DAB) mechanism to reduce RACH competition. The solution prioritizes the devices and adds delays to the traffic in three different scenarios of low, medium and high. [21] suggests Q-Learning experience based BS selection by the devices. The solution implicates devices in overlapping areas to have prior knowledge of BS. Standardization organizations like 3GPP have included ACB and Extended ACB mechanism to reduce collisions in LTE cellular networks [22]. However, there is still a lot of room to improve and reduce additional delays for the devices.

Table I outlines existing literature on state of the art solutions and ideas related to the smart grid and RESs. The table also includes existing solutions to tackle RACH congestion in communications resource access. Exhaustive study and observation of existing literature deduce that there is a need for an adequate and suitable architecture for smart RES management and control with continuous communications, which is not yet proposed. The strength of most of the existing solution is to enable a barring or prioritized system. We are the first to design and propose an outright architecture which not only discusses smart meter design and monitoring for RES but also tackles the communication challenges.

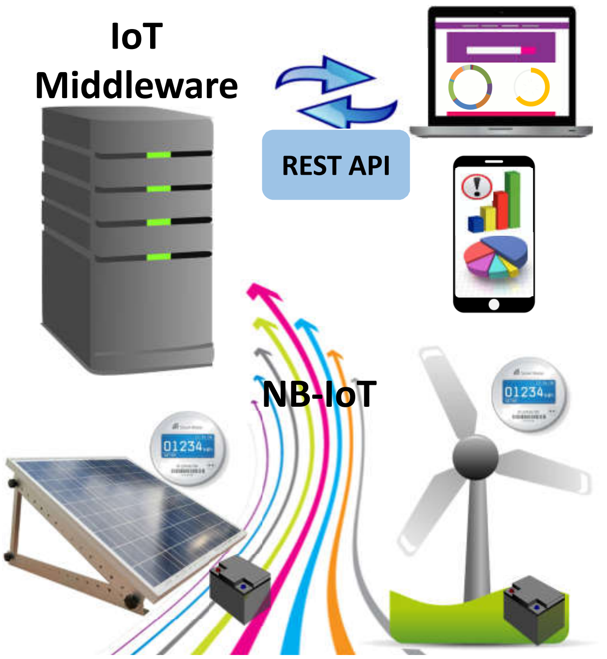


Fig. 3. Proposed paradigm of smart off-grid

III. Smart Off-Grid Solution using IoT

Our design of smart meter considers a number of sensors like humidity, motion, etc. for continuous monitoring of the RES equipment. We propose that the information is then fed to a middleware service like ThingSpeak[2], which not only provides data storage but also generates interactive graphs for deeper insights. Moreover, a mobile application or a web app can utilize ThingSpeak REST API to read the accessible and authorized data. Each communication message between a user and an RES equipment requires middleware server (ThingSpeak) and communications network. Figure 3 highlights middleware communications with the RES monitoring architecture over NB-IoT and 4G. The number of smart meters increases the access intensity in NB-IoT and 4G LTE RACH causing delays and the proposed scheme counters that by aggregating several smart meter requests. We propose that each device identifies itself as an aggregator using a boolean check variable Aggregator. If the device is an aggregator, it accepts the data, accumulates all messages and requests for RACH. On the contrary, the device broadcasts request to nearby devices, selects the first response of candidate aggregator and send data for accumulation.

A 4G BS shares resources using a number of random preambles which are accessed by the devices. Let V={v_1, v_1, ..., v_k} be the total number of k devices, cuncurrently requesting to the 4G BS for the (M) RACH resources. Assuming that each of the devices compete with an equal opportunity, the collision probability (P_α) and success probability (P_β) in the legacy network is desribed by [4] as:

$$P_\alpha = 1 - \left(1 - \frac{1}{M}\right)^{k-1} \tag{1}$$

Algorithm 1 Smart meter aggregation algorithm

```
if Aggregator==True then
    Receive and acknowledge messages from neighboring smart meters
    Request for RACH
    Aggregate and send data
else
    Broadcast aggregator request and store Acknowledgments.enqueue(acknowledging address)
    if Acknowledgments=! $\emptyset$ then
        Send data to device at Acknowledgments.dequeue()
    else
        Aggregator=True
    end if
end if
```

$$P_\beta = \left(1 - \frac{1}{M}\right)^{k-1} \tag{2}$$

Moreover, the 4G LTE uplink Signal-Interference-plus-Noise-Ratio (SINR) (δ_α) of a smart meter with k other devices with P transmission power, g channel gain, and μ Additive White Gaussian Noise off a quasi-static Rayleigh fading channel, is described in [5] as:

$$\delta_\alpha = \frac{gP}{\mu + \sum_{n=1}^{k-1} g_n P_n}, \tag{3}$$

where interference by other k devices is represented by $\sum_{n=1}^{k-1} g_n P_n$.

The data rate (ψ_α) with BW bandwidth of 4G network, can be estimated as:

$$\psi_\alpha = BW \log(1 + \delta_\alpha) \tag{4}$$

Let R_B be the coverage radius of a BS and R_S be the range of an aggregating smart meter. If another smart meter is randomly placed in the radius of the BS then the probability of its placement within the range of an aggregator can be calculated as $P_\rho = \frac{\pi {R_S}^2}{\pi {R_B}^2}$. We can extend the probability for randomly placed λ aggregators, as:

$$P_\rho = \frac{\sum\limits_{a=1}^{\lambda} \pi {R_{S,a}}^2 - \left[\left(\sum\limits_{b=1}^{\lambda}\sum\limits_{c=1}^{\lambda} \pi {R_{S,b}}^2 \cap \pi {R_{S,c}}^2\right)/2\right]}{\pi {R_B}^2} \tag{5}$$

Because, the aggregators are randomly placed, they can overlap each other. Above equation first calculates total radius of all λ aggregators as $\sum\limits_{a=1}^{\lambda} \pi {R_{S,a}}^2$. Subsequently, the overlapping radius points are substracted using second part, $\left[\left(\sum\limits_{b=1}^{\lambda}\sum\limits_{c=1}^{\lambda} \pi {R_{S,b}}^2 \cap \pi {R_{S,c}}^2\right)/2\right]$. Combining both parts give us total favourable outcomes which then divided by total possible outcomes $\pi {R_B}^2$ estimates probability of a randomly

[2] ThingSpeak is the open IoT platform with MATLAB analytics. Online: [https://thingspeak.com/]

placed smart meter within the range of an aggregator. However, for better understanding and simplicity of complex equation, we assume that the aggregators are disjoint to each other, which makes second term of above equation $\left[\left(\sum\limits_{b=1}^{\lambda}\sum\limits_{c=1}^{\lambda}\pi {R_{S,b}}^2 \cap \pi {R_{S,c}}^2\right)/2\right]$ equal to 0. Moreover, it also puts a constraint that to total radius points of λ aggregators must not exceed the total possible outcomes, i.e. $\sum\limits_{a=1}^{\lambda}\pi {R_{S,a}}^2 < \pi {R_B}^2$.

Considering that there are k devices randomly placed within a BS, the total number of devices in λ aggregators can be estimated as:

$$\rho = \frac{\sum\limits_{a=1}^{\lambda}\pi {R_{S,a}}^2}{\pi {R_B}^2} \times k + \lambda \tag{6}$$

Considering that the proposed aggregation reduces the access intensity by ρ devices, the collision probability (P_γ) and success probability (P_ω) in the proposed architecture can be estimated as:

$$P_\gamma = 1 - \left(1 - \frac{1}{M}\right)^{k-1-\rho} \tag{7}$$

$$P_\omega = \left(1 - \frac{1}{M}\right)^{k-1-\rho} \tag{8}$$

Moreover, the reduced competition also reduces the interference in the network, ergo the SINR in the proposed system (δ_β) becomes:

$$\delta_\beta = \frac{gP}{\mu + \sum_{n=1}^{k-1-\rho} g_n P_n}\ , \tag{9}$$

Subsequently, better SINR (δ_β) increases the data rate (ψ_β) for a device in the proposed paradigm with same BW bandwidth. Mathematically:

$$\psi_\beta = BW \log(1 + \delta_\beta) \tag{10}$$

The proposed system provides lower collision and higher successful access to the devices in the presence of smart meters. Our claim stands if following hypothesis holds true: $X = \frac{Proposed\ Collision\ Probability}{Existing\ Collision\ Probability} = \frac{P_\gamma}{P_\alpha} > 1$. Replacing values of P_γ from Equation 7 and P_α from Equation 1 in the equation, gives:

$$\left[X = \frac{1 - \left(1 - \frac{1}{M}\right)^{k-1-\rho}}{1 - \left(1 - \frac{1}{M}\right)^{k-1}}\right] > 1 \tag{11}$$

Applying $log[X]$ and reducing:

$$\left[\begin{array}{l}\log[X] = (k-1-\rho)\times\log\left[1 - \left(1 - \frac{1}{M}\right)\right] - \\ (k-1)\times\log\left[1 - \left(1 - \frac{1}{M}\right)\right]\end{array}\right] > 0 \tag{12}$$

TABLE II. SYMBOLS AND SIMULATION PARAMETERS

Symbol	Description	Simulation Value
k	Number of users	25-250
M	Number of available resources (RACH)	64
BW	4G LTE, NB-IoT	20 MHz, 180 KHz
ρ	Number of aggregated users	N/A
λ	Number of aggregator	4-20
P_α	Collision probability in existing system	N/A
P_β	Collision probability in proposed system	N/A
P_γ	Success probability in existing system	N/A
P_ω	Success probability in proposed system	N/A
δ_α	SINR in the existing system	N/A
δ_β	SINR in the proposed system	N/A
ψ_α	data rate in the existing system	N/A
ψ_β	data rate in the proposed system	N/A
P	Transmission power	250mW
g	Channel gain	15 dBi
μ	Additive White Gaussian Noise	-101 dBm
R_B	Range of 4G base station (BS)	500m
R_S	Range of smart meter aggregator	50m

$$\left[\begin{array}{l}\log[X] = \log\left[1 - \left(1 - \frac{1}{M}\right)\right]\times \\ \left[(k-1-\rho) - (k-1)\right]\end{array}\right] > 0 \tag{13}$$

$$\left[\log[X] = \log\left[1 - \left(1 - \frac{1}{M}\right)\right] \times -\rho\right] > 0 \tag{14}$$

The term in above equation ($\log\left[1 - \left(1 - \frac{1}{M}\right)\right] \times -\rho$) will always provide a positive value where $\rho > 1$ and $M > 1$. It should be noted that in every possible scenario, a BS will have more than one resources (M) and more than one aggregations (ρ). Thus proving $\log[X] > 0 \Rightarrow X > 1 \Rightarrow P_\gamma > P_\alpha$, where $\rho > 1$ and $M > 1$. It is safe to assume that a similar hypothesis also holds true for success probability and SINR values.

IV. PERFORMANCE EVALUATION

Our Monte-Carlo simulation based experiments and results for the existing and proposed system includes a total of 25 to 250 devices and 64 RACH preambles (M) for 20 MHz (LTE) and 180 KHz (NB-IoT) bandwidth (BW) based 4G LTE BS with 500 m communications radius [23]. The number of aggregators is defined between 4 to 20, having 50 m radius each, outlined in Table II. Each device/ smart meter is programmed to have a 250 mW transmission power, 15 dBi channel gain and -101 dBm noise factor. Comparative analysis is carried out using existing legacy system benchmarks presented in [4] and [5]. The existing legacy models are implemented with similar parameters and on similar Monte-Carlo simulation settings.

The increase in the number of requests and devices impacts the network performance. Assume a scenario of crowded urban environment where people are almost continuously communicating through mobile devices. Thus, increasing the accessing intensity and RACH collision which result in packet drops and additional delays. These delays become critical in delay intolerant applications. Our system aggregates the requests automatically and reduces the access intensity without barring any device. Figure 4 shows that the proposed smart

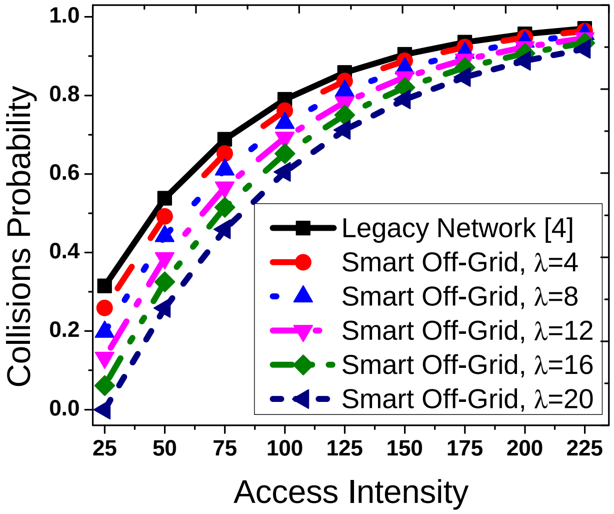


Fig. 4. Collision probability in the proposed and legacy system

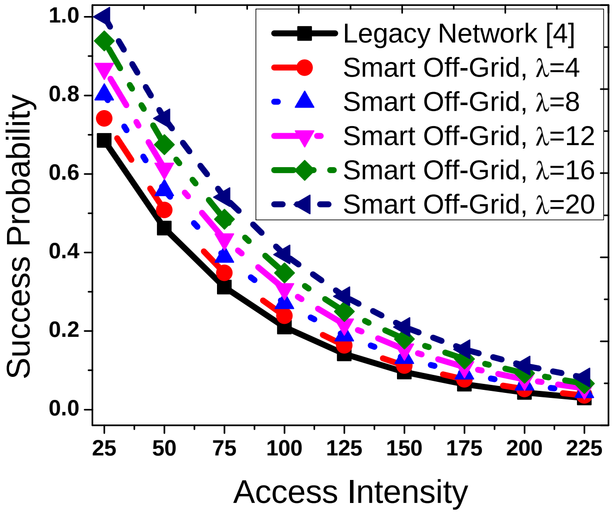


Fig. 5. Success probability in the proposed and legacy system

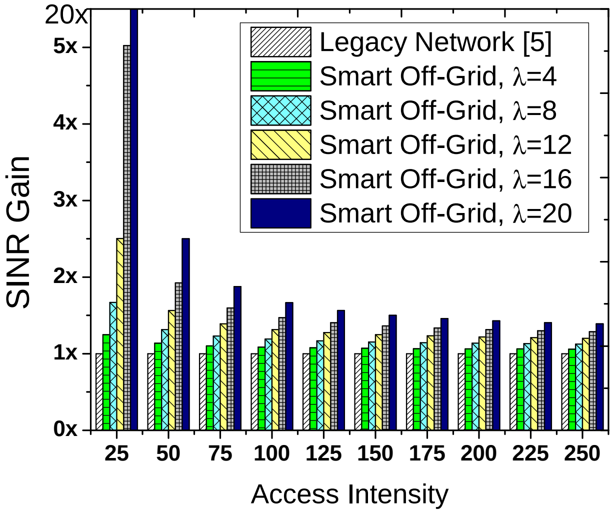


Fig. 6. SINR gain in the proposed and legacy system

off-grid system reduces collision probability by 5% to 25% for 4 to 20 aggregators. The existing collision probability is calculated using Equation 1 and proposed collision probability is estimated using Equation 7. The legacy network suffers from higher collisions due to the increase in the number of devices, whereas, the proposed smart off-grid reduces access intensity by aggregating requests. However, the increase in access intensity equally impacts both systems and increases collisions.

Figure 5 illustrates that the proposed system provides access to all devices at access intensity equals 25. The results for existing and proposed success probabilities are estimated using Equation 2 and 8. The successful resource allocation reduces in the proposed scheme with the increase in the number of devices. However, the proposed scheme outperforms the legacy system by 25% success probability. Figure 6 highlights the SINR gain of the proposed scheme over the existing network using Equation 3 and 9. The number of aggregators (λ) has a huge impact on the SINR value. The proposed scheme with 4 to 20 aggregators achieves approximately 1.3× to 20× gain over the legacy system. Figure 7 shows that the increase in devices equally impact on legacy network and proposed scheme with all variations. With a bandwidth of 20 MHz BW and only λ=4, the proposed scheme outperforms the legacy system by $\sim$ 2 Mbps. The increase in the value of λ increase the performances of the proposed scheme, i.e. with $\lambda = 20$, the data rate soars high as 43.97 Mbps. The channel quality indicator (CQI) of devices impact bandwidth distribution, ergo reducing data rate per device. Considering NB-IoT limited bandwidth of 180 KHz, we have experimented the data rate for 25 to 250 devices, competing against each other. NB-IoT specific evaluation in Figure 8 also present similar facts that the proposed scheme outperforms the legacy system by providing three times higher value.

A. Complexity Analysis

The time and order complexity of proposed aggregation algorithm in a best-case scenario includes that the device is aggregator and shares only one message. In response, receives neighboring messages and request for RACH. Thus, using Big-O notation, the total complexity of the best case scenario is $O(1) + O(N) + O(1)$. On the other hand, if the device is not an aggregator, it requests to all neighboring devices ($O(N)$) and receives all possible acknowledgment, but chooses the first response. Nevertheless, the total complexity becomes $O(N) + O(N) + O(1)$. In both cases, the number of devices plays a major role to increase complexity but this number also increase the chances of successful aggregation. Thus, the message exchange becomes acceptable for all those devices requiring a successful connection and communications.

V. Conclusion

This article presents and provides a complete architecture and communications paradigm to keep track of distributed RES equipment. Our major contributions include exhaustive literature review and innovative solution. We propose to continuously keep track of the performance and efficiency of the RES equipment using a smart meter. Our innovative smart meter not only tracks the energy generation but also monitors the

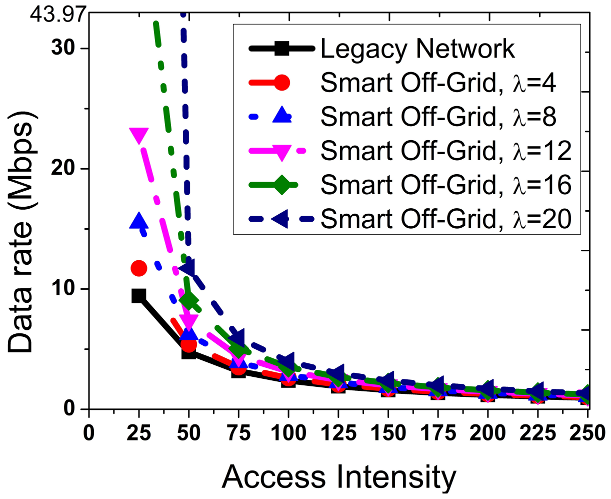


Fig. 7. Data rate of 4G LTE in the proposed and legacy system

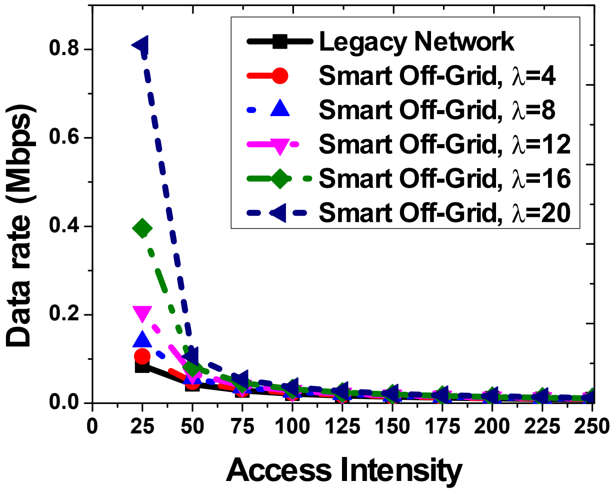


Fig. 8. Data rate of NB-IoT in the proposed and legacy system

health of the equipment. Moreover, the communication model for pervasive communication over NB-IoT is modeled and an aggregation scheme is presented. Our analytical model and respective proof of concept corroborate our claims of pervasive communications over 4G LTE and show clear advantages over the existing systems. Smart meter aggregation in NB-IoT and 4G environment achieve 25% reduced collisions and increased success probability. The communication also gains 1.3 to 20 times higher SINR value and three times higher data rate for a device (in LTE and NB-IoT, both), which leads to uninterrupted access. Our work and mathematically tractable equations pave a bridge to the co-existence of the NB-IoT in LTE and future 5G communications network. Moreover, in a device-centric architecture the trustworthy cooperation and aggregation of the devices is also a major future work.

References

[1] Zhiyong Li, Lin Chen, and Guofang Nan: ‘Small-Scale Renewable Energy Source Trading: A Contract Theory Approach’, *IEEE Transcation on Industrial Informatics*, vol. 14, no. 4, pp. 1491-1500, 2018

[2] S.K. Tiwari, Bhim Singh, and P.K. Goel, ‘Design and Control of Micro-Grid fed by Renewable Energy Generating Sources’, *IEEE Trans. on Industry Applications* , vol. 54, no. 3, pp. 2041-2050, 2018.

[3] Hui-Ju Hung, Ting-Yu Ho, Shi-Yong Lee, Chun-Yuan Yang, and De-Nian Yang, ‘Relay Selection for Heterogeneous Cellular Networks with Renewable Green Energy Sources’, *IEEE Trans. on Mobile Computing*, vol. 17, no. 30, pp. 1-14, 2018.

[4] K.S. Ko, M.J. Kim, K.Y. Bae, D.K. Sung, J.H. Kim, and J.Y. Ahn, ‘A novel random access for fixed-location machine-to-machine communications in OFDMA based systems’. *Communications Letters, IEEE*, vol. 16, no. 9, pp. 1428-1431, 2012.

[5] F. H. Kumbhar, N. Saxena, A. Roy, ‘Reliable Relay Autonomous Social D2D Paradigm for 5G LoS Communications’. *IEEE Communications Letters*, vol. 21, no. 7, pp. 1593-1596, 2017.

[6] N. Jiang, Y. Deng, M. Condoluci, W. Guo, A. Nallanathan and M. Dohler, ‘RACH Preamble Repetition in NB-IoT Network’. *IEEE Communications Letters*, vol. 22, no. 6, pp. 1244-1247, 2018.

[7] N. Saxena, M. Agiwal, H. Ahmad and A. Roy, ‘D2D-based Survival on Sharing (SoS) for Enhanced Disaster Time Connectivity’, *IEEE Technology and Society Magazine*, vol.37, no. 3, pp. 64-73, Sept. 2018.

[8] Rui Li, Wei Wang, and Mingchao Xia, “Cooperative Planning of Active Distribution System With Renewable Energy Sources and Energy Storage Systems”, *IEEE Access*, vol. 6, pp. 5916-5926, 2017.

[9] Shaghayegh Kazemlou, and Shahab Mehraeen, “Novel Decentralized Control of Power Systems With Penetration of Renewable Energy Sources in Small-Scale Power Systems”, *IEEE Transactions on Energy Conversion*, vol. 29, no. 4, pp. 851-861, 2014.

[10] Yuan Wu, Vincent K. N. Lau, Danny H. K. Tsang, Li Ping Qian, and Limin Meng, “Optimal Energy Scheduling for Residential Smart Grid with Centralized Renewable Energy Source”, *IEEE Systems Journal*, vol. 8, no. 2, pp. 562-576, 2014.

[11] Virginia Pilloni, Alessandro Floris, Alessio Meloni, and Luigi Atzori, “Smart Home Energy Management Including Renewable Sources: A QoE-driven Approach”, *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2006-2018, 2016.

[12] Mattia Marinelli, Fabrizio Sossan, Giuseppe Tommaso Costanzo, and Henrik W. Bindner, “Testing of a Predictive Control Strategy for Balancing Renewable Sources in a Microgrid”, *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1426-1433, 2014.

[13] C. Perera, C.H. Liu, S. Jayawardena, and M. Chen, “A survey on internet of things from industrial market perspective”, *IEEE Access*, vol. 2, pp. 1660-1679, 2014.

[14] L. M. Bello, P. D. Mitchell and D. Grace, ‘Intelligent RACH Access Techniques to Support M2M Traffic in Cellular Networks,’ in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8905-8918, Sept. 2018

[15] G.Farhadi, and A.Ito, “Group-based signaling and access control for cellular machine-to-machine communication”, *In IEEE Vehicular Technology Conference (VTC Fall), IEEE 78th, 2013* (pp. 1-6), 2013

[16] C.H. Chang, and H.Y. Hsieh, “Not every bit counts: A resource allocation problem for data gathering in machine-to-machine communications”. *In IEEE Global Communications Conference (GLOBECOM)*, pp. 5537-5543, 2012.

[17] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, “Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications”, *Wireless Communications, IEEE*, vol. 21, no. 3, pp. 12-18, 2014.

[18] H.Wu, C. Zhu, R.J. La, X.Liu, and Y. Zhang, FASA, “ Accelerated S-ALOHA using access history for event-driven M2M communications, *Networking, IEEE/ACM Transactions on*, vol. 21, no. 6, pp. 1904-1917, 2013

[19] S.Y Lien, T.H. Liau, C.Y. Kao, and K.C Chen, “Cooperative access class barring for machine-to-machine communications”. *Wireless Communications, IEEE Transactions on*, vol. 11, no. 1, pp. 27-32, 2012.

[20] T. M. Lin, C. H. Lee, J. P. Cheng and W. T. Chen, “PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks," *IEEE Transactions on Vehicular Technology,* vol. 63, no. 5, pp. 2467-2472, Jun 2014.

[21] M. Hasan, E. Hossain, and D. Niyato, “Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches”. *Communications Magazine, IEEE*, vol. 51, no. 6, pp. 86-93, 2013.

[22] 3GPP. Technical Specification 37.868, Study on RAN Improvements for Machine-type Communications. 2011, URL: www.qtc.jp/3GPP/Specs/37868-b00.pdf [Access Date: 5 July 2017]

[23] F. H. Kumbhar, A. Roy and N. Saxena, “RoBiN: Random Access using Border Routers in Cellular Netw.”, *Mobile Netw. and Applications*, vol. 21, no. 4, pp. 620-634, 2016.

The SMH Algorithm : An Heuristic for Structural Matrix Computation in the Partial Least Square Path Modeling

Odilon Yapo M. Achiepo[1]
Agropastoral Management School
UPGC University of Korhogo
BP 1328 Korhogo, Cote d'Ivoire

Edoete Patrice Mensah[2]
Dpt. of Maths. and Computer Science
INPHB Institut of Yamoussoukro
BP 1093 Yamoussoukro, Cote d'Ivoire

Edi Kouassi Hilaire[3]
Lab. of Maths. and Computer Science
UNA University of Abidjan
02 BP 801 Abidjan 02, Cote d'Ivoire

***Abstract*—The Structural equations modeling with latent's variables (SEMLV) are a class of statistical methods for modeling the relationships between unobservable concepts called latent variables. In this type of model, each latent variable is described by a number of observable variables called manifest variables. The most used version of this category of statistical methods is the partial least square path modeling (PLS Path Modeling). In PLS Path Modeling, the specification of the relashonships between the unobservable concepts, knows as structural relationships, is the most important thing to know for practical purposes. In general, this specification is obtained manually using a lower triangular binary matrix. To obtain this lower triangular matrix, the modeler must put the latent variables in a very precise order, otherwise the matrix obtained will not be triangular inferior. Indeed, the construction of such a matrix only reflects the links of cause and effect between the latent variables. Thus, with each ordering of the latent variables corresponds a precise matrix.The real problem is that, the more the number of studied concepts increases, the more the search for a good order in which it is necessary to put the latent variables to obtain a lower triangular matrix becomes more and more tedious. For five concepts, the modeler must test $5! = 120$ possibilities. However, in practice, it is easy to study more than ten variables, so that the manual search for an adequate order to obtain a lower triangular matrix extremely difficult work for the modeler. In this article, we propose an heuristic way to make possible an automatic computation of the structural matrix in order to avoid the usual manual specifications and related subsequent errors.**

Keywords*—Structural equations modeling; PLS algorithm; latents variables; structural matrix; R programming language

I. Introduction: PLS Path Modeling in R

The PLS Path Modeling in a structural equation modelling with latent variables (SEMLV), is a method in which the partial least square (PLS) algorithm is used to estimate the model ([1], [2], [3]). Generally, the structural equation models (SEM) are describe graphically by specifying the latent variables (inobservable). For each latent variable, the manifest variables (observable) that are related to it are also specified. Latent variables represent concepts such as loyalty, quality, poverty, abilities, etc. The manifest variables are indicators that describe these latent variables and they are collected in a dataset. An example of such model, called European Customer Satisfaction Index (ECSI) Model, that can be found in [4], is giving in the Figure 1 below:

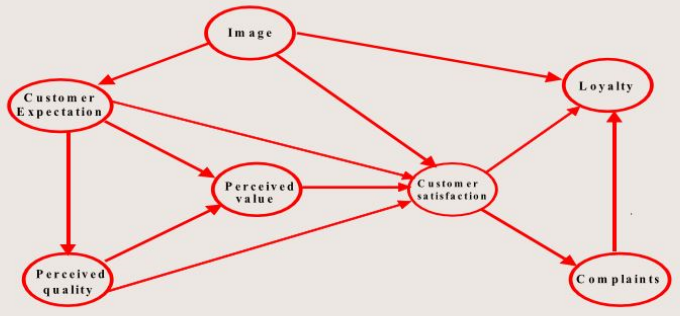


Fig. 1. The European Customer Satisfaction Index Model.

The Figure 1 shows an example of the structural relationship between latent variables. It is known as the European Customer Satisfaction Index model (ECSI model) and is often used in marketing studies. This article focuses on the specification of this kind of relations in practice. When we use a computer to estimate the model, the graph is often specified as binary low triangular matrix. The operation may be time-consuming because one has to find the best order of the latent variables in a table in order to get the lower triangular matrix. The goal of this paper is to give a method which automatically get the right order and automatically compute the structural relationship matrix.

II. Conceptualization : Main idea behind the SMH Algorithm

A square (lower triangular) boolean matrix representing the inner model (i.e. the path relationships between latent variables) is a matrix of zeros and ones that indicates the structural relationshipsbetween latent variables. This path matrix must be a lower triangular matrix that has a 1 when column j affects row i, and a 0 otherwise.

The latent variables can be classified in three categories according to their roles in the structural equations in which they appear. The SMH is based on the following classification:

- **The exogenous variables :** It's the latent variables which have no other latent variables related to them.
- **The endogenous variables :** It's the latent variables which are not related to any other latent variables.

- **The neutral variables :** It's the latent variables that are related to others in both directions.

The main idea of this heuristic is to classify all the latent variables within these differents groups (Exogenous, Endogenous, Neutral) and find a way to order them to obtain a lower triangular matrix. To find the rigth order of the latents variables, we can remark than the exogenous latents variables must be ordered first (left side), then the neutral latents variables must follow them (middle), and finally, the endogenous latent variables must be the last ones to use (right side). This groups order is found by analyzing some simple cases.

For a formal purpose, let consider the following mathmatical notations :

- N the numbers of latent variables
- ξ_j the j^{th} latent variable
- Θ_j the endogenous statut of the latent variable ξ_j
- Γ_j the exogenous statut of the latent variable ξ_j
- E_j the number of latents variables that the variable ξ_j is related to
- F_j the number of latents variables that are related to ξ_j
- K_j the numbers of exogenous, latent variables that are related to the variable ξ_j
- μ_j the order score of the latent variable ξ_j

The variables Θ_j and Γ_j can be express using the kroneker notation :

$$\Theta_j = \begin{cases} 1 & \text{if } \xi_j \text{ is endogenous} \\ 0 & \text{if } \xi_j \text{ is not endogenus} \end{cases} \tag{1}$$

$$\Gamma_j = \begin{cases} 1 & \text{if } \xi_j \text{ is exogenous} \\ 0 & \text{if } \xi_j \text{ is not exogenous} \end{cases} \tag{2}$$

This conceptualisation, will be use to find an ordered metric for each variable. The variables will be orderd according to the value of this metric. Hight the metric's value of a variable is, hight will be it rank.

III. Computing : The order metric of the SMH Algorithm

The heuristic method is based on three general empirical principles where it foundation can be seen.

A. About the Exogenous Variables

The exogenous latent variables are the only ones with $\Gamma_j = 1$ and they must have the lowest values μ_j to be in the first position in the structural matrix. Different exogenous latent variables are distinguished according to the number of latent variables F_j they are related to. The higher F_j is, the lower the score μ_j has to be. Some variables that an exogenous latent variable is related to can be endogenous. Therefore, exogenous variables are to be characterized by the number of endogenous variables they belongs to (K_j) they are related to. The higher K_j is, the higher the score μ_j has to be. To take into account these realities, the order score of the endogenous latent variables is taken to be $-10^4 F_j + K_j$. In this case, the minimum score is obtained when all latent variables are exogenous except for one which is exogeneous $(F_j = N-1, K_j = 1)$ and the maximum score is obtained when all the latent variables are endogenous except for one which is endogenous $(F_j = 1, K_j = N-1)$. The scores of the exogenous latent variables are in the interval $[-10^4(N-1)+1, -10^4+N-1]$.

B. About the Endogenous Variables

The endogenous latent variables are the only ones with $\Theta_j = 1$ and they must have the highest values of μ_j to be in the last position in structural matrix. Different endogenous latent variables are distinguished according to the number of latent variables (E_j) related to them. The higher E_j is, the higher the score μ_j must be. To take into account this reality, the order score of the endogenous latent variables is taken to be $10^4 E_j$. In this case, the maximum score is obtained when all latent variables are endogenous except for one $(E_j = N-1)$ which is exogenous and the minimum score is obtained when all the latent variables are exogenous except for one $(E_j = 1)$ which is endogenous. The scores of the exogenous latent variables are in the interval $[10^4, -10^4(N-1)]$.

C. About the Neutral Latent Variables

The neutral latent variables are the ones with the $\Theta_j + \Gamma_j = 0$. They must have the values of μ_j which are higher than the highest exogenous variable value and less than the lowest endogenous variable value in order to be between exogenous and endogenous latent variables in the structural matrix. Different neutral latent variables are distinguished according to the number of latent variables(F_j) they are related to. The higher F_j is, the lower the score μ_j must be. Some variables that a neutral latent variable are related to can be endogenous. Therefore, exogenous variables are to be characterized by the number of neutral variables(K_j) they are related to. The higher K_j is, the higher the score μ_j have to be. Neutral variables are also distinguished according to the number of latent variables(E_j) they are related to. The higher E_j is, the higher the score μ_j have to be. To take into account all these realities, the order score of the endogenous latent variables is taken to be $10^{3/2}E_j - 10F_j + K_j$. In this case, the maximum score is obtained when all latent variables are endogenous except for one $(E_j = N-1, F_j = 1, K_j = 1)$ which is exogenous and the minimum score is obtained when all latent variables are exogenous except for $(E_j = 1, F_j = N-1, K_j = N-1)$ which is endogenous. The scores of the exogenous latent variables are in the interval $[10^{3/2}(N-1) - 9, 10^{3/2} - 9(N-1)]$.

D. Order Score Computation

To compute the structural matrix, the latent variables must be ordered properly. The correct order give a lower triangular matrix. As it has been said before the main objective of the heuristic is to find the best set of ordered variables to compute the correct structural matrix. This order is based on the score that can be defined by

$$\mu_j = \begin{cases} 10^4 E_j & \text{if } \xi_j \text{ is endogenous} \\ 10^{3/2} E_j - 10F_j + K_j & \text{if } \xi_j \text{ is neutral} \\ -10^4 F_j + K_j & \text{if } \xi_j \text{ is exogenous} \end{cases} \quad (3)$$

Mathematically, these descriptions can be summarize in the single function defined as :

$$\mu_j = 10^4 E_j \Theta_j + (10^{3/2} E_j - 10F_j + K_j) \\ *(1 - \Theta_j - \Gamma_j) - (10^4 F_j - K_j)\Gamma_j \quad (4)$$

The latent variables are then ordered based on their μ scores. For two latent variables ξ_i and ξ_j , the position of ξ_i in the structural matrix is before ξ_j if $\mu_i \leqslant \mu_j$.

The problem solved by our method is a similar problem to that of the well-known traveling salesman problem in operations research ([5], [6]). However, the metaheuristics used in operational research, such as tabu search, simulated annealing, genetic algorithms, etc. have the disadvantage of requiring significant resources in terms of calculation. In addition, the implementation of these algorithms is very complex and require a good mastery of their operating principles. Compared to these methods, the approach developed in this article is very easy to use. The method is limited to a simple classification of latent variables and manifests variables, to their enumeration and to the application of a simple arithmetic formula to obtain scores for ordering latent variables. The computation time is more than one hundred lower than that of conventional optimization metaheuristics. Our approach is therefore an optimization metaheuristic that applies to a very particular problem, namely, the search for a structural matrix in the PLS Path Modeling.This heuristic is the core method of used in the R package *plspm.formula* ([7]) we have already developed and which is available for free download on the mirror sites of the R software. The following Figure 2 shows the performance of the heuristic when the number of latent variables is growing :

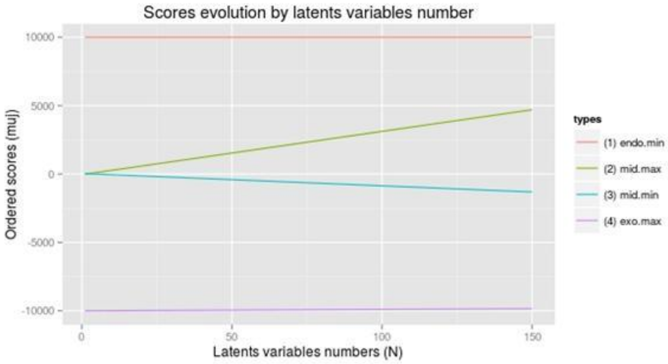


Fig. 2. The exogeneous minimum and maximum

According to the figure, the heuristic is able to give correct response with more than 100 latent variables.Based on this result, we can state that the heuristic method is very robust since the reasonable numbers of latent variables one can use in practice is generally less than twenty.

IV. Programming : The SMH algorithm code in R

A. The plspm.shm R Function

This section present the implementation of the SMH algorithm in R language [8]. The fonction is based on the R Package *plsmp* ([9]) basis of this scientific computing language can be found in . The SMH algorithm in R is as follows:

```
require(plspm)
plspm.shm <- function(latents,latlist,
                      mat=TRUE,iplot=TRUE)
{
   N <- length(latents)
   vldroite <- unique(unlist(latlist[[2]]))
   vlgauche <- latlist[[1]]
   calc.nfois.exo <- function(vlat) {
      nbfois.exo <- function(vtot) {
         return(sum(vlat %in% vtot))
      }
      return(sum(sapply(latlist[[2]],
             nbfois.exo)))
  }
  calc.nfois.exo <- Vectorize(calc.nfois.exo)
  vlexo <- latents[1-as.numeric(
                   latents %in% vldroite)]
  calc.equ.exo <- function(vlat){
     indx <- which(latlist[[1]] == vlat)
     if(length(indx) < 1){res <- 0}
     else {
        res <- sum(as.numeric(
        vlexo %in% latlist[[2]][[indx]]))
     }
     return(res)
  }
  calc.equ.exo <- Vectorize(calc.equ.exo)
  ntotF <- sapply(latlist[[2]], length)
  calc.nbequ <- function(vlat){
     indx <- which(latlist[[1]] == vlat)
     if (length(indx) < 1) {res <- 0}
     else {res <- ntotF[indx]}
     return(res)
  }
  calc.nbequ <- Vectorize(calc.nbequ)
  Thetaj <- 1-as.numeric(latents %in% vldroite)
  Ej <- as.vector(calc.nbequ(latents))
  Gammaj <- 1-as.numeric(latents %in% vlgauche)
  Fj <- as.vector(calc.nfois.exo(latents))
  Kj <- as.vector(calc.equ.exo(latents))
  muj <- 10^4*Ej*Thetaj+(10^(3/2))*Ej-10*Fj+Kj)
         *(1-Thetaj-Gammaj)-(10^4*Fj-Kj)*Gammaj
  olatents <- latents[order(muj)]
  reslist <- list(mu = muj, ordre = olatents)
  if(mat){
    matlist <- function(vect){
       return(as.numeric(olatents %in% vect))
    }
    Mlist <- lapply(latlist[[2]], matlist)
    mat.vect <- function(j){
        indj <- which(
                 latlist[[1]] == olatents[j])
        if (length(indj) < 1){
           return(rep(0, N))
        }
        else {return(unlist(Mlist[indj]))}
   }
   mat.vect <- Vectorize(mat.vect)
   Mat <- t(mat.vect(1:N))
   rownames(Mat) <- olatents
   reslist <- c(reslist, list(matrice = Mat))
 }
 if (iplot) {innerplot(Mat)}
 return(reslist)
}
```

B. The Parameters and Results of the plspm.shm Function

The algorithm take essentially two inputs:

latent : a character vector containing the latent variable names

latlist : a list to specify which latents variables explain another

The parameter latlist is a R list structure and must contain two R objects:

1) a vector of the endogenous latent variables.
2) a list of vector objects for each endogenous variable. For an endogenous variable, the vector contains exogenous latent variables which are related to it. The order of vector objects in the internal list must correspond to the one of the endogenous variable.

The main output of the plspm.shm function is an ordered vector of all the latent variables. This order is the one one can use to have a structural matrix in the form of lower trianguler binary matrix needed to estimate PLS Path Model, for example the *plspm()* function in the *plspm* R package (*plspm*). But, the functions have the logical parameter mat that permits to compute the corresponding inner matrix *(mat=TRUE)* or not *(mat=FALSE)*. This prevents from using a manual ordered latent variables vector to find the matrix. By default, the function compute that matrix. The function also have an other logical parameter name igraph that specifies if the relationship graph must be compute *(igraph=TRUE)* or not *(igraph=FALSE)*.

V. Illustration: Test of the SMH Algorithm in R

A. Applications on a Relative Simple Problem

To show the simple usage of SMH algorithm, we generate four (4) latent variables ”A1”, ”A2”, ”A3” and ”A4”. Weassume that the relations between these latent variables can be described by two rules :

- First: ”A1” and ”A4” have an impact on ”A2”
- Secondly: ”A3” have an impact on ”A1” and ”A2”.

The implementation in R is giving by the code below :

```
R> lvect <- paste("A",1:4,sep="")

R> lvlist <- list(
paste("A",1:3,sep=""),
list("A3", c("A1","A3","A4"),"A4")
)

R> res <- plspm.shm(lvect,lvlist,
                   mat=TRUE,iplot=TRUE)
```

The different results obtained in R concerning the latent variables vector, the latent variables list and the structural matrix are :

```
R> print(lvect)
[1] "A1" "A2" "A3" "A4"

R > print(lvlist)
[[1]]
[1] "A1" "A2" "A3"
[[2]]
[[2]][[1]]
[1] "A3"
[[2]][[2]]
[1] "A1" "A3" "A4"
[[2]][[3]]
[1] "A4"

R> print(round(res,2))
$mu
[1]  21.63 30000.00 11.62 -20000.00
$ordre
[1] "A4" "A3" "A1" "A2"
$matrice
    [,1] [,2] [,3] [,4]
A4    0    0    0    0
A3    1    0    0    0
A1    0    1    0    0
A2    1    1    1    0
```

We can then see that the algorithm is capable of finding the correct order of the latent variables and capable of giving the correct structural matrix (triangular inferior). The graph Figure 3 given by the algorithm is :

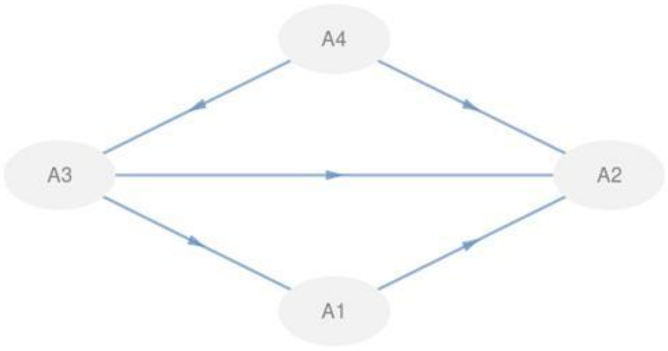


Fig. 3. The inner graph of the simple example

This graph is the graphical version of the structural matrix and it use makes easier the understanding of the structural relationships. Notice that in this example, we have four latent variables. The next example will use six latent variables and is concerned with a real example of the ECSI model as presented in the plspm package on the satisfaction dataset.

B. Application on a More Complex Problem

In this second example, the latent variables are denoted : image (”IMAG”), expectations (”EXPE”), quality (”QUAL”), value (”VAL”), satisfaction (”SAT”) and loyalty (”LOY”). The relations between the latent variables are more complex and can be described by the following five rules :

- Image have an influence on expectations, satisfaction and loyalty
- Expectation have an influence on quality, value and satisfaction
- Quality have an influence on value and satisfaction
- Value have influence on satisfaction
- Satisfaction have influence on loyalty.

The implementation of these different rules and the application of the RSH algorithm in R are given by the following code :

```
R> satvect <- c("IMAG", "EXPE", "QUAL",
                "VAL", "SAT", "LOY")

R> satlist <- list(
               c("EXPE","QUAL", "VAL","SAT", "LOY"
               list(
                 c("IMAG"),
                 c("EXPE"),
                 c("EXPE","QUAL"),
                 c("IMAG", "EXPE", "QUAL", "VAL"),
                 c("IMAG", "SAT"))
               )

R> satres <- plspm.shm(satvect,satlist,
                       mat=TRUE,iplot=TRUE)
```

The different results obtained in R and concerning the latent variables vector, the latent variables list and the structural matrix are :

```
R> print(satvect)
[1] "IMAG" "EXPE" "QUAL" "VAL"  "SAT"  "LOY"

R> print(satlist)
[[1]]
[1] "EXPE" "QUAL" "VAL"  "SAT"  "LOY"
[[2]]
[[2]][[1]]
[1] "IMAG"
[[2]][[2]]
[1] "EXPE"
[[2]][[3]]
[1] "EXPE" "QUAL"
[[2]][[4]]
[1] "IMAG" "EXPE" "QUAL" "VAL"
[[2]][[5]]
[1] "IMAG" "SAT"
```

```
R> print(satres)
$mu
[1] -30000.0  2.6  11.6  53.2  117.5  20000.0
$ordre
[1] "IMAG" "EXPE" "QUAL" "VAL"  "SAT"  "LOY"
$matrice
```

```
       [,1] [,2] [,3] [,4] [,5] [,6]
IMAG     0    0    0    0    0    0
EXPE     1    0    0    0    0    0
QUAL     0    1    0    0    0    0
VAL      0    1    1    0    0    0
SAT      1    1    1    1    0    0
LOY      1    0    0    0    1    0
```

We can again see that the algorithm is capable of finding the correct order of the latent variables and capable of giving the correct structural matrix (triangular inferior). The graph Figure 4 given by the algorithm is :

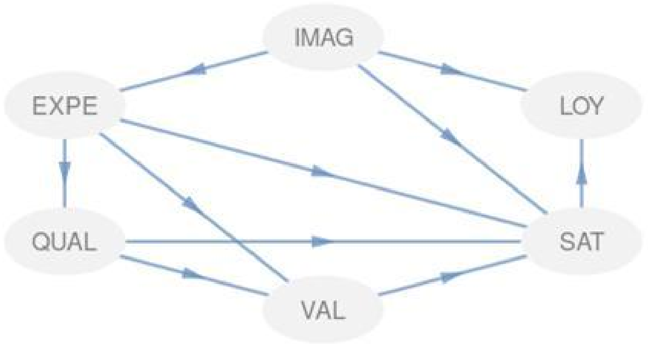


Fig. 4. The inner graph of the complex example

This graph is the graphical version of the structural matrix. It confirms the fact that the heuristic is able to handle problems with large variables.

VI. Conclusion

In the field of PLS Path modeling, the task of specifying structural matrices has always been tedious because of its purely manual nature. The method proposed in this article freed the modeler of this constraint by providing a means of automatic search of the correct order in which the latent variables must be placed in order to obtain a lower triangular matrix. The algorithm even calculates this matrix directly, which saves time and avoids errors related to the manual specification of such matrices. The heuristic described in this paper makes easier the process of finding automatically the PLS Path Modeling specifications. The simulations carried out show that, theoretically, this heuristic can easily be used for models involving more than one hundred latent variables. This possibility increases the scope of the PLS Path Modeling that was, until now, used on a limited number of latent variables because of the difficulties related to the manual specification of the structural relationships. However, one must take care of the fact that the structural relation rules are not circular because the matrix, in this case, is not triangular and that the problem can be misspecified in practice. The SMH heuristic also avoids the need of exploring all of the possible ordered latent variables configurations. It is an elegant solution to this combinatory problem. The use of this heuristic avoids the test of all arrangements of latent variables in order to find the best which gives the correct structural matrix. Future work will focus on the generalization of the principle of our method on the traveling salesman problem. Such a generalization will allow the algorithm to apply a much larger set of problems.

In this case, the study of the algorithmic complexity of the method and its comparison with the existing heuristics will make it possible to better understand its advantages over the optimization metaheuristics known to deal with the traveling salesman problem.

References

[1] Avkiran, N. K., Ringle, C. M., Low, R. K. Y. (2018); *Monitoring Transmission of Systemic Risk: Application of Partial Least Squares Structural Equation Modeling in Financial Stress Testing*. Journal of Risk, forthcoming, 2018.

[2] Ahrholdt, D. C., Gudergan, S. P., Ringle, C. M.; *Enhancing Service Loyalty: The Roles of Delight, Satisfaction, and Service Quality*. Journal of Travel Research, Volume 56, Issue 4, pp. 436-450, 2017.

[3] Mikko Ronkk, Cameron N. McIntosh, John Antonakis, Jeffrey R. Edwards (2016); *Partial least squares path modeling: Time for some serious second thoughts*. Journal of Operations Management, Elsivier, 2016.

[4] D. Christian (2009); *Free Model for Generalized Path Modeling and Comparison with Bayesian Network*, EDF Research and Development, 2009.

[5] Ahmad Fouad El-Samak, Wesam Ashour (2015); *Optimization of Traveling Salesman Problem Using Affinity Propagation Clustering and Genetic Algorith*. JAISCR, Vol. 4, No. 4, pp. 239-245, 2015.

[6] Johann Dréo, Alain Pétrowski, Patrick Siarry, Eric Taillard (2003); *Métaheuristiques pour l'optimisation difficile*. Eyrolle, 2003.

[7] Odilon Yapo M.,Achiepo (2015); *plspm.formula: Formula Based PLS Path Modeling in R*, R package version 1.0., 2015.

[8] R Core Team (2015); *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2015.

[9] G. Sanchez and L., Trinchera and G. Russolillo (2015);*plspm: Tools for Partial Least Squares Path Modeling*,R package version 0.4.7, 2015.

scaleBF: A High Scalable Membership Filter using 3D Bloom Filter

Ripon Patgiri[1], Sanbuzima Nayak[2], Samir Kumar Borgohain[3]
Dept of Computer Science & Engineering
National Institute of Technology Silchar
Assam-788010, India

***Abstract*—Bloom Filter is extensively deployed data structure in various applications and research domain since its inception. Bloom Filter is able to reduce the space consumption in an order of magnitude. Thus, Bloom Filter is used to keep information of a very large scale data. There are numerous variants of Bloom Filters available, however, scalability is a serious dilemma of Bloom Filter for years. To solve this dilemma, there are also diverse variants of Bloom Filter. However, the time complexity and space complexity become the key issue again. In this paper, we present a novel Bloom Filter to address the scalability issue without compromising the performance, called scaleBF. scaleBF deploys many 3D Bloom Filter to filter the set of items. In this paper, we theoretically compare the contemporary Bloom Filter for scalability and scaleBF outperforms in terms of time complexity.**

***Keywords*—Bloom filter; membership filter; scalable bloom filter, duplicate key filter; hashing; data structure, membership query.**

I. Introduction

Burton Howard Bloom introduces a data structure on approximate membership query in 1970 [1], hence, it is named as Bloom Filter. Bloom Filter is an extensively experimented to enhance a system's performance since its inception. Moreover, Bloom Filter is also applied numerous areas, namely, Big Data, Cloud Computing, Networking, Security [2], Database, IoT, Bioinformatics, Biometrics, and Distributed system. However, Bloom Filter is inapplicable in hard real-time system, and password management system [3] due to accuracy issues. Applications of Bloom Filter take the lion's share in Computer Networking which includes Named Data Networking (NDN), Content-Centric Networking (CCN), Software-defined Network (SDN), Domain Name System (DNS), and Computer Security. The Bloom Filter reduces space consumption in an order of magnitude as compared to a conventional hash algorithm. However, Bloom Filter cannot stand itself. Bloom Filter is used as enhancer of a system. For example, BigTable uses Bloom Filter to reduce the number of disk accesses which improves the performance drastically [4]. Similarly, in Cassandra [5].

A. Motivation

Several variants of Bloom Filters have been developed to address some issues [6]. However, most of the Bloom Filters are developed to address scalability issue. Guanlin Lu et al. [7] proposes a Forest-structured Bloom Filter (FBF). The FBF is a combination of RAM and flash memory. Similarly, Debnath et al. [8] develops a very high scalable Bloom Filter based on RAM and flash memory. BloomStore is also another highly scalable Bloom Filter [9]. However, these solutions are hierarchical, and thus, lookup and insertion cost is very high. It takes $O(logn)$ time complexity in insertion and lookup operations as demonstrated in Table I.

B. Contribution

To address scalability issues, we propose a novel scalable Bloom Filter, called scaleBF. scaleBF is a very simple data structure yet powerful. scaleBF increases its scalability without compromising the performance. scaleBF takes $O(1)$ time complexity in lookup and insertion operations, which is compared in Table I. scaleBF uses chaining hash data structure for implementing the scalability. Also, scaleBF deploys 3DBF [12] to inherit the performance and low memory consumption.

Table I depicts the most scalable Bloom Filters. BloomFlash [8], and FBF [7] uses hierarchical structures to indexed the Bloom Filters. BloomStore [9] uses linear chain data structure (not open hashing data structure) to store the Bloom Filters in Flash memory. Moreover, BloomStore is designed to perform parallel lookup operation. On the contrary, scaleBF uses chaining hash data structures to achieve higher scalability without compromising the performances. TB^{2}F [10] deploys tree-bitmaps and Bloom Filter, and used for name lookup in Content-Centric Network (CCN). The input is split into a T-segment of fixed size and a B-Segment of variable size. The T-segment key is inserted into bitmap-trie, and the B-segment is inserted into Bloom Filter. However, maintaining trie data structure is costly in terms of space as well as time. On the other hand, Bloofi [11] uses tree structured Bloom Filter which cuases costly in insertion and lookup. The scalability of BloomFlash [8], FBF [7], BloomStore [9], scaleBF is higher than TB^{2}F and Bloofi [11].

C. Organization

The article is organized as follows- Section II presents the proposed system, called scaleBF. The architecture of scaleBF is demonstrated in Section II. Section III presents a theoretical analysis on scaleBF. Also, every aspect of scaleBF is analyzed in Section III. Article discusses cons of scaleBF in Section IV. Finally, the article is concluded in Section V.

II. scaleBF: The Proposed System

A. 3D Bloom Filter (3DBF)

The 3-Dimensional Bloom Filter (3DBF) is similar to conventional Bloom Filter except array structure [12]. The

TABLE I. COMPARISON OF VARIOUS SCALABLE BLOOM FILTER

Name	Types	Insertion	Lookup	Scalability	Platform	Algorithm
BloomFlash [8]	Hierarchical	Logarithmic	Logarithmic	High	RAM & Flash	Serial
FBF [7]	Hierarchical	Logarithmic	Logarithmic	High	RAM & Flash	Serial
BloomStore [9]	Linear Chaining	Constant	Constant	High	RAM & Flash	Parallel
TB^{2}F [10]	Hierarchical	Logarithmic	Logarithmic	Medium	RAM	Parallel
Bloofi [11]	Hierarchical	Logarithmic	Logarithmic	Medium	RAM	Serial
scaleBF	3D	Constant	Constant	High	RAM	Serial

3DBF uses 3D arrays and it is a static Bloom Filter in nature. The static Bloom Filter does not change the size at run time. Also, static Bloom Filter does not readjust with ever growing data. However, a new 3DBF is created to address the scalability issue.

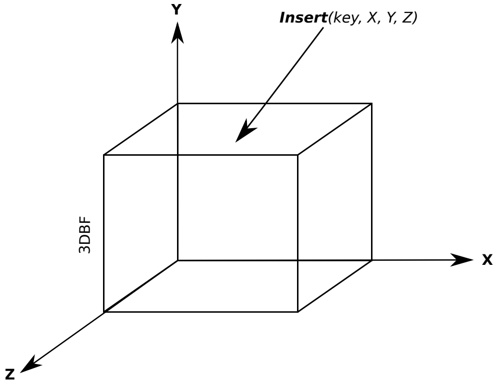


Fig. 1. 3DBF architecture

Figure 1 depicts the architecture of 3DBF. The 3D Bloom Filter uses four modulus operator using prime numbers instead of hashing a key into k different places. These modulus reduces the false positive probability. Thus, 3DBF is independent from number of hash functions k. Let, $\mathbb{B}_{X,Y,Z}$ bet the 3DBF where X, Y and Z be the dimension of the filter. The dimensions are prime numbers, otherwise, false positive increases. Let, $\mathbb{B}_{i,j,k}$ be a cell of the 3DBF. The cell stores **long int** which occupies $64 - bits$. Let us insert a key κ. The 3DBF uses Murmur hashing [13] to generate a hash-value of input item κ. Let, h be the generated hash-value by Murmur hashing. Now, $i = h\%X,\ j = h\%Y,\ k = h\%Z$, and $\rho = h\%63$, where ρ is the bit position of the cell $\mathbb{B}_{i,j,k}$. 3DBF sets a bit using Equation (1)-

$$\mathbb{B}_{i,j,k} \leftarrow \mathbb{B}_{i,j,k}\ OR\ (1 << \rho) \tag{1}$$

where OR is bitwise OR operator. Equation (1) is invoked to insert an input item into 3DBF. The lookup operation requires similar calculation. Equation (2) is invoked to perform the lookup operation in a 3DBF.

$$Flag \leftarrow (\mathbb{B}_{i,j,k} \oplus (1 << \rho))AND(1 << \rho) \tag{2}$$

If $Flag$ is assigned by '1', then 3DBF returns true, otherwise, it returns false. Each item requires a single bit in 3DBF as disclosed in Equation (1), and each cell has $63 - bits$. Therefore, 3DBF consumes the lowest memory as compared to other variants of Bloom Filter. Moreover, 3DBF features detection of the fullness of the filter. 3DBF defines the criticality factor to consider whether the filter is full or not [12].

B. Insertion operation in scaleBF

scaleBF deploys chaining mechanism of conventional hashing data structure for highly scalable. scaleBF deploys many 3DBFs.

1) Insertion of Bloom Filter: A Bloom Filter is formed by three 3DBFs. Each Bloom Filter is formed by three 3DBF. However, Bloom Filter can be formed by augmenting more 3DBF, but we have chosen three for simpler illustration. Each key is inserted into three 3DBF. Let, η be the chain size, and input item κ to be inserted. There are η chains in scaleBF. A new Bloom Filter (three 3DBF) is inserted into the chain if the Bloom Filter (three 3DBF) in particular chain is full.

2) Insertion of a Key: Insertion of the key is performed using Equation (1) and hashes the key into the particular chain. If a Bloom Filter (three 3DBF) size is full, then move to the last Bloom Filter (three 3DBF). Insert the key using Equation (1). A key is hashed into particular slot of the chain. There are many Bloom Filters in the particular slot linked with each other as shown in Figure 2. If first three Bloom Filter is full, then create and link three 3DBF as demonstrated in the figure.

C. Lookup operation in scaleBF

Figure 3 depicts the lookup operations of the scaleBF. A key is hashed into particular chain and lookup all Bloom Filters sequentially. As a comparison, three 3DBF is searched. If the first three Bloom Filter returns true, then the key is member of Bloom Filter. Otherwise, move forward to the next three 3DBF, and so on.

III. ANALYSIS

There is no significant difference between 3DBF and conventional Bloom Filter analysis of number bits consumed, except $k = 1$ in 3DBF. Therefore, scaleBF is analyzed through the conventional Bloom Filter. Let, m be the size of Bloom Filter, n be the number of entries, and $k = 1$ be the number of hash function, then the probability of a bit to be '0' is

$$\left(1 - \frac{1}{m}\right)^n$$

Therefore, probability of total bit to be '1' is

$$\left(1 - \left(1 - \frac{1}{m}\right)^n\right)$$

Since, scaleBF uses 3DBF, thus, $m = X \times Y \times Z \times 63$. F. Grandi [14] present a new way to calculate the false positive probability using $\delta - transformation$. Let, X be the random

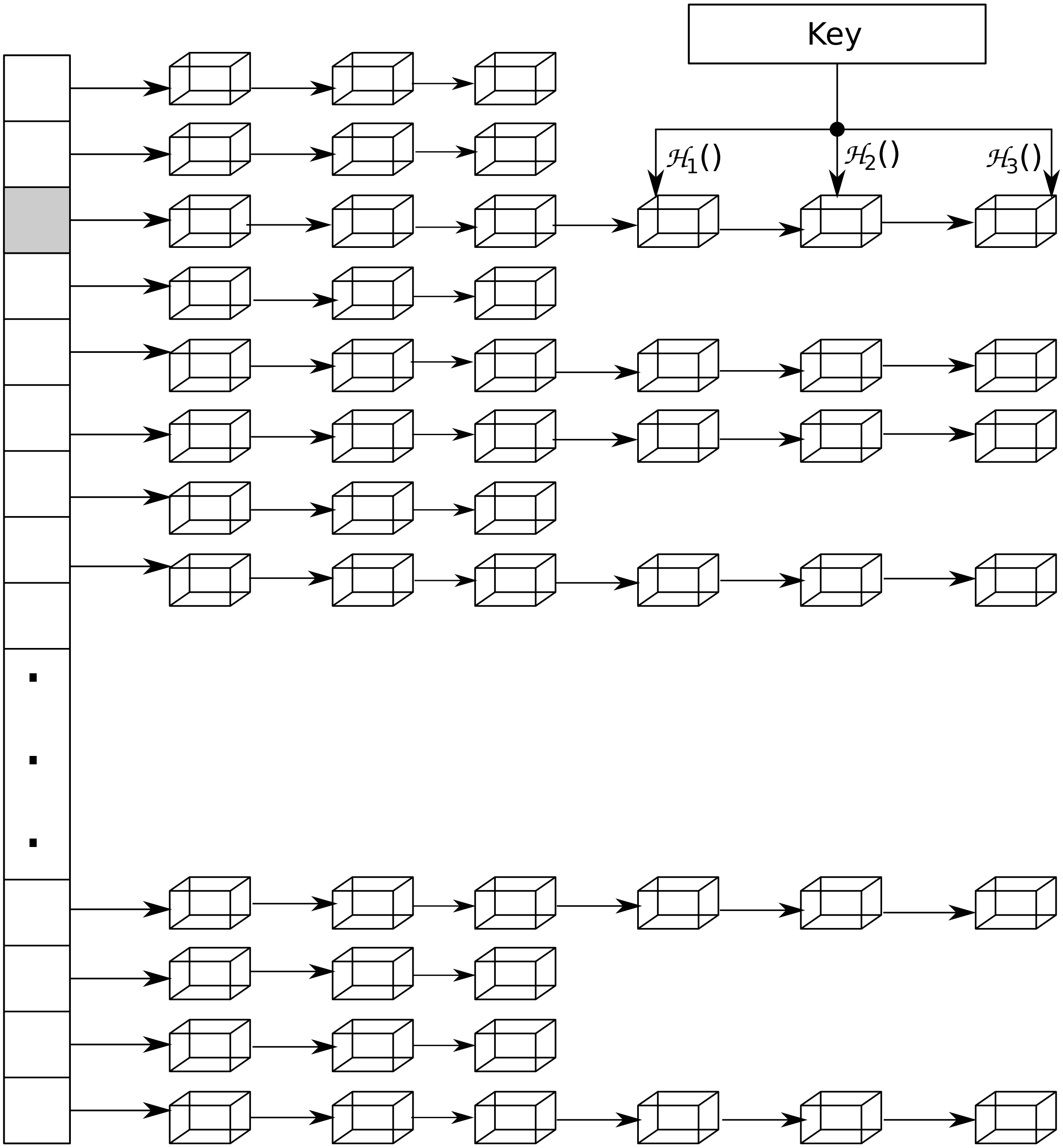


Fig. 2. Insertion of an input item and increment of filter size using conventional chaining in scaleBF.

variable to represent the total number of '1' in the Bloom Filter, then

$$E[X] = m\left(1 - \left(1 - \frac{1}{m}\right)^n\right)$$

The probability of false positive is conditioned to a number by $X = x$, then

$$Pr(FP|X = x) = \left(\frac{x}{m}\right)$$

Therefore, false positive probability is

$$FPP = \sum_{x=0}^{m} Pr(FP|X = x)Pr(X = x)$$

$$FPP = \sum_{x=0}^{m} \left(\frac{x}{m}\right) f(x)$$

where $f(x)$ is probability mass function of X. F. Grandi [14] applies $\delta - transformation$ to calculate $f(x)$ and

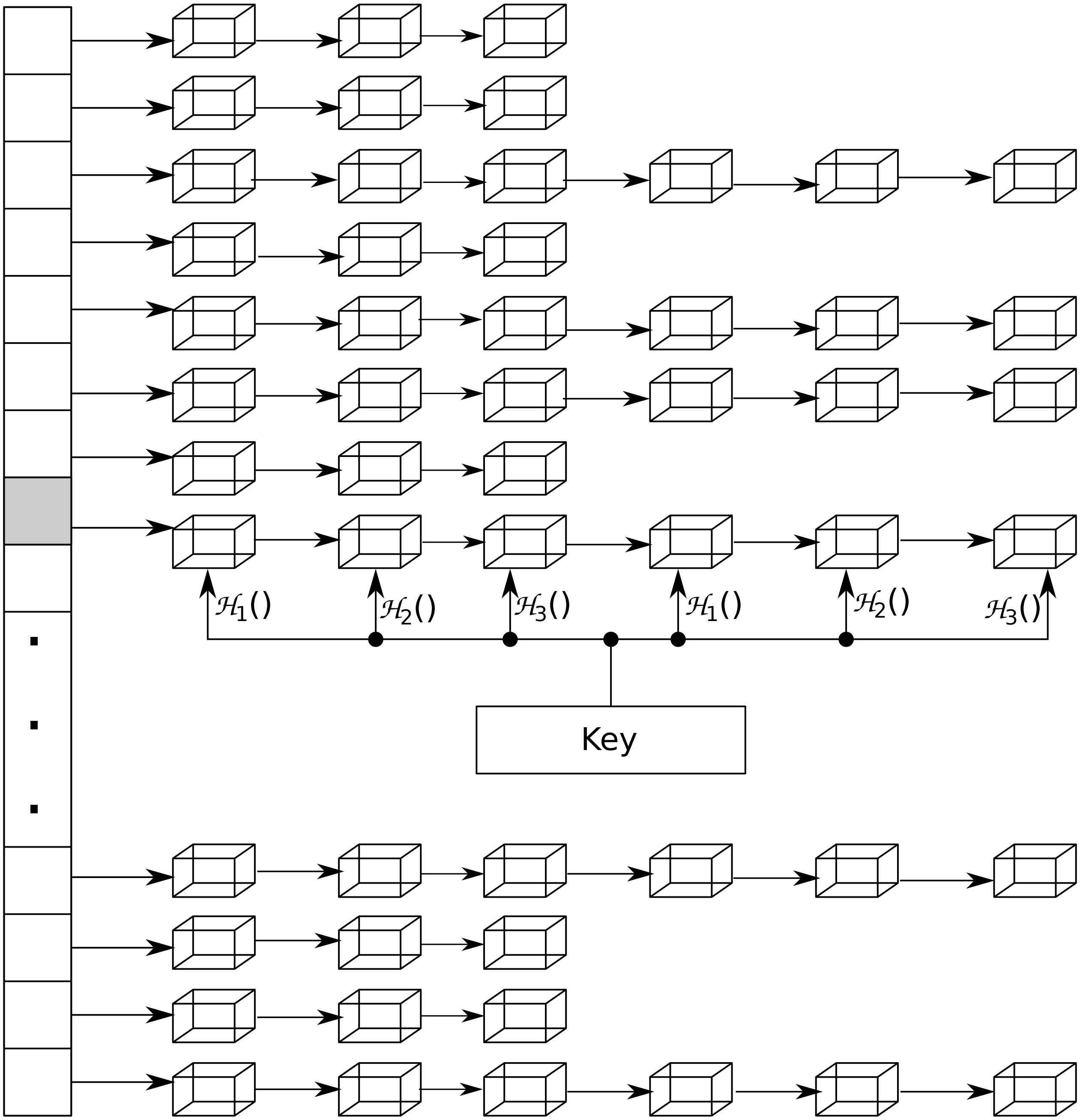


Fig. 3. Lookup of an input item in scaleBF.

presents FPP as follows-

$$FPP = \sum_{x=0}^{m} \left(\frac{x}{m}\right) \binom{m}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(\frac{x-j}{m}\right)^n \quad (3)$$

Equation 3 is calculation of false positive probability of a 3DBF. scaleBF deploys three 3DBF. Therefore, the false positive probability of a Bloom Filter having three 3DBF is

$$FPP = \prod_{p=1}^{3} \left(\sum_{x=0}^{m} \left(\frac{x}{m}\right) \binom{m}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(\frac{x-j}{m}\right)^n \right) \quad (4)$$

Let, there are n Bloom Filter (three 3DBF each), and their false positive probabilities are FPP_i, where $i = 1, 2, 3, \ldots, n$. From Equation (4), the average false positive probability of scaleBF is

$$FPP_{avg} = \frac{\sum_{i=1}^{m} FPP_i}{n} \quad (5)$$

Equation (5) presents the false positive probability of scaleBF.

A. Scalability

Scalability is the key barrier to the modern Bloom Filter. There are numerous Bloom Filter that addresses the scalability issue. However, scalable Bloom Filters are developed based on reordering, hierarchical and forest structure. scaleBF uses simple hashing scheme to enhance the scalability of Bloom Filter. The chaining is the most used hashing data structure. However, chaining has linear search in the worst case, i.e., $O(n)$ time complexity. In other words, all keys are hashed into single chain location. However, it is once in a blue moon in real-world. Besides, most of the chain remains unused. Therefore, the chaining size must be a prime number to avoid the above situation.

Undoubtedly, the scalability is achieved using chaining data structure in scaleBF. The RAM size of the system also plays a vital role in scaleBF. 3DBF allocates memory dynamically which requires few memory blocks be contiguous to satisfy the request by the most modern programming language. Therefore, there is less worry about the unavailability of memory blocks. However, scaleBF does not guarantee the availability of the memory.

Let, P be the slot size and Q be the number of chains to be stored in chaining. The load factor $\alpha = \frac{Q}{P}$, where P is a prime number, and Q is the total Bloom Filter to be inserted. Therefore,

$$Q = \sum_{i=1}^{T} \frac{m_i}{3} \tag{6}$$

where m_i is the size of i^{th} 3DBF. Then, the load factor becomes

$$\alpha = \frac{\sum_{i=1}^{T} \frac{m_i}{3}}{P} \tag{7}$$

The total available bits in scaleBF are

$$\frac{\tau \times X \times Y \times Z \times \left(\sum_{i=1}^{T} \frac{m_i}{3}\right)}{P} \tag{8}$$

where τ is the threshold that depends on the requirements, X, Y, and Z are the dimensions. The τ is calculated by $1, 2, 3, \ldots, \beta$ and β be the number of bits per cell in a 3DBF [12]. For high accuracy, τ is set to 1. However, $\tau = \beta$ defines that false positive is insignificance.

B. Time and Space Complexity

The time complexity is also a key barrier in the scalable Bloom Filter. Hierarchical Bloom Filter or Forest Structured Bloom Filter takes $O(logn)$ time complexity in lookup and insertion operation. Other variants of scalable Bloom Filters also decrease the performance. scaleBF uses $O(1)$ time complexity to lookup and insertion operation on an average case. However, the worst case time complexity is $O(n)$ and it is impractical.

Let, a key κ to be inserted into scaleBF. The κ is hashed into a particular slot of chain and insert into the key κ in desired Bloom Filter (three 3DBF). If the first Bloom Filter is full, then move to the next and so on. Let, the maximum, the size of a particular chain is $\mathcal{C}$. scaleBF uses prime number P to evenly distribute the keys as disclosed in Equation (7). Thus, the size of $\mathcal{C}$ is small. Let us, there are 70% slots empty even if prime number P. That is, 30% slots are filled. Then, each slot has at least 30% of Q which is also very small. However, the P is a prime number, and thus, the distribution is fair enough to fill each slot. Thus, $\mathcal{C}$ is very small and the total time complexity is $O(1)$ on an average. Similarly, lookup cost also $O(1)$ on an average case.

C. Performance

scaleBF also inherits the performance of 3DBF [12]. The insertion and lookup cost depends on the cost of Equation (1) and (2). Equation (1) and (2) uses Murmur hashing [13], which is known as a very fast string hashing. The computational complexity of Murmur hashing is $O(1)$, since, the length of a string is constant and small. Therefore, the Equation (1) and (2) also cost $O(1)$ time complexity. 3DBF enhances the performance by reducing the total number of complex arithmetic operations. Thus, scaleBF increase its scalability without compromising the performance.

IV. Discussion

scaleBF provides impressively very high scalability. However, the initial cost of memory consumption can be high. For instance, insert a key which mapped to the slot 3 of chaining, and creates new three 3DBF. Another insertion key also triggers creation of new three 3DBF which is mapped into a slot, say 2. Thus, the initial cost of memory is high. However, scaleBF is ideal for very large scale membership filtering. Moreover, scaleBF also ideal solution of large memory allocation due to dynamic memory allocation system. scaleBF also depends on the size of 3DBF.

V. Conclusion

Deduplication requires very high scalable Bloom Filter, since, deduplication processes trillions of keys. Moreover, there are diverse applications of high scalable Bloom Filter, for instance, DNA Assembly. In this paper, we have presented a very high scalable Bloom Filter without comprising the performances. In addition, scaleBF also provides insertion and lookup cost of $O(1)$. scaleBF outperforms Bloofi [11], BloomFlash [8], FBF [7], and TB2F [10] in terms of computational time complexity while maintaining higher scalability. However, the scaleBF does not support deletion of an item. Thus, there is no false negative. Interestingly, scaleBF can be applied many research areas to boost up the performance and scalability, and its applicability not limited to NDN, but also Big Data, Cloud Computing, Database, Distriubuted System, IoT, and Computer Networking.

References

[1] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[2] R. Patgiri, S. Nayak, and S. K. Borgohain, "Preventing ddos using bloom filter: A survey," *EAI Endorsed Transactions on Scalable Information Systems*, 2018.

[3] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet mathematics*, vol. 1, no. 4, pp. 485–509, 2004.

[4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, 2008.

[5] A. Lakshman and P. Malik, “Cassandra: A decentralized structured storage system,” *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, 2010.

[6] R. Patgiri, S. Nayak, and S. K. Borgohain, “Shed more light on bloom filter’s variants,” in *Proceedings of the 2018 International Conference on Information and Knowledge Engineering*. CSREA Press, 2018, pp. 14–21.

[7] G. Lu, B. Debnath, and D. H. C. Du, “A forest-structured bloom filter with flash memory,” in *2011 IEEE 27th Symposium on Mass Storage Systems and Technologies (MSST)*, 2011, pp. 1–6.

[8] B. Debnath, S. Sengupta, J. Li, D. J. Lilja, and D. H. C. Du, “Bloomflash: Bloom filter on flash-based storage,” in *2011 31st International Conference on Distributed Computing Systems*, 2011, pp. 635–644.

[9] G. Lu, Y. J. Nam, and D. H. C. Du, “Bloomstore: Bloom-filter based memory-efficient key-value store for indexing of data deduplication on flash,” in *2012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, 2012, pp. 1–11.

[10] W. Quan, C. Xu, A. V. Vasilakos, J. Guan, H. Zhang, and L. A. Grieco, “Tb2f: Tree-bitmap and bloom-filter for a scalable and efficient name lookup in content-centric networking,” in *2014 IFIP Networking Conference(IFIP NETWORKING)*, vol. 00, 2014, pp. 1–9.

[11] A. Crainiceanu and D. Lemire, “Bloofi: Multidimensional bloom filters,” *Information Systems*, vol. 54, pp. 311 – 324, 2015.

[12] R. Patgiri, S. Nayak, and S. K. Borgohain, “rDBF: A r-dimensional bloom filter for massive scale membership query,” *Journal of Network and Computer Applications*, vol. Personal communication.

[13] A. Appleby, “Murmurhash,” Retrieved on August 2018 from https://sites.google.com/site/murmurhash/, 2018.

[14] F. Grandi, “On the analysis of bloom filters,” *Information Processing Letters*, vol. 129, pp. 35 – 39, 2018.

Discovery of Corrosion Patterns using Symbolic Time Series Representation and N-gram Model

Shakirah Mohd Taib[1],
Zahiah Akhma Mohd Zabidi[2],
Izzatdin Abdul Aziz[3]
and Farahida Hanim Mousor[4]
Department of Computer and
Information Sciences
Universiti Teknologi Petronas
32610 Seri Iskandar, Perak, Malaysia

Azuraliza Abu Bakar[5]
Center for Artificial Intelligence
Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor, Malaysia

Ainul Akmar Mokhtar[6]
Department of Mechanical Engineering
Universiti Teknologi Petronas
32610 Seri Iskandar, Perak, Malaysia

***Abstract*—There are many factors that can contribute to corrosion in the pipeline. Therefore, it is important for decision makers to analyze and identify the main factor of corrosion in order to take appropriate actions. The factor of corrosion can be analyzed using data mining based on historical datasets collected from monitoring sensors. The purpose of this study is to analyze the trends of corroding agents for pipeline corrosion based on symbolic representation of time series corrosion dataset using Symbolic Aggregation Approximation (SAX). The paper presents the analysis and evaluation of the patterns using N-gram model. Text mining using N-gram model is proposed to mine trend changes from corrosion time series dataset that are transformed as symbolic representation. N-gram was applied for the analysis in order to find significant symbolic patterns that are represented as text. Pattern analysis is performed and the results are discussed according to each environmental factor of pipeline corrosion.**

***Keywords*—Pipelines corrosion analysis; Symbolic Aggregation Approximation (SAX) representation; corrosion patterns; corrosion factor**

I. Introduction

Time series information is substantial in numerous application areas such as financial market modeling, weather forecasting, sensor systems and motion tracking. The purpose of analyzing the time sequence is to discover hidden knowledge as to figure the future patterns. In the oil and gas industry, pipeline is the main transportation to deliver the products. Pipeline corrosion usually happens in oil and gas industry as a lot of equipment is made from steel. Besides, the natural existence of corroding agents can initiate the chemical reaction that accelerates the corrosion process. Corrosion is the degradation of the material that attacks every component at every stage in the oil and gas industry and could occur because of the chemical reaction with the environmental. Corrosion is a threat to the oil field structures in pipelines, casing and tubing [1]. This problem can be the cause of pipeline's leakage that brings a vast impact to the operation process and system infrastructure cost [2]. Therefore, effective management of tools and equipment maintenance is important to avoid high maintenance cost. Hence, the analysis and monitoring of the pipeline's system is required to discover the important corrosion patterns and predict the consequences of the system's failure. The failure can be measured from the data collection of the static equipment such as sensors. Different types of sensor devices in oil and gas industry produce large-scaled of data that can reach several terabytes per day [3]. This historical time series data requires space and implementation of big data strategy. Therefore, efficient and effective time series representation and similarity searching become one of important issues in analyzing data from sensor devices. Numerous dimensionality reduction methods have been proposed for sufficient time series data representation including Symbolic Aggregate Approximation (SAX). SAX is a symbolic representation method for time series data [4], [5] that offers simple and efficient dimensionality reduction. In this study, SAX is used to transform numerical time series data into a symbolic data representation. The time series data that were recorded by sensor devices in a pipeline contains sensor readings based on several agents that might contribute to the corrosion. In order to discover hidden patterns in the transformed symbolic time series, N-gram model was used as a tool to analyze pattern trends in the corroding agents behavior. The content of this paper was organized as follows. The background of study section contains a discussion on the corrosion issue and data pre-processing method used in this study. The pre-processing consists of data cleaning by fixing the missing values and sorting out the important data that required for analysis. This section includes a further discussion of SAX method, N-gram models as well as Markov chains assumptions. The next section describes a methodology for the study that starts from the data collection until evaluation. In Result and discussion section, the result of the analysis patterns from both SAX and N-gram model will be discussed based on symbolic patterns that are discovered using real pipeline sensors time series data. In the final section of the paper, some conclusive remarks and directions for future work are put forward.

II. Background of Study

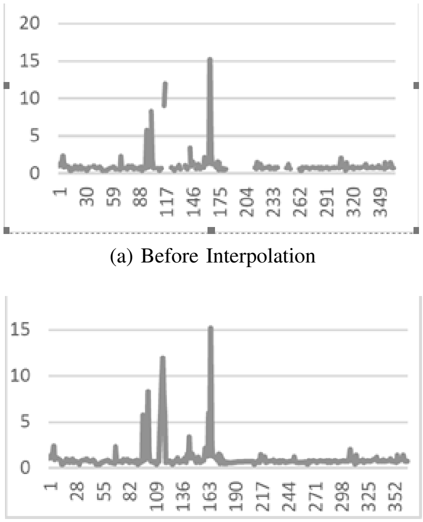
A. Corrosion in Pipelines

Many important electrochemical, chemical, hydrodynamic and metallurgical parameters have been identified as main corrosion factors in pipelines [6], [7]. The effect of main factors such as pH, temperature and ion migration give major influence to the corrosion rates. Solution pH and chloride concentration have a significant relationship that affects the corrosion process [8]. Chloride often occurs in the pipeline as a negative ion when it dissolved in the water phase present in the pipeline [9]. This chloride ion will bond with other elements that allow the corrosion to occur especially in subsea pipelines. Thus, the acidic rain that contains chloride also might be a factor that contributes to the increasing of corrosion rate [10]. Besides that, high temperature in a pipeline can cause internal corrosion. Temperature is a platform to accelerate the chemical and electrochemical processes occurring in the pipeline [11]. A low temperature makes the corrosion rate slowly increases due to the continuous dissolution of ion in a pipeline.

Acquiring data from corroding agents using sensor devices is a practical way for corrosion monitoring and early warning of structural failure as well as prediction of pipelines life [12]. However, large stream data from sensor needs cleansing and analysis to extract meaningful information from it. The feasibility of real-time pipeline monitoring and inspection system using acoustic sensors has been investigated by [2]. They found that this system can provide early detection such as corrosion and leaks of the pipelines but needs improvement on the poor quality of signal measurement and noise that may lead to inaccurate data transmission.

B. Time Series Representation

Time series analysis has been widely used in various fields of research. Esling and Agon [13] defined time series as a collection of values obtained from sequential measurements over time and stored as large dataset which causes the major issue for the high dimensionality of data. Time series tasks can be categorized as prediction, clustering, classification and segmentation [14]. Time series data can be univariate or multivariate when several series simultaneously span multiple dimensions within the same time range. Therefore, a well-defined and approximated representation for the original data is very important in the analysis of time series [5]. There are many approaches that has been highlighted for time series data representation such as Discrete Fourier Transform (DFT) [15], Discrete Wavelet Transform (DWT) [16], Singular Value Decomposition (SVD) [17], Piecewise Aggregate Approximation (PAA) [18] and the symbolic representation approach which is Symbolic Aggregation Approximation (SAX)[5]. Symbolic representation allows the application of data structures and algorithms from the text processing and bio-informatics research. Lin at.al [4] have proposed SAX as a time series representation method by transforming numeric values into alphabet sequence. The data is transformed by PAA representation before it is being symbolize into a discrete string. Therefore, the algorithm extends the PAA-based approach acquiring the calculation and low computational many-sided quality while giving acceptable flexibility for data mining. SAX is the first symbolic representation that offers a dimensionality reduction and a lower bound of the Euclidean distance [4], [19]. Implementation of SAX method in this study is briefly discussed in the next section.



(a) Before Interpolation

(b) After Interpolation

Fig. 1. Interpolation for missing values

C. Data Pre-proessing

In data mining, data pre-processing is an important stage to clean and transform raw data into appropriate format for further mining tasks. The basic pre-processing techniques used in this study are data cleaning and data normalization. Missing data is a common problem in time series dataset due to equipment failures during recording process. In order to transform the raw data into SAX, the missing data need to be filled in to maintain the consistency of the result. One of the missing data solutions for time series is interpolation technique. Linear interpolation can estimate the missing values based on the continuity in a single sequence [20]. Linear interpolation evaluates the estimation of a capacity between two known esteems. Linear interpolation requires assessing a new value by connecting two adjacent values with a straight linear as shown in (1).

$$y = y_1 + (x - x_1)\frac{y_2 - y_2}{x_2 - x_1} \tag{1}$$

Fig 1. shows the example of time series data with missing values and a cleaned data after interpolation using [21] in Microsoft Excel.

D. N-gram Model

N-gram model is a simple text mining model that assigns probabilities to sentences and sequences of words. The concept of N-gram can be demonstrated from the chain rule of probability [22] as shown in (2).

$$P(W) = P(w_4 \mid w_1, w_2, w_3) \tag{2}$$

Whereby, P(W) is a sequence of words and

$$P(w_4 \mid w_1, w_2, w_3)$$

is the conditional probability of word w_4 given the sequence w_1, w_2, w_3 .The sequence of *N* will be represented as $w_1 ... w_n$. Based on equation (2), it can be computed into (3).

$$P(w_1^n = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1})) \\ = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1}) \qquad (3)$$

The chain rules outline the connection between the joint probability of a sequence and the conditional probability of a word given past words. From (2),the estimation of words can be calculated by multiplying a number of conditional probabilities[22]. However, using chain rule is not suitable for this study as the long sequence of symbolic data from SAX words cannot be computed the exact probability, $P(w_n \mid w_1^{n-1})$. Therefore, bigram model is used to predict the conditional probability of the next word. Instead of calculates the probability of the previous symbolic data $P(w_n \mid w_1^{n-1})$, it can be calculated using the conditional probability of the preceding word $P(w_n \mid w_{n-1})$. For example, instead of computing the probability into (4) it can be computed into (5).

$$P(c \mid cbcccccacccc) \qquad (4)$$

$$P(c \mid c), P(c \mid b), P(c \mid a) \qquad (5)$$

Therefore, bigram model is used to predict the conditional probability of the next word [22] and the approximation as shown in (6).

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-1}) \qquad (6)$$

This assumption that depends on the previous frequency of words sequence is called Markov assumption. Markov models are the class of probabilistic model that can accept and anticipate the probability of some future unit without looking too far into the past [22]. Therefore, from bigram, it can derives to the trigram and to the N-gram which takes *N-1* words into the past. Hence, the common equation for N-gram estimation to the conditional probability of the next word in a sequence as shown in (7).

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \qquad (7)$$

III. Methodology

The methodology of this study consists of six stages as shown in Fig. 2.

This study was started by collecting time series data from the oil and gas company. The data is a record of sensor readings for five different corroding agents. Corroding agents in the dataset are described as Agent A, Agent B, Agent C, Agent D and Agent E. The agents are selected environmental factor that may contribute to corrosion rate of the pipeline in oil and gas industry. Each dataset was recorded from March 2010 to December 2016. The data was recorded for one-year analysis. However, as to standardize a count for one-year analysis, the data was selected between January 2011 to December 2016. Fig. 3 shows the actual data that was collected in a numeric form of time series. The data consists of daily sensor's reading for corroding agent A to E and the corrosion rate.

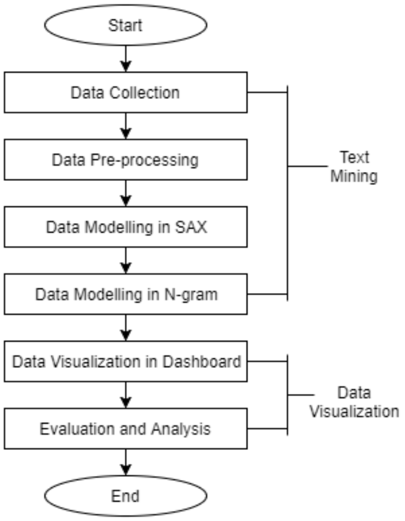
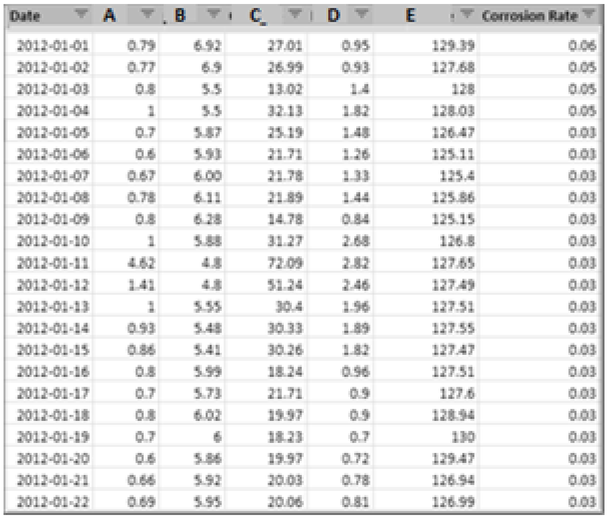


Fig. 2. Research Methodology



Date	A	B	C	D	E	Corrosion Rate
2012-01-01	0.79	6.92	27.01	0.95	129.39	0.06
2012-01-02	0.77	6.9	26.99	0.93	127.68	0.05
2012-01-03	0.8	5.5	13.02	1.4	128	0.05
2012-01-04	1	5.5	32.13	1.82	128.03	0.05
2012-01-05	0.7	5.87	25.19	1.48	126.47	0.03
2012-01-06	0.6	5.93	21.71	1.26	125.11	0.03
2012-01-07	0.67	6.00	21.78	1.33	125.4	0.03
2012-01-08	0.78	6.11	21.89	1.44	125.86	0.03
2012-01-09	0.8	6.28	14.78	0.84	125.15	0.03
2012-01-10	1	5.88	31.27	2.68	126.8	0.03
2012-01-11	4.62	4.8	72.09	2.82	127.65	0.03
2012-01-12	1.41	4.8	51.24	2.46	127.49	0.03
2012-01-13	1	5.55	30.4	1.96	127.51	0.03
2012-01-14	0.93	5.48	30.33	1.89	127.55	0.03
2012-01-15	0.86	5.41	30.26	1.82	127.47	0.03
2012-01-16	0.8	5.99	18.24	0.96	127.51	0.03
2012-01-17	0.7	5.73	21.71	0.9	127.6	0.03
2012-01-18	0.8	6.02	19.97	0.9	128.94	0.03
2012-01-19	0.7	6	18.23	0.7	130	0.03
2012-01-20	0.6	5.86	19.97	0.72	129.47	0.03
2012-01-21	0.66	5.92	20.03	0.78	126.94	0.03
2012-01-22	0.69	5.95	20.06	0.81	126.99	0.03

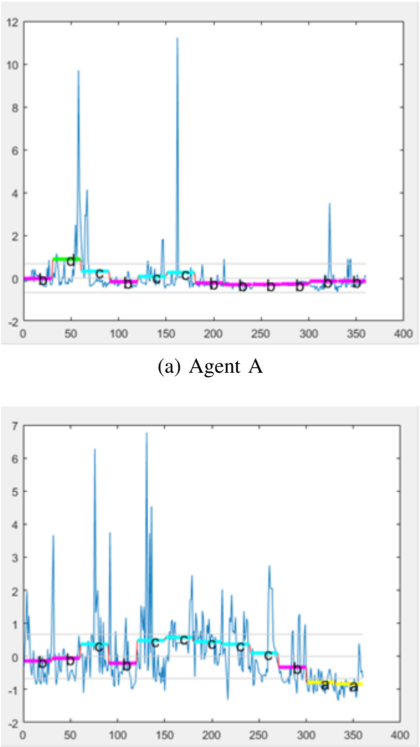
Fig. 3. Time series dataset of corroding agent

The second stage is data pre-processing where the raw data was analysed to detect errors and missing values. Linear interpolation method was applied to impute the missing values in the raw data. Data quality is important in order to achieve a better result in mining task. Other than imputation process, data was normalized and selected data was prepared for the next stage which is data representation using SAX. The result of the translated data is presented in Fig. 4 in the next section. The data are being standardized into one-year analysis for six years starting from 2011 to 2016. Next stage is the N-gram modeling where the transformed SAX data is clustered using N-gram model. N-gram model can assign probabilities to sentences and sequences of words. Therefore, N-gram was used to evaluate yearly trend of each corroding agent based on the co-occurrence of symbolic sequence patterns. In order to visualize the result of N-gram analysis, a dashboard was

TABLE I. SAX REPRESENTATION OF CORRODING AGENT TIME SERIES DATA.

Corroding Agent	SAX Representation
A	b b b d b c b b b b b b
	b b b b b c c c b b b c
	b b b b b d d b b b b b
	b d c b c c b b b b b b
	c c c c b b b c b b b b
	b b b b b c c c b b c c
B	c c c c c b c c c b b b
	b b c c d c d b b a c c
	b b b b a b b d b c c d
	c c b b c b c c b b b c
	b b b b c c c c c c c b
	c b c c b c b c c c c c
C	d b b c b c b b b b b b
	b b b c c d b c b c b b
	b d c c c d d b b a a a
	b b c b c c c c c b a a
	b b b c c c d c b b b b
	b b b b c c d d b b b b
D	c b b c c c b c b b b d
	c c c b b d b b b d b b
	b b c b c c c b b b c c
	c c c c c b b c c b b a
	b b b b b b b b b b c d
	c c c b b c b b b b c b
E	c b c c c c c a c c c c
	c c a c c c a c c c c c
	c c c c c a c c c c c c
	c c c c c c c c c c a a
	b a a b b b d d b d c c

developed using R language. In the final stage, the overall result for this study were evaluated and analysed.



(a) Agent A

(b) Agent C

Fig. 4. SAX patterns for Agent A and Agent C in 2014

A. Result and Discussion

This study set out to investigate the hidden patterns from symbolic time series of pipeline corroding agents using N-gram model. The following discussion will focus on the symbolic representation and analysis of corroding agents' behavior based on N-gram results. A clean dataset was transformed into symbolic representation after interpolation process was completed. TABLE I shows symbolic representation result after transforming numeric time series data using the SAX algorithm. Each corroding agent has six strings that represents the symbolic SAX patterns generated from 2011 to 2016.

From the SAX representation shown in TABLE I, it can be analysed that each word represents a certain range value. This is because SAX uses the concept of Piecewise Aggregate Approximation to get the range of breakpoints for each symbol. For this study, the symbolic range is from *a* to *d* whereby *d* is the highest range value and *a* is lower than 0. Fig. 4 shows the SAX graph for Agent A and Agent C.

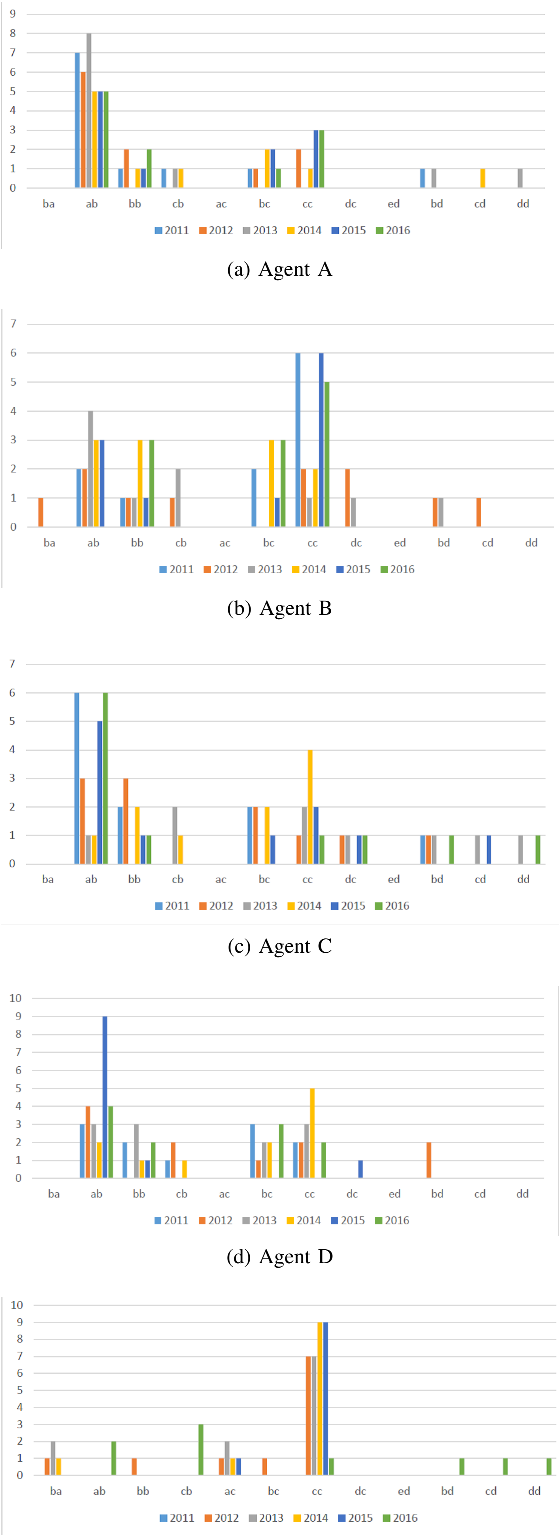
Both agents shown in Fig. 4 have different patterns in 2014. Agent A has minimal changes in the pattern throughout 2014 except for increase trend in the first quarter of the year. Meanwhile Agent B has a decreasing trend in the last quarter of 2014. After transforming the data, the symbolic representation data was used to classify the sequence of each symbolic pattern using N-gram model. According to Bhakkad [23], to predict for which word comes next for particular pattern of document is based on the occurrence of different bigram frequency. Therefore, bigram was used to cluster the pattern trend for each corroding agent in order to discover co-occurrence behavior. The pattern was analyzed by counting the frequency of each bigram that yearly found in SAX graph. Fig 5 shows the bigram frequency for each corroding agent from Agent A to Agent E within year of 2011 until 2016.

Term frequency-inverse document frequency (*tf-idf*) has been conducted in order to evaluate the pattern for all the environmental factors of corrosion rate. Term frequency-inverse document frequency (*tf-idf*) is a statistical measure that commonly used in information retrieval and text mining to evaluate the important of a text document. Term frequency is used to measure the frequently term that occurs in a document while inverse document frequency measures how important a term is. Term frequency took the more frequent words while inverse document frequency also took along the rare words that occur. In this study, term frequency-inverse document frequency is used for pattern categorization. Term frequency in this study will take account on the more occurrence pattern while inverse document frequency will take the rare pattern less occur using log to measure the *tf-idf*. The 0 value in *tf-idf* is consider as the pattern is not very informative as the pattern often occurs throughout the analysis. *Tf-idf* can be categorized into two conditions that are high and low frequency. The high *tf-idf* shows that the bigram frequency is more important compared to othes. Fig. 6 illustrates the *tf-idf* patterns for each corroding agent.

Based on the *tf-idf* result, it shows that Agent C and Agent E have more significant bigrams compared to the other three agents. Agent D has only one important bigram while Agent A has two bigrams that have same *tf-idf* value.

B. Corrosion Rate Pattern

All corroding agents contribute to the corrosion in a pipeline throughout the year. Based on the SAX graphs, the



(a) Agent A

(b) Agent B

(c) Agent C

(d) Agent D

(e) Agent E

Fig. 5. Bigram Frequency of all corroding agents from 2011 - 2016

corrosion rate has different trend for different year. Fig. 7 shows SAX graph of corrosion rate in 2011 and 2016.

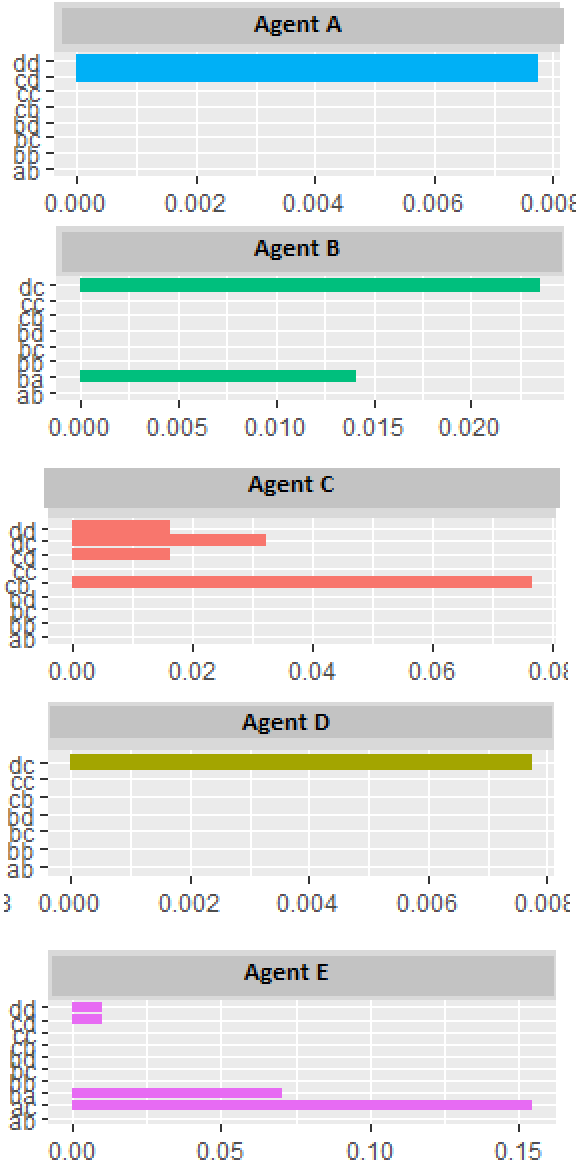
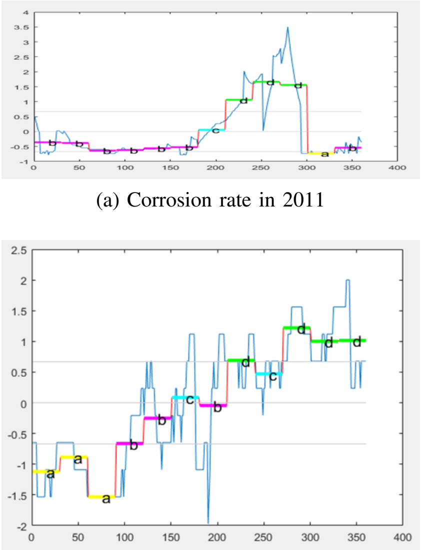


Fig. 6. Terms Frequency Inverse Document (tf-idf) Patterns

C. Relationship between the Bigram Pattern of Environmental Factor and Corrosion Rate

Based on the *tf-idf* result, it can be analyzed that Agent C and Agent E have more patterns than other environmental factors. From 2011 to 2016, there were changes in both agents that accelerated the corrosion rate in a pipeline. Agent D represents only *dc* pattern. From SAX graph, it shows the pattern has decreased from *d* to *c*. The similar pattern can be found in 2014, 2015 and 2016. This pattern shows that there was a decrease time series for Agent D in 3 years. Therefore, it can be concluded that Agent D has less contribution to the corrosion rate throughout the 3 years. Agent A pattern includes *cd* and *dd*. There is an increasing pattern from *c* to *d* and then the pattern remains unchanged in SAX graph. Therefore, the present of Agent A in a pipeline can cause an increasing in corrosion rate for year 2011, 2012, 2015 and 2016. As shown



(a) Corrosion rate in 2011

(b) Corrosion rate in 2016

Fig. 7. SAX patterns for Agent A and Agent C in 2014

in the previous section, SAX graph of corrosion rate for Agent B has decreased from *d* to *c* and significant bigrams based on *tdf-idf* is *dc* and *ba*.

Similar trend is represented by the other pattern, *ba*. Thus, it can be concluded that Agent B does not affect the corrosion rate. Based on the *tf-idf* result, the Agent E pattern is *ac ba cd dd*. The pattern shows an increasing in corrosion rate from *a* to *c* in year 2015. Therefore, it can be concluded that Agent E accelerated the corrosion process in 2015. The pattern shows a decreasing trend from *b* to *a* in 2013. However the it increased from *c* to *d* and become constant until end of year. This shows that environmental factor of Agent E contributes the most of corrosion rate throughout the 6-year analysis. The pattern for Agent C is constant from *c* to *b* in 2013, 2014 and 2016 but it shows a decreasing trend from *d* to *c* for year 2015 and 2016 while it started to increase again from *c* to *d* in year 2016. Throughout the six years analysis, symbolic pattern of Agent C shows that it accelerated the corrosion rate in some particular years while some other years the corrosion rate decreased.

Agent C and Agent E are the two-major environmental factors that contributed to the corrosion rate throughout the six years analysis. The pipeline might contain more substance from Agent C and Agent E that affect internal corrosion.SAX graph for Agent E in 2015 shows high values in Agent E and the increase of corrosion rate in the pipeline. According to [24], corrosion rate is low if the pipeline contains lower amount of some particular agents for several metals. For this case it might be because Agent E is a substance that accelerates the chemical and electrochemical processes occurring in the pipeline [11].

IV. Conclusion

This work describes a symbolic time series approach to analyse corrosion rate in the context of oil and gas industry. Five corroding agents were determined to investigate the measure of corrosion rate. Different types of corroding agents or factors give different rate of pipeline corrosion throughout 6-years observation. The original time series collected with missing values must be processed and represented as symbolic time series representation using SAX. A text-based method which is N-gram is able to define important SAX sequences using term frequency-inverse document frequency (*tf-idf*). This technique allows the pattern to be more descriptive as *tf-idf* reflects the important trend that can occur in the data analysis. Further work on this topic is currently being carried out. Other types of imputation techniques suggested by other researchers may be investigated to improve data quality and to have more accurate result. The understanding of the complexity of the dataset particularly of corrosion process with multiple environmental factors is the most important direction for the future work. A method to predict the correlation between the multiple corrosion factors need to be identified based on the symbolic representation and text mining techniques .

Acknowledgment

The authors gratefully acknowledge for the financial support of this research under internal grant of Universiti Teknologi PETRONAS (UTP).

References

[1] D. Brondel, R. Edwards, A. Hayman, D. Hill, S. Mehta, and T. Semerad, "Corrosion in the oil industry," *Oilfield review*, vol. 6, no. 2, pp. 4–18, 1994.

[2] M. Golshan, A. Ghavamian, and A. M. A. Abdulshaheed, "Pipeline monitoring system by using wireless sensor network," *IOSR J. Mech. Civ. Eng*, vol. 13, no. 3, pp. 43–53, 2016.

[3] R. M. Aliguliyev and Y. N. Imamverdiyev, "Big data strategy for the oil and gas industry: General directions," *Problems of information technology*, vol. 2017, no. 2, 2017.

[4] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and knowledge discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[5] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "Extended sax: Extension of symbolic aggregate approximation for financial time series data representation," *DEWS2006 4A-i8*, vol. 7, 2006.

[6] S. Nešić, "Key issues related to modelling of internal corrosion of oil and gas pipelines–a review," *Corrosion science*, vol. 49, no. 12, pp. 4308–4338, 2007.

[7] Z. Ahmad, *Principles of corrosion engineering and corrosion control*. Elsevier, 2006.

[8] Y. Wang, G. Cheng, W. Wu, Q. Qiao, Y. Li, and X. Li, "Effect of ph and chloride on the micro-mechanism of pitting corrosion for high strength pipeline steel in aerated nacl solutions," *Applied Surface Science*, vol. 349, pp. 746–756, 2015.

[9] A. Fu and Y. Cheng, "Effects of alternating current on corrosion of a coated pipeline steel in a chloride-containing carbonate/bicarbonate solution," *Corrosion science*, vol. 52, no. 2, pp. 612–619, 2010.

[10] M. Ilman *et al.*, "Analysis of internal corrosion in subsea oil pipeline," *Case Studies in Engineering Failure Analysis*, vol. 2, no. 1, pp. 1–8, 2014.

[11] D. Ghazali, "A study of corrosion to the carbon steel in the present of carbon dioxide," Ph.D. dissertation, University Malaysia Pahang, 2010.

[12] M. Tan, F. Varela, Y. Huo, F. Mahdavi, K. Wang *et al.*, "An overview of recent progresses in acquiring, visualizing and interpreting pipeline corrosion monitoring data," in *CORROSION 2018*. NACE International, 2018.

[13] P. Esling and C. Agon, "Time-series data mining," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, p. 12, 2012.

[14] C. Ratanamahatana, E. Keogh, A. J. Bagnall, and S. Lonardi, "A novel bit level time series representation with implication of similarity search and clustering," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2005, pp. 771–777.

[15] S. Wu, Y. Wu, Y. Wang, and Y. Ye, “An algorithm for time series data mining based on clustering,” in *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, vol. 3. IEEE, 2006, pp. 2155–2158.

[16] F. Mörchen, “Time series feature extraction for data mining using dwt and dft,” 2003.

[17] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, *Fast subsequence matching in time-series databases*. ACM, 1994, vol. 23, no. 2.

[18] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, “Dimensionality reduction for fast similarity search in large time series databases,” *Knowledge and information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[19] P. Gao, S. Yan, L. Xie, and J. Wu, “Dynamic reliability analysis of mechanical components based on equivalent strength degradation paths,” *Strojniški vestnik-Journal of Mechanical Engineering*, vol. 59, no. 6, pp. 387–399, 2013.

[20] “Linear interpolation with excel,” http://www.datadigitization.com/dagra-in-action/linear-interpolation-with-excel/, accessed: 2018-09-30.

[21] “Method to calculate interpolation step value in excel,” https://support.microsoft.com/en-us/help/214096/method-to-calculate-interpolation-step-value-in-excel., 2017, accessed: 2018-09-1.

[22] D. Jurafsky and J. H. Martin, “Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition,” pp. 1–1024, 2009.

[23] A. Bhakkad, S. Dharamadhikari, and P. Kulkarni, “Efficient approach to find bigram frequency in text document using e-vsm,” *International Journal of Computer Applications*, vol. 68, no. 19, 2013.

[24] B. S. Pijanowski and I. Mahmud, “A study of the effects of temperature and oxygen content on the corrosion of several metals,” Catholic University Of America Washington Dc Inst. Of Ocean Science And Engineering, Tech. Rep., 1969.

Impact of Android Phone Rooting on User Data Integrity in Mobile Forensics

Tahani Almehmadi[1]
Technical College for Girls in Jeddah
Technical and Vocational Training Corporation
Jeddah, KSA

Omar Batarfi[2]
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, KSA

***Abstract*—Modern cellular phones are potent computing devices, and their capabilities are constantly progressing. The Android operating system (OS) is widely used, and the number of accessible apps for Android OS phones is unprecedented. The increasing capabilities of these phones imply that they have distinctive software, memory designs, and storage mechanisms. Furthermore, they are increasingly being used to commit crimes at an alarming rate. This aspect has heightened the need for digital mobile forensics. Because of the rich user data they store, they may be relevant in forensic investigations, and the data must be extracted. However, as this study will show, most of the available tools for mobile forensics rely greatly on rooted (Android) devices to extract data. Rooting, as some of the selected papers in this research will show, poses a key challenge for forensic analysts: user data integrity. Rooting per se, as will be seen, is disadvantageous. It is possible for forensic analysts to extract useful data from Android phones via rooting, but the user data integrity during data acquisition from Android devices is a prime concern. In suggesting an alternative rooting technique for data acquisition from an Android handset, this paper determines whether rooting is forensically sound. This is particularly due to the device's modification, which a root often requires, that may violate the data integrity.**

***Keywords*—Android; rooting; data integrity; mobile forensics**

I. Introduction

It is scarcely fitting to refer to the device that many people use while receiving the occasional call as a telephone currently. This device's capabilities are growing by no less than the number of mobile subscribers using them. For instance, as of October 2012, about one-third of the US populace (121 million subscribers) had a smartphone [1]. These modern mobile handsets not only match low-priced computers in regards to computing capacity but can also store and generate sizeable quantities of data. Due to the devices' computing capacities (and hardware attributes), the gamut of download-accessible usages and the array of tasks that they can accomplish is astounding. These usages/apps are capable of storing data locally in the modern handset [2]. Among these modern (mobile) handsets, the Android OS has recently become the preferred OS [3]. The mobile devices' capacities promote a rapid uptake in consumer and business settings, and Android's open-source nature thus enables scientific research and "reproducibility" [[3], p. 1937].

The increasing prevalence of smartphones has, however, not been without negative consequences. Smartphones have been (and are being) increasingly used in crimes. These devices have been located at crime scenes in the course of investigations. Criminals have used smartphones to commit email fraud, harass others via texts (SMS), for child trafficking and child pornography, and in narcotics-related communications [4]. They have also become shrewd enough to wipe all traces of their activity. This trend has heightened the necessity of digital smartphone forensics, with Android OS-based devices being no exception. To justify this, the data deposited in smartphones can be very valuable to experts during investigations. Smartphones have already proven themselves to carry a sizeable quantity of probative data that is linked to their users based solely on phonebook contacts, SMS and call histories, instant messaging logs, email threads, and browser history. It is probable that these phones have more probative data that can be traced to a user per byte than the majority of PCs, and the acquisition of these data is harder via forensically appropriate methods. This problem is partly due to the overabundance of cell handsets that are currently available. It is worth noting a large number of Android-based phones, the numerous features they possess, the numerous apps specific to them, and, similarly, the valuable data that can be acquired from local storage. There were approximately 1.4 billion in-use Android phones globally as of September 2015 [2]. Coupled with this overabundance are the general scarcity of hardware and software, and the (deficient) standardization of interfaces in the industry. The multiplicity of Android smartphones implies a variation in the models' features, ranging from the media for data storage, the file system, the OS version and the efficacy of some tools. Even separate Android smartphone models produced by the same maker may require separate data cables and software to access the phone data.

Furthermore, the fact that criminals can wipe their activity off of their smartphone's memory, thereby making it difficult for law-enforcement experts to retrieve data from the devices, has become an investigative challenge [5]. It could be that the existing criminal investigation techniques are still immature. It has already been noted that digital smartphone forensics tools are necessary for investigations since the quality collection and analysis of mobile device data depends on them. However, forensic data extraction methods do not usually validate alterations to subscriber data. The forensic acquisition of data is, to a considerable extent, an "invasive" activity because, typically, investigators "crack" the phone to obtain the needed data. This is often done minus the device owner's consent. As such, cracking the device without exposing the integrity of the needed data is a complicated endeavor. This study focuses on the aspect of user data integrity by exploring whether "rooting" an Android device which is the gaining of administrative privileges before data extraction from Android devices, threatens the user data integrity. The focus on Android devices is due to that Operating System become dominant.

II. Background and Related Work

Although the problem of forensic data acquisition is not new, the majority of expert-designed forensics tools were created out of necessity, and their focus was singularly on the Microsoft Windows OS (a platform that dominated the market for the past 20 years) [3]. Conversely, the cell phones' (factually) comparatively small market share and the differences in (their) hardware and software specifications have hampered the creation of similar tools for cell phones. Smartphones' enhanced capabilities, in comparison to conventional "feature phones," are more intricate. Mobile devices today have features similar to those of computer systems. Android and iOS, the current dominant platforms for smartphones, are built on modern, hardy OSs (Linux for Android and OSX/FreeBSD for iOS) [3]. Even so, these devices' hardware and software are different from those of Windows PCs, for which the present forensics tools and processes are intended. Smartphones, for instance, have no modular hardware (hard drives and detachable RAM cards) that typify modern PCs. Cellular phones may incorporate removable SD memory expansion modules, which can easily be examined via methods similar to those executed on conventional PC systems, but they only serve as auxiliary storage modules. Plus, "many manufacturers are moving away from their use" [[3], p. 1937]. Likewise, cellular phones often run "exotic" file systems and deploy different low-level protocols for accessing data storage modules "that make better use of the embedded non-volatile memory" [[3], p. 1937]. These inbuilt distinctions weaken proper criminal investigations involving cellular phones by using existing tools; thus, novel tools are needed to effectively deal with the new challenges being posed by modern cell phones.

Scrivens and Lin [2] identified the critical elements in forensic investigations on mobile devices, viz. the location(s) for data storage, data mining, and data analytics. The investigator must specifically know where the data are deposited, how the data are deposited, and any attendant file permissions before attempting an extraction. Once these particulars are identified, data extraction must be done since it is an essential part of forensic investigations. Extraction is so critical that using a wrong technique may mess up an investigation. According to Vidas, Zhang, and Christin [6], the prevalence of Android OS devices facilitates the usage of shared attributes to reduce the variety (which digital forensics tools should have) while simultaneously exploiting the capacity for sound data extraction. Makers and network providers tend to maintain competitive advantages by including bonus features in and offering support services to mobile handsets. However, Android handsets have a common framework that is used during acquisition. Specific to Android phones, rooting, in which the investigator or user gains root/administrative privileges where s/he is supposed only to gain unprivileged access, usually involves taking advantage of a security flaw (which is typically dependent on the device and the firmware version) with the intent being installation of unsupported software in the phone. The reasons for rooting Android devices are varied and include the ideological want by users to have control, bypassing controls that are specific to carriers that inhibit the use of particular software, and firmware upgrades (installing an Android version that is higher than that currently supported by the carrier). Rooting, as Grover [1] contends, essentially enables the user to implement elevated-privilege functions on the handset that are usually unavailable in regular user modes. It may be used legitimately or illegitimately. The user may desire to circumvent security controls or to interfere with the data collected via security apps. Overall, rooting can consequently undermine the phone's operating system's security, alter parts of the phone that may collect users' data, diminish interoperability and endanger the device provider's warranty.

Nevertheless, despite the apparent compromises to user data integrity, root access may be inevitable when forensic investigators legitimately deploy it for data extraction. This is contingent upon the situation and the needed data. Whenever possible, root access ought to be avoided.

A. Related Work

Android phones are usually made up of some partitions that are usually mapped to Memory Technology Device (MTD)-type devices. The exact partitioning scheme is dependent on the vendor configuration, but generally, Android phones typically have six partitions. The most common partitions are the /system, /user data, /cache, /boot, and /recovery [5], [6]. The /user data partition is the most forensically pertinent because all the data generated from apps installed by the user is deposited in this partition. As such, wiping it out is like performing a factory reset. It is from the /user data partition where evidence files are often acquired. Alternatively, the /recovery partition, which is "the alternative "system" partition" [[5], p. 288], can be exploited when the system booting fails or when the custom ROM has to be flashed. Forensic investigators use this partition when acquiring a system partition image. Notably, in normal mode, no application data is deposited in the /recovery partition; therefore, data corruption or overwriting there has no likelihood of altering data on the phone that may subsequently be used in a criminal case.

Acquiring data from Android phones is generally categorized into physical and logical acquisition techniques [7]. Logical acquisition methods (in which the focus of this study lies) include file/folder copying, Content Providers, and Recovery Mode [7], whereas physical acquisition techniques involve data partition imaging. Son et al. [7] focused on the Recovery Mode. In determining whether the Android Recovery Mode maintains the integrity of the user data during its acquisition, the authors justified that the Mode can grant administrator access while the phone is in a state where the corruption of the user data can be reduced. Conversely, for (the) imaging of the data partition containing the user data and/or copying files/folders, the phone must be rooted first. In this case, the phone must be booted normally. Normal booting, as Son et al. argue, may not ensure the integrity of the user data or that of unallocated data. Therefore, the authors detailed a process intended to lessen the time and extra work required for the forensic investigation of a suspect Android phone. From the procedure, they developed a tool (Android Extractor) to automatically execute the process via a series of experiments using several Android device models. Their tests confirmed the preservation of the integrity of the user data. Comparatively, the JTAG (Joint Test Action Group), which the authors used for physical data acquisition, was effective in fully acquiring the device data. When the JTAG is used first before the Android Extractor, they concluded (based on the JTAG-compatible devices that were used) that JTAG also maintained user data integrity. However, Hazra and Mateti [5] noted that the JTAG forensics technique of acquiring memory data is executed only when data acquisition via physical or logical extraction is unsuccessful and that it is risky. Although it is useful in extracting locked data, the risk of losing evidence is always there.

In [4], the authors noted that imaging the device's memory is critical in mobile forensics because the memory may contain useful data. Its access can be possible by rooting the device. They detailed a procedure for acquiring all the information from Android Negated AND (NAND) flash(ing). One method suggested the facilitated collection of a byte-by-byte duplicate of the NAND flash per se to recover deleted data. The process required rooting the device to extract a dd image of the appropriate partition(s) and store it in a detachable SD card mounted in the phone, after which the (memory) dumps were examined for prospective evidence. Its disadvantage is that a microSD card slot must be present, which is a deficiency present in many popular Android phone models.

Moreover, extracting a dd image file is likely "when permissions are altered to gain access to the root directory" [[4], p. 3]. As such, rooting is not forensically reliable. Furthermore, root access to obtain the dd image requires the installation of a 3rd-party program in the phone. This would make the acquired data is used as evidence, inadmissible in court. It must be noted that there are other ways to gain administrative privileges on other Android phones that require no 3rd-party software installations. Rooting via 3rd-party installation(s) could be customized to be forensically sound if alternative ways of gaining root privileges are found.

In [6], the authors outline a process for acquiring the logical and physical images of phone storage via the custom recovery image (CRI) technique, and its focus is on Android phones' /recovery partition and the Android Recovery Mode. It requires altering the /recovery partition. Nevertheless, as discussed earlier, the /user data partition is the partition of interest since much of the data that forensic analysts are interested in is found there. As such, the alteration of the /recovery partition will not affect the data. Its operational outline is as follows: (i) acquire a CRI that incorporates the special utilities that facilitate the recovery of the data, ADB, and superuser; (ii) flash the CRI to the Android phone; (iii) reboot the phone in /recovery mode; and (iv) use the command "ADB shell" from the forensic computer terminal "to execute data recovery binaries from the recovery image" [[2], p. 5]. Some data dumping utilities may be utilized, which are contingent upon the flash storage technology in use. Many Android phones use MTD [5]. The Media Technology Device system is an extraction layer for raw (NAND) flash phones that grants software permission to use one interface in accessing multiple flash technologies or a device driver used for directly accessing NAND flash storage. The nanddump for MTD phones may be executed to acquire "NAND data independent of the higher-level filesystem deployed on the memory" [[6], S17]. For phones with no MTD mechanism, other acquisition methods must be used. The dd utility, for instance, may be utilized for copying data. Both of these utilities may be deployed in the recovery of a physical image. Additionally, it is worth noting that not all files are necessarily warehoused in the onboard memory since many Android phones support one microSD module. While the user can install particular apps and store specific data on their phones, some makers may opt to install the /user data partition in its entirety on the module.

The work of Son et al. [7] continues that of [6], although their focus is on the issue of data integrity. After the creation of the custom recovery mode image, the phone must be booted in the flash mode for the image to be flashed to /recovery (or /boot). Here, Son et al. emphasize a crucial aspect associated with the data integrity. If the image is flashed to the /recovery space, the phone ought to shift to Recovery Mode after being flashed. However, the phone "must be manually entered into Recovery Mode" [[7], S7]. In the case that booting into Recovery Mode does not work, the phone will go on to boot normally, thus using the /user data partition and possibly compromising the integrity of user data.

Conversely, in the case that the image is flashed to the /boot partition, the phone may subsequently, instantly and automatically go into recovery mode. With root permission in this mode, forensic investigators can obtain device access via the use of the Android Debug Bridge (ADB) command. From here, investigators can acquire all the needed data. This mode was the basis of the Extractor that was developed and deployed [[7], S8]. Based on the two primary data acquisition methods outlined earlier (data partition imaging and file copying for emphasis), mounting the partition to acquire the (targeted) partitioned unit is unnecessary. Nevertheless, for the file unit to be acquired, the /user data partition ought to be mounted in read-only mode. In this way, data acquisition can be made via the ADB pull command and, more importantly, data integrity is guaranteed.

III. Research Motivation

Data extraction from smartphones during a forensic investigation poses a number of challenges for forensic experts. By using the proper techniques and tools, it is possible to mine useful data from call logs, contact lists, SMS and email threads and browser history. However, the integrity of users' data during acquisition is a major issue for forensic analysts. This need is what has prompted the design of this study, its specific focus on rooting, and the data integrity concerns that have been posted. Therefore, we seek to compare user data integrity when an Android phone is rooted with data extracted from the phone via a custom recovery image, which is believed to affect only the recovery partition without the user data partition. In addition, we compare them with the basic data extracted from the phone before rooting.

1) Hypothesis: If versatile, high-reliability rooting software is used on an Android phone and user data is extracted using forensic software, all the data can be acquired without changing its integrity. These data can thus be used as reliable evidence during forensic investigations.

IV. Experiments

The provision of a proper environment for performing the (intended) experiment is crucial to ascertain that the findings drawn from it are correct. The data acquisition tools are detailed below. Table.I for hardware tools and Table.II for software tools. It should be noted that all the programs used in the experiment's implementation are licensed.

TABLE I. Hardware Tools

Hardware	**Specification**
Dell Inspiron 15 7000	Intel Core i7, 2.80 GHz, 16 GB
Samsung Galaxy S4	GT-I19505
USB Cable	Micro USB Data Charger Cable
MicroSD card	64 GB

TABLE II. Software Tools

Software	**Specification**
Microsoft Windows10	64-bit
SAMSUNG USB Driver for Mobile Phones	Driver definitions to connect to the computer
Android Debug Bridge (ADB)	Access the mobile data on the computer
KingoRoot [8]	PC Version
Odin v3.09	A utility developed by Samsung to flash a custom recovery image to a Samsung Android device
TWRP recovery image [9]	Custom Recovery Image (CRI)
Belkasoft Evidence Center v9.2 - Trial version [10]	It analyzes digital evidence stored in computers and mobile devices
FileAlyzer v2.0	Tool to analyze files

A. Data Acquisition

The experiment will use ADB commands, the custom recovery image, and rooting techniques for data acquisition. A comparison will then be made to determine the effect of Android device rooting on user data integrity. The first step is shown in Fig.1. In detail, a backup was taken from an Android phone using ADB before any rooting operations were performed on the device. ADB is one of the command line tools that constitute the Android SDK package. It allows communication with Android devices and performs actions such as app installation and debugging and aids the safe backup of device and app data on PCs, regardless of the OS. Thus, after enabling developer options and connecting the Android phone to a PC, we ran the command-line interface to make a backup using ADB commands.

For the 2nd stage, a custom recovery image (CRI) was used in data acquisition. The last acquisition method focused on modifying the recovery partition. However, the important content is in the /user data partition, and so modifying the /recovery partition will not affect these data. The data can be acquired from the partition via the ADB pull command or by using the copy process to the MicroSD card from the TWRP homepage. We used the process of copying to the MicroSD card to interface with the smartphone while in recovery mode and extract all files and folders. In the 3rd and last stage, the researchers rooted the device using KingoRoot. KingoRoot Android works on Windows. It supports almost any Android device and version, is risk-free and can unroot at any time. After successfully rooting the Android phone, we used the Belkasoft Evidence Center backup that based on a dd command to gather data. As shown in Table.III the backup file characteristics and their corresponding hash value.

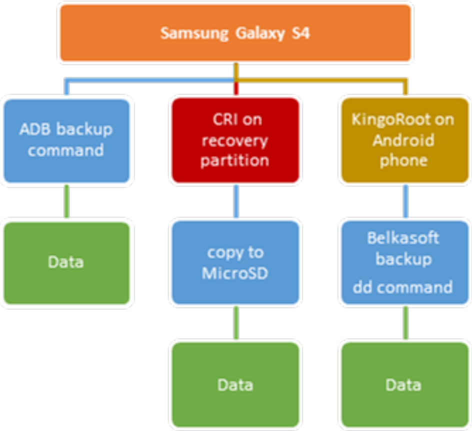


Fig. 1. Step 1 in data acquisition

TABLE III. The Backup File Characteristics and their Corresponding Hash Value

ADB Backup GalaxyOriginalAndroid.ab - 4.45 GB 7BB2BA975D0E69E1CEFE5CCE2965CC1726597525
CRI GalaxyCRMI- 12.6 GB 5609BB28440CB5B20F5C1A25AA750F972BEFAB8A
KingoRoot - Belkasoft Backup GalaxyRootedAndroid.dd - 14.6 GB 0CF0458CB55CADDF495DA8E45A6A9DB8710C3453

B. Data Analysis

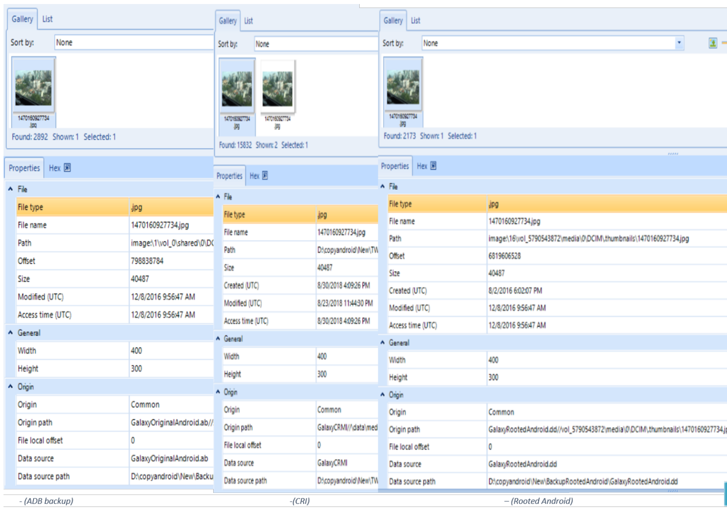
The Belkasoft Evidence Center program was used to extract and analyze the digital evidence from the three Android backups. The Images and Memos files were analyzed by selecting a random file from the extracted folder in the first phase and comparing the hash value of the file with the corresponding file extracted in the second and third phases of the experiment.

Images and Memo Files: The sample file (1470160927734.jpg) was extracted from the Images folder and analyzed using the Belkasoft Evidence Center as shown in Fig.2.

From the sample file extracted from the Images folder in the three acquisition states that were executed, it can be seen that the image's name, shape, identity, and actual path are retained. It can also be noted from the FileAlyzer report that the examined Memo also has the same hash values as shown in Fig.3.

C. Main Points of the Analysis

The results of the illustrated analyses indicate that no data changes occurred during the rooting process or during data extraction. This result is consistent with the results achieved recently, despite the different experiences and programs used [11]. Nevertheless, the results of the folder analysis show an apparent discrepancy in the amount of data that was retrieved using the Belkasoft Evidence Center. The reason for the



Found: 2892 Shown: 1 Selected: 1

File	
File type	jpg
File name	1470160927734.jpg
Path	image:\1\vol_0\shared\0\DC
Offset	798838784
Size	40487
Modified (UTC)	12/8/2016 9:56:47 AM
Access time (UTC)	12/8/2016 9:56:47 AM
General	
Width	400
Height	300
Origin	
Origin	Common
Origin path	GalaxyOriginalAndroid.ab//
File local offset	0
Data source	GalaxyOriginalAndroid.ab
Data source path	D:\copyandroid\New\Backu

- (ADB backup)

Found: 15832 Shown: 2 Selected: 1

File	
File type	jpg
File name	1470160927734.jpg
Path	D:\copyandroid\New\TW
Size	40487
Created (UTC)	8/30/2018 4:09:26 PM
Modified (UTC)	8/23/2018 11:44:30 PM
Access time (UTC)	8/30/2018 4:09:26 PM
General	
Width	400
Height	300
Origin	
Origin	Common
Origin path	GalaxyCRMI//\data\med
File local offset	0
Data source	GalaxyCRMI
Data source path	D:\copyandroid\New\TW

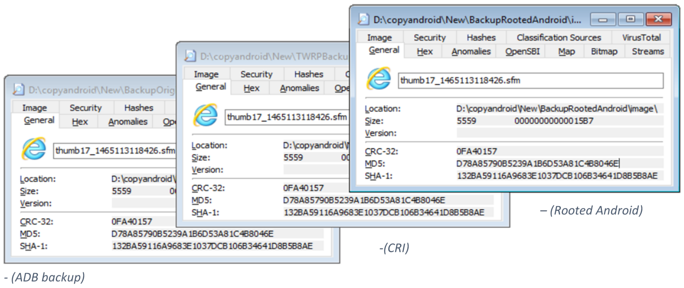
-(CRI)

Found: 2173 Shown: 1 Selected: 1

File	
File type	jpg
File name	1470160927734.jpg
Path	image:\16\vol_5790543872\media\0\DCIM\.thumbnails\1470160927734.jpg
Offset	6819606528
Size	40487
Created (UTC)	8/2/2016 6:02:07 PM
Modified (UTC)	12/8/2016 9:56:47 AM
Access time (UTC)	12/8/2016 9:56:47 AM
General	
Width	400
Height	300
Origin	
Origin	Common
Origin path	GalaxyRootedAndroid.dd//vol_5790543872\media\0\DCIM\.thumbnails\1470160927734.jpg
File local offset	0
Data source	GalaxyRootedAndroid.dd
Data source path	D:\copyandroid\New\BackupRootedAndroid\GalaxyRootedAndroid.dd

– (Rooted Android)

Fig. 2. Analyze sample file (1470160927734.jpg)



D:\copyandroid\New\BackupOrig

thumb17_1465113118426.sfm

Location:	D:\copyandroid\
Size:	5559 0
Version:	
CRC-32:	0FA40157
MD5:	D78A85790B5239A1B6D53A81C4B8046E
SHA-1:	132BA59116A9683E1037DCB106B34641D8B5B8AE

- (ADB backup)

D:\copyandroid\New\TWRPBacku

thumb17_1465113118426.sfm

Location:	D:\copyandroid\
Size:	5559 00
Version:	
CRC-32:	0FA40157
MD5:	D78A85790B5239A1B6D53A81C4B8046E
SHA-1:	132BA59116A9683E1037DCB106B34641D8B5B8AE

-(CRI)

D:\copyandroid\New\BackupRootedAndroid\i...

thumb17_1465113118426.sfm

Location:	D:\copyandroid\New\BackupRootedAndroid\image\
Size:	5559 00000000000015B7
Version:	
CRC-32:	0FA40157
MD5:	D78A85790B5239A1B6D53A81C4B8046E
SHA-1:	132BA59116A9683E1037DCB106B34641D8B5B8AE

– (Rooted Android)

Fig. 3. Analyze sample file (thumb17_1465113118426.sfm)

different amounts of data goes back to the repeated files, where we notice the height of the images in the backup using CRI Fig.5, while for the backup using rooting Fig.6, we see a high number of documents. While this amount of data is not shown in the backup using ADB as shown in Fig.4. It is also worth noting that the tools that are used to install the root is 3rd-party utilities on the Android device. Nevertheless, the utilities did not affect the final data that was recovered.

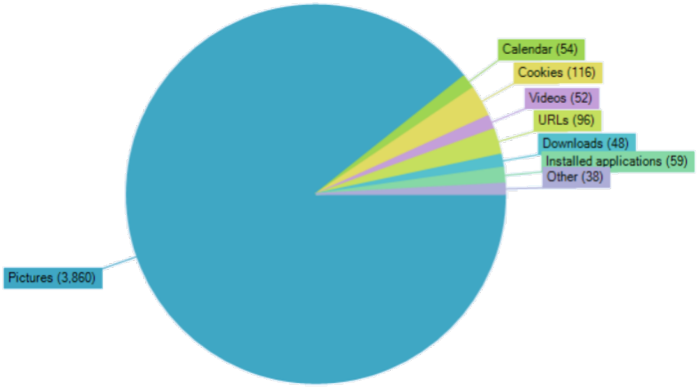


Fig. 4. Samsung Galaxy S4 – ADB Backup

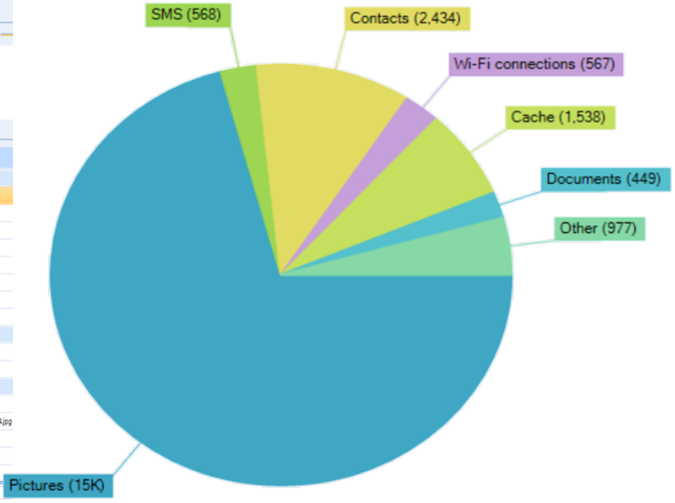


Fig. 5. Samsung Galaxy S4 – CRI

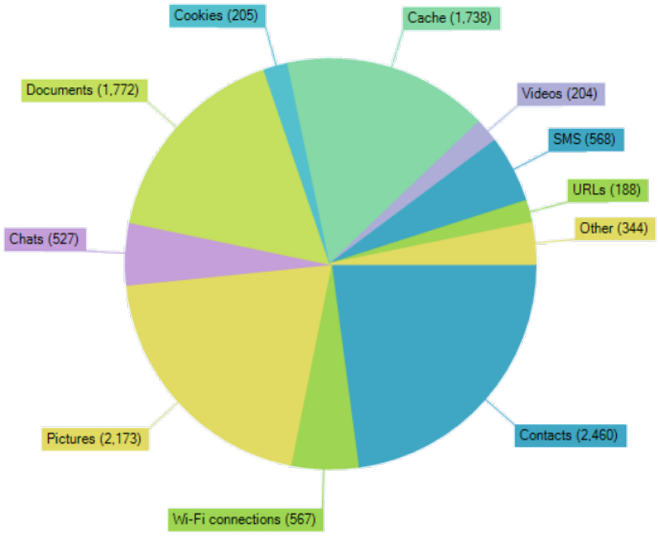


Fig. 6. Samsung Galaxy S4 – KingoRoot - Belkasoft Backup

V. Conclusion

The use of Android devices around the world is growing exponentially. Unfortunately, this rapid growth has led to the misuse of these devices. Similarly, smartphones are now important in criminal investigations. The data stored in different applications in smartphones can be used by forensic experts during the investigation of a crime. There are different tools and methods used to get and extract data from Android smartphones.

This paper sought to investigate the impact of rooting Android phones on the integrity of user data and the search for any damage resulting from the rooting of the device since Android device rooting to acquire physical data necessitates modifications to the device data. Herein, we did not notice any effect on the user data during the process of rooting. Believe it is preferable to document the processes and events during the extraction process and to avoid unnecessary changes to the user data.

The rooting process is therefore legally valid. In addition, the evidence extracted from android devices as a result of the rooting process is sound, reliable evidence of sentencing in criminal cases.

Acknowledgment

We thank Belkasoft LLC who provided support that greatly assisted the research.

References

[1] J. Grover, “Android forensics: Automated data collection and reporting from a mobile device,” Digital Investigation, vol. 10, pp. S12–S20, 2013.

[2] N. Scrivens and X. Lin, “Android digital forensics: data, extraction and analysis,” in Proceedings of the ACM Turing 50th Celebration Conference-China, 2017, p. 26.

[3] D. Votipka, T. Vidas, and N. Christin, “Passe-partout: A general collection methodology for Android devices,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 1937–1946, 2013.

[4] J. Lessard and G. Kessler, “Android Forensics: Simplifying Cell Phone Examinations.,” 2010.

[5] S. Hazra and P. Mateti, “Challenges in Android Forensics,” in International Symposium on Security in Computing and Communication, 2017, pp. 286–299.

[6] T. Vidas, C. Zhang, and N. Christin, “Toward a general collection methodology for Android devices,” digital investigation, vol. 8, pp. S14–S24, 2011.

[7] N. Son, Y. Lee, D. Kim, J. I. James, S. Lee, and K. Lee, “A study of user data integrity during acquisition of Android devices,” Digital Investigation, vol. 10, pp. S3–S11, 2013.

[8] “KingoRoot for Android, the best One Click Root Tool/APK for free.” [Online]. Available: https://www.kingoapp.com/. [Accessed: 22-Dec-2018].

[9] “Download TWRP for jfltexx.” [Online]. Available: https://dl.twrp.me/jfltexx/. [Accessed: 22-Dec-2018].

[10] “Belkasoft: Evidence Search and Analysis Software for Digital Forensic Investigations.” [Online]. Available: https://belkasoft.com/. [Accessed: 22-Dec-2018].

[11] M. Hassan and L. Pantaleon, “An investigation into the impact of rooting android device on user data integrity,” in Emerging Security Technologies (EST), 2017 Seventh International Conference on, 2017, pp. 32–37.

Neighbour-Cooperation Heterogeneity-Aware Traffic Engineering for Wireless Sensor Networks

Christopher Mumpe⋆,1, Da Tang⋆,2, Muhammad Asad†,3, Muhammad Aslam‡,4,
Jing Chen⋆,5, Jinsi Zhu⋆,6, Luyuan Jin⋆,7
⋆School of Computer Science and Technology, Dalian University of Technology. Dalian - China
†Department of Computer Science and IT, Superior University, Gold Campus, Lahore, Pakistan
‡Department of Computer Science, COMSATS University Islamabad, WAH CANTT, Pakistan

***Abstract*—Extending the operational duration is a major field of interest in Wireless Sensor Networks (WSNs). This lifetime enhancement task challenges researchers to design an energy efficient traffic engineering which minimizes the dissipation energy and retain the expected quality of routing protocols. Network lifetime can be prolonged by balancing the energy optimization throughout the network period over which sensors relay data traffic towards Base Station (BS). Existing techniques of continuous and autonomous reporting sensor nodes, offer an opportunity to design the sensing and reporting co-operation between sensor nodes. Nearby nodes with similar reading environment can cooperate with each other to avoid transmission redundant information. In this paper we propose "Adaptive Inter-Networking Improved (AINI)" multi-hop routing protocol with co-operate sensing of inter and intra cluster communication by exploiting the concept of tripling the sensor nodes. Proposed routing protocol improved the reliability of whole network by improving the reliability of inter-cluster multi-hoping. Sensor nodes use the shortest path to deliver data to CH using intra-cluster multi-hoping and these CHs are accountable to forward this data to BS using inter-cluster multi-hop communication. Proposed routing protocol resolves the certain issues of WSN like network lifetime, network stability and CHs selection technique. To prove the efficiency of our proposed model we compared the simulation results with existing state-of-the art routing protocols such as, LEACH, LEACH-C, SEP, ESEP and DEEC. Experimental results shows the benefits of neighbour cooperation and heterogeneity-aware by the performance of proposed protocol over existing state-of-the-art routing protocols.**

Keywords*—*Wireless Sensor Networks; energy efficient; clustering; multi-hop; routing protocol

I. Introduction

Wireless Sensor Network (WSN) comprises a large number of sensor nodes which used to measure and monitor the field such as health monitoring, battlefield surveillance, sensing of light, sound, traffic monitoring, industrial control, vibration, humidity, temperature, etc [1]. After sensing, sensor nodes forward this information to the Base Station (BS). Sensor nodes are randomly deployed in hostile environments which are equipped with limited battery power and processing capabilities with the objective of data collection from sensing field and deliver it to user interface for analysis [2], [3]. Energy efficient routing is a major issue to resolve in WSNs as efficient routing helps to prolong the connectivity of sensor nodes with the network [4]. Stability and lifetime enhancements for WSNs without negotiating the requirement of network are the main aim of on-going research. In order to achieve these aims of WSNs, the network can be divided into groups called clusters. Dynamic CH selection was considered to select a suitable CH which can be implemented with a centralized or a distributed approach in WSNs [5]. Whereas, heterogeneous routing protocol comes under the hierarchy of centralized and distributed clustering [6]. It is noticed that clustering can also improve the load balancing, scalability and connectivity of the network [7], [8]. Certain Cluster Head (CH) selection criterion is used to select a CH for a specific cluster. This CH is responsible for special tasks to perform for its specific cluster such as, assigning TDMA slots to sensor nodes for receiving data, compress this data with complex calculation and transmit it to BS. In this way, energy of CH dissipates much earlier than other sensor nodes which leads to the uneven energy dissipation among the network [9]. For this purpose, CHs should be elected dynamically in each round for load balancing and maintain the energy level of network [10]. Various routing protocol have been proposed using distributed clustering [11], these clustering routing protocols generate random CH selection based on probabilistic method, which usually results into un-even cluster formation [12]. This distributed clustering choose CHs randomly, by ignoring the residual energy which leads to earlier expiration of low energy nodes [13]. Thus, uniform distributed clustering is expected to resolve this uneven energy dissipation of CHs by allocating the variable to highest energy level nodes inside particular cluster [14], [15]. Figure 1 shows the example of distributed clustering.

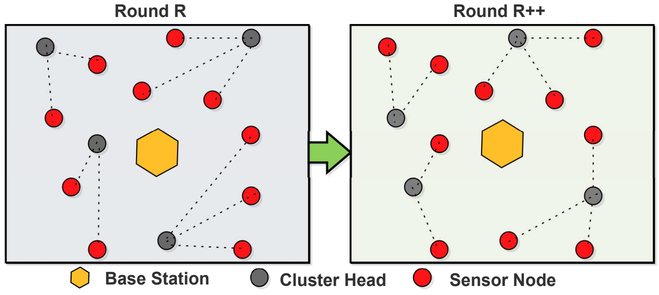


Fig. 1. Sample of Distributed Clustering

In order to sense data in complex environment heterogeneous sensor nodes are used to enhance the nodes capability and functionality [16]. For example, sensor nodes sparsely deploy at higher hierarchy layers are expected to execute more complex tasks. In this paper an heterogeneous clustering routing protocol named "AINI: Adaptive Inter-Networking Improved" routing protocol for WSNs is proposed. To improve the lifetime and stability of network (intra-cluster for nodes to CHs and inter-cluster for CHs to BS) multi-hop communication is used. Existing dynamic state-of-the art clustering routing protocol like LEACH, LEACH-c, SEP, ESEP, and DEEC for WSNs have achieved better performance of network. Trailing their ideas, we propose a new concept of tripling the sensor nodes into Child Sensor Nodes (CSN) which has further improved the CH selection technique stability and lifetime of network.

The rest of the paper is organized as follows: Section 2 contains the brief literature review of existing work. Section 3 described the network model of proposed routing protocol. Proposed model, network operations and CH selection technique is briefly explained in Section 4. Section 5 defines the performance measures. Experimental results and simulations are discussed in Section 6. Experimental methodology is given in Section 7. Conclusion and future work are described in Section 8.

II. Related Work

Extensive research on routing protocols has been proposed based on structure and hierarchy of the WSNs. Low Energy Adaptive Clustering Hierarchy (LEACH) [17], is a pioneer clustering routing protocol, many routing protocols were proposed based on it. In this literature review, we are going to mention few protocols from which our protocol is encouraged and was able to overcome the problems over previous routing protocols. LEACH divides the network into clusters and assigns a CH to each cluster for communication with BS. These CHs are randomly chosen and the rest of the nodes are called member nodes of cluster. These CHs have added responsibility to receive data from all nodes and transmit it to BS which causes CHs to dissipate more energy than member nodes and causes the early death. The problem in LEACH is, it is distributed and only designed for homogeneous network. In [18], LEACH-centralized is proposed to overcome the existing issue, though it improves the network lifetime over LEACH but due to the limited scalability LEACH-c could not perform well in large scale networks. In [19], Stable Election Protocol (SEP) two-level heterogeneous network is proposed. SEP introduces two types of nodes, nodes with higher energy are referred as advance nodes and the other nodes are called normal nodes. Most probability to become a CH is of advance nodes as compared to normal nodes. Problem in SEP was, it performs distributed cluster-formation which results the uneven number of clusters and secondly it only allows for two-level heterogeneity. In order to further enhance SEP protocol Enhanced-SEP [20] protocol was proposed with three level heterogeneity. In [21] Distributed Energy Efficient Clustering (DEEC) proposed heterogeneous network with distributed property. In DEEC, CHs are chosen according to the initial and residual energy of nodes; furthermore, DEEC holds multi-level heterogeneity but the problem with DEEC is, it generates uneven CHs in each round. In [22], Hierarchical Cost Effective LEACH (HCEL) protocol proposed three-level heterogeneous network model. This protocol intends to perform better in heterogeneous network in terms of network lifetime and stability. But the simulation results of HCEL shows gradual reduction of deployment cost ratio in dense network. In [23], Advanced-Multi-Hop LEACH proposed for heterogeneous network, which uses optimal path for communication between CH and BS. The authors claim that using optimal number of clusters and hops using multi-hoping can results better network lifetime and stability period over single-hop LEACH. In [24], Threshold sensitive Energy Efficient sensor Network (TEEN) routing protocol proposed for reactive networks. It was simple temperature sensing protocol, major drawback of this protocol was that, it continuously sense the field and transmit only when there is a change in data values. To enhance the TEEN protocol, Distance Adaptive Threshold sensitive Energy Efficient sensor Network [25] (DAPTEEN) was proposed. DAPTEEN enhanced the network lifetime and CH selection technique of TEEN. In this literature review, various routing protocols are presented and each research contribution tries to overcome the energy dissipation problem and added some improvements in network lifetime of WSNs. Some protocols enhance the stability of network while some protocols added more heterogeneity in nodes. In this regard we propose AINI: Adaptive Inter-Networking Improved routing protocol, with advance CH selection and improved routing technique. Proposed protocol and its models are described in next sections.

III. Network Model for Proposed Protocol

Proposed routing protocol develops cluster-formation and then proceeds tripling connectivity among selected CHs. In this way the proposed model allows general network model settings and shows enough flexibility of execution within any random deployment of sensor nodes. Proposed network model considers some key assumptions such as; (1) linear radio model of radio characteristics to carry on wireless communications, (2) static nodes deployment to ensure nodes maintain certain location-ID throughout the network operations, (3) contain heterogenous characteristics according to the energy resources during network operation and (4) development of multi-hop intra cluster communication and multi-hop inter cluster communications during transmission of sensed information. Our goal is to minimize the energy dissipation in order to improve the network stability and lifetime. Execution of proposed model create unique clustering in which nodes perform tripling bonding and then start communication with selected CHs. Network topology of proposed protocol is shown in Figure 2

IV. Proposed Model of AINI Protocol

Our proposed model AINI routing protocol provides distributed cluster-formation of sensor nodes to enable dynamic self-organized management among sensor nodes based on probabilistic cluster-head selection. Proposed model AINI designs unique distributed cluster-formation in which initially sensor nodes select themselves as Parent Sensor Nodes (PSNs) and later on these PSNs develop unique collaborations among neighbour PSNs to form tripling bonding among every three closer PSNs. These tripling-based bonded PSNs are defined as Child Sensor Nodes (CSNs). Major purpose of formation of CSNs bonding is to introduce higher level assistance among CSNs which result into creation of single Leading Sensor Node (LSNs) which will transmit the information towards selected CHs on behalf of its bonded CSNs. This unique two layer distributed neighbour-formation of PSNs and CSNs offer reasonable energy efficiency by assistance of LSNs, which is almost three times more than some conventional distributed routing protocols.

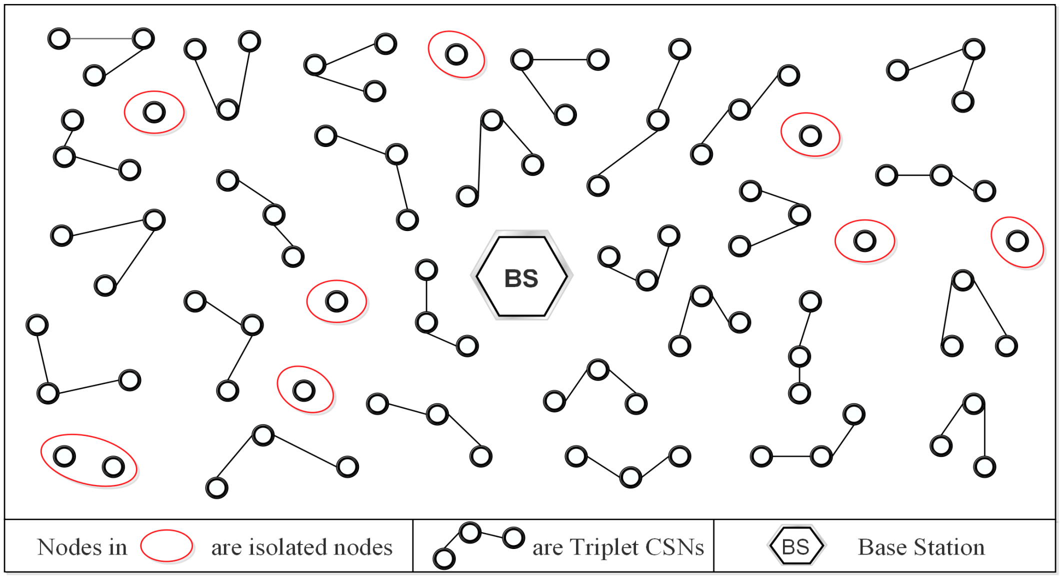


Fig. 2. Network Topology of AINI

During each round of network, all CSNs switch between ‘Sleep’ and ‘active’ modes. After forming of CSNs, one LSN will act as active node and becomes responsible for data transmissions towards CHs. During this transmission period of active nodes, transceiver of other nodes will be in sleep-mode, this sleep-mode will stop communication with their CHs and nodes only sense the environment. Only one node will be active in each CSNs for one round, other two nodes will be in sleep-mode. In this way, nodes will rotate the transmission responsibility and will serve after consecutive two rounds and will save its residual energy considerably. These two nodes in sleep-mode will save additional energy by avoiding idle listening. Nodes which are not in any CSNs will be active and communicate continuously with CHs in all rounds until their energy depleted. Networking operation of proposed protocol is divided into R_n rounds and each round consists of four phases: Initialization phase, Network setup phase, Transmission phase and Termination phase.

AINI protocol is mainly proposed to generate an energy efficient distributed clustering routing protocol which improves highly demanded QoS, reliability, better network lifetime and network stability by minimizing the overall energy consumption. Furthermore, multi-hop routing minimizes the packet drop ratio and retransmission of packet. AINI enforces adaptive CH selection which is accountable to update nodes-mode.Figure 3 shows the four major network operations in flowchart of proposed routing protocol. The main communication steps of AINI protocol are following:

A. Cluster Head Selection

Proposed routing protocol is completely distributed and nodes help themselves to become CHs. Selection process require iterations to run which should be long enough to receive message from the nodes in the range of cluster. Initially nodes are deployed in heterogeneous mode with different energy levels. As in [17], [18], [19], [20], [21], [24], [26], [25] every node is forced to become a CH because of rotating

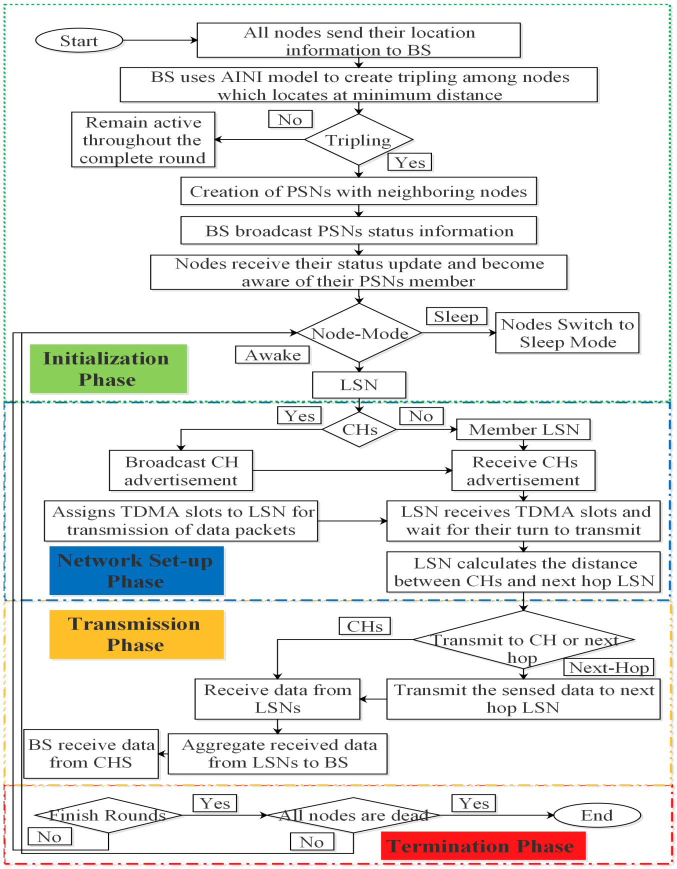

Start
All nodes send their location information to BS
BS uses AINI model to create tripling among nodes which locates at minimum distance
Remain active throughout the complete round
No
Tripling
Yes
Creation of PSNs with neighboring nodes
BS broadcast PSNs status information
Nodes receive their status update and become aware of their PSNs member
Sleep
Node-Mode
Nodes Switch to Sleep Mode
Initialization Phase
Awake
LSN
Yes
CHs
No
Member LSN
Broadcast CH advertisement
Receive CHs advertisement
Assigns TDMA slots to LSN for transmission of data packets
LSN receives TDMA slots and wait for their turn to transmit
Network Set-up Phase
LSN calculates the distance between CHs and next hop LSN
Transmission Phase
CHs
Transmit to CH or next hop
Next-Hop
Receive data from LSNs
Transmit the sensed data to next hop LSN
BS receive data from CHS
Aggregate received data from LSNs to BS
Finish Rounds
Yes
All nodes are dead
Yes
End
No
No
Termination Phase

Fig. 3. Network Operations and Flowchart of AINI Routing Protocol

epoch. If the value of epoch is same for few low energy level nodes and high energy level nodes, they both have equal chances to become a CH. Decision based on preferred ratio of being a CHs per round CH_{pr}. Above mentioned protocols allows each node to become a CH after every $1/CH_{pr}$ round. As the nodes are consuming energy in every round, so the energy level cannot be the same after first round [27]. After first round CH selection of proposed protocol is based on residual energy of nodes. Active nodes for the first round will participate in CH selection process, and higher energy nodes will be selected as CHs. In order to select CHs, threshold is calculated with the following equation:

$$T = \begin{cases} \frac{CH_{pr}}{1-CH_{pr}\times(r\times\frac{1}{CH_{pr}})\times d} & \text{if } n \in \alpha \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Where α shows the set of active-nodes in first round which are appropriate for CH, r is the round and d is distance to BS . After calculating threshold, random number is generated at each node. If random number of node is $<$ the value of threshold, this node will be selected as a CH. This CH broadcast announcement to whole network, only active-nodes will be able to receive this announcement. When a node chooses CH in its cluster range, this node will transmit a message using CSMA MAC protocol to evade collisions. This message contains the location and energy level of node. After receiving the message from nodes, CH will calculate the energy level and distance of each node to BS and choose the active nodes for next round and labelled them as AN_{nr} in a table and broadcast this table to all active nodes in CCs. Furthermore, CH will assign TDMA slots to each active-node for communication and each node will transmit the sensed data in its assigned TDMA slot. Algorithm 1 shows the precise CH selection of proposed AINI protocol.

```
Input  : Leading Sensor Nodes (LSNs)
Output: Cluster Heads (CHs)
Repeat;
Initialization;
while Threshold is calculated on each node do
    Random number is generated on each node;
end
if Random number > threshold value then
    Sustain as LSN;
else
    Select as CH;
end
```

Algorithm 1: Cluster Heads Selection Algorithm

B. Multi-Hop Routing

Our proposed model develop unique two-layer multi-hop transmission phase in which both intra-cluster and inter-cluster communication adopt multi-hoping communication among sensor nodes to BS. The communication system utilizes first order radio model and furthermore nodes have capability to switch to multi-level transmission levels. If nodes have to participate in just intra-cluster communications then these nodes switch their transmission level to lowest one to save the energy resources which are being wasted during amplifications cost of transmissions. Similarly if nodes share the responsibilities of inter-cluster communications then these nodes switch to maximum transmission level to accommodate highest range of inter-cluster communications. Basic assumptions of first order radio model indicates the communications skills of sensor nodes which utmost can offer coverage of whole network region to small network area.

1) Intra-Cluster Multi-hop Routing: Each node in the network is able to determine its location, own distance from CH and neighbour node closer to CH. If Node m wants to transmit data packet, it will take the shortest path to transmit data to CH and creates the first route. Using the multi-hop routing, AINI protocol transmits routes with data in the overhead of packet, which improve the network lifetime and reduce the energy dissipation in hotspot area. Routes have adequate choice of pledge to fully disjoint multiple routes in condition of utilizing opportunistic routing. Trigonometric ratio and law of cosine is used to calculate the route width R_w with transmitting range r and distance d. By calculating the route width opportunistic routing can be utilized between the nodes in the range. Pseudo-code of intra-cluster multi-hop routing is given in Algorithm 2. In order to transmit data through intra-cluster multi-hoping following number of hops are required:

$$N_{h_a^b} = \left[\frac{a\times r_{max}}{b\times r}\right] \tag{2}$$

where, $N_{h_a^b}$ shows the require number hops to transmit data from source node a towards the next-hop node b and r_{max} is the maximum range of node.

2) Inter-Cluster Multi-Hop: For small networks inter-cluster multi-hop routing may not be useful because data travels among the CHs which often consume more energy than direct transmission. But for the large scale network with the large amount of sensor nodes requires more clusters and CHs in the network, these CHs transmit data to BS with significant amount of energy [28]. In order to minimize the energy consumption in long range transmission, we use inter-cluster multi-hop routing in our proposed protocol. Assuming that all CHs have the same range R of transmission which helps CHs to communicate with each other. CHs located near to BS (their range covered by the BS) have the number of hops $N_{h_a^b}$ = 0, while CHs locates away from BS calculate the distance of neighbouring CHs which is in the direction of BS. First hop of CH will be labelled as $N_{h_a^b}$ = 1 and the next hop will add its hop distance in $N_{h_a^b}$ and pass it to other neighbour CH. Algorithm keeps choosing the hop distance until all CHs are connected to each other. After calculating the hop distance CHs will transmit data to the next hop CH, next hop CH transmits this data with its own data to the next hop CH, in this way data keep moving among CHs until it reached to BS. Figure 4 and 5 shows the multi-hop routing with respect to rounds.

C. Data Transmission

To transmit data, proposed model adopts the lossless step-by-step multi-hop transmission [29]. In this transmission model , the transmission of i bit message with node n_i is executed successively. Such as; node transmit the received data with its own sensed data after arrival. δ denotes the non-transmitted data (or source data) packet of node n_i; $\varepsilon(n_i, n_j)$ denotes as intermediate transmission result of node n_i and node n_j. ψ denotes the final transmission of node n_i of all received data and its own data. When node n_i receive data ψ from node n_j , node n_i transmit ψ_j with its own data (may be the source

```
Input  : Location of Source Node ($S_n$), Number of
         Neighbour Nodes ($N_n$), List of Neighbour,
         Distance from $S_n$ to distance $d$, Packet
         Transmission Ratio $PTR$, Transmission
         Range ($T_R$)
Output: Find the shortest route and deliver data to
         CHs
Compute the number of total nodes;
Select $n$ number of nodes;
for each node $m$ do
    Calculate the node range;
    $R_n = d \times \sqrt{1 - (2d^2 - r^2/2 \times d^2)^2}$
    $R_w = 2 \times R_n$
    Node $o$ is a neighbour of node $m$ with longest
      distance from $S_n$;
    Node $m$ forward its sensed data to node $o$;
    Random hop $R_h$ from $\{0, r_{max}/r\}_0 = 0$;
    while $R_{h_0} < R_h$ do
        Farthest node from node $o$ is $o_i$ at
          leftmost(rightmost) routes and a distance d
          from BS is a same number of hops as node $o$;
        Node $o$ transmits packet to node $o_i$;
        $R_{h_o} = R_{h_o} + 1$ , and $o = o_i$;
    end
end
if node $m$ is closest to CH then
    Transmit packet directly to CH and create major
      route;
    Node $q$ is selected as random node $R_n$ from
      major route in non-hotspot area;
    $N_{h_a^b}$ hops routed from node $q$ towards left route;
    $N_{h_a^b}$ hops routed from node $q$ towards right route;
else
    Before convention of same-hop route, transmits
      data packets to CH;
end
All data packets are reached at CHs;
CHs aggregate all data packets to BS;
```

Algorithm 2: Multi-Hoping Pseudo-code of AINI

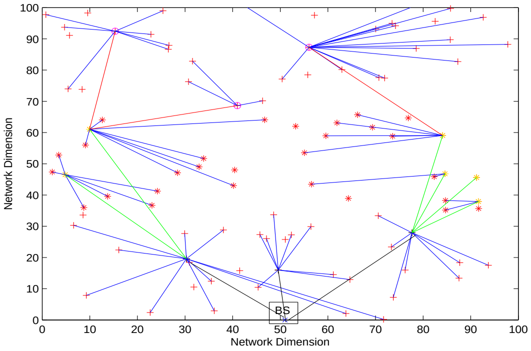


Fig. 4. Multi-hop Routing in Round (R)

data δ or intermediate data ε). If node n_i transmitting data packet ε and if data received from n_j then $\psi_j = \varepsilon_j$, data to be transmitted are both source data, then the transmission formula will be:

$$\varepsilon(n_i, n_j) = max(\delta_i, \delta_j) + (1 - c)min(\delta_i, \delta_j) \quad (3)$$

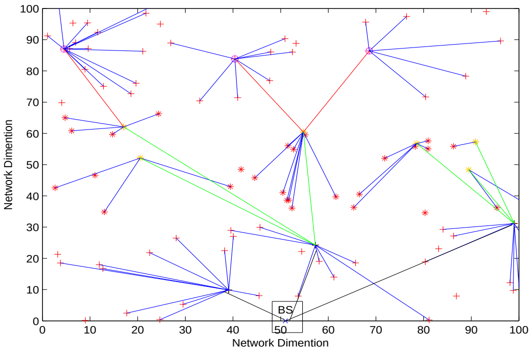


Fig. 5. Multi-hop Routing in Round (R++)

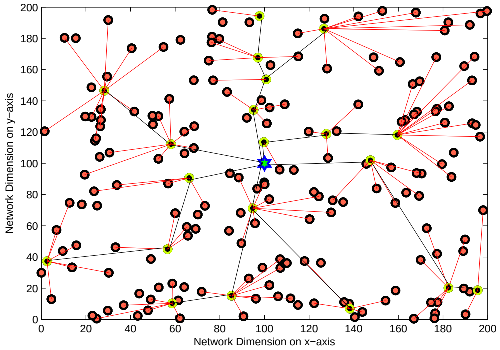


Fig. 6. Data Transmission between Intermediate Nodes

If transmitted data is not from source node, then intermediate transmission formula will be:

$$\varepsilon(\varepsilon_i, \psi_j) = max(\varepsilon_i, \psi_j) + \epsilon(1 - c)min(\varepsilon_i, \psi_j) \quad (4)$$

ϵ is called forgetting factor. Its a decimal between (0, 1), and ε_i, ψ_j are the intermediate and final transmission of active nodes respectively. Figure 6 shows the data transmission between intermediate nodes.

D. Energy Dissipation

Active-mode nodes transmits the sensed data to CH using intra-cluster multi-hop routing in their allocated TDMA slots and sleep-mode nodes remain sleep until the next round. CHs received data from nodes and transmit it to BS. Depending on the distance between transceiver and receiver, we use the first order radio model and free space model in our proposed protocol [30]. We use this model for better comparison with existing protocols [17], [18], [19], [20], [21]. Parameter used by radio models are $E_{elec} = 50nJ/bit, E_{fs} = 10pJ/bit/m^2 and E_{amp} = 0.0013pJ/bit/m^4$. To transmit a i bit message to distance d, the radio expands:

$$E_{Tx}(i, d) = \begin{cases} iE_{elec} + E_{fs}id^2, & \text{if } d < do. \\ iE_{elec} + E_{amp}id^4, & \text{if } d \geq do. \end{cases} \quad (5)$$

Using the radio models, noticeable energy is saved in data transmission. If there are N number of nodes in the network, and N_{CH} is the best possible number of CH, to calculate the average number of nodes in each cluster can be calculated by following equation:

$$\left(\frac{N}{N_{CH}} - 1\right) \tag{6}$$

Transmission of Non-CH node, radio dissipates E_{tx} and E_{amp} for transmitting to attain Signal-to-Noise ratio (SNR). Therefore, transmission of i bit data packet by non-CH node will be:

$$E_{n_{CH}} = \left(\frac{N}{N_{CH}} - 1\right)\left(E_{tx} \times i \times E_{amp} \times i \times d^2\right) \tag{7}$$

$E_{n_{CH}}$ represents non-CH nodes, E_{tx} and E_{amp} is transmitting energy and amplifier of nodes respectively, while d is a distance of node towards BS. Receiving data by CHs from $N_{n_{CH}}$ nodes, the equation will be:

$$E_{rcv} = \left(E_{rx} \times i\right)\left(\frac{N}{N_{CH}} - 1\right) \tag{8}$$

E_{rcv} is the total received data at any CH, and E_{rx} is the total energy dissipated by CH while receiving this data. In order to aggregate this data towards BS, energy dissipates by CH will be:

$$E_{T_{BS}} = \left(E_{data} \times i\right)\left(\frac{N}{N_{CH}}\right) \tag{9}$$

$E_{T_{BS}}$ is the transmission data towards BS, and E_{data} is the total data received by any CH. In order to calculate the total dissipated energy by any CH to transmit this data, the equation will be:

$$E_T = \left(E_{tx} \times i_{AD} \times E_{amp} \times d^2\right) \tag{10}$$

E_T is the dissipated transmission energy of any CH to transmit the total data i_{AD} received from all associated nodes. To calculate the total dissipated energy in one round, we use the following equation:

$$E_{CH} = E_{T_{BS}} + E_T + E_{rcv} \tag{11}$$

Total dissipated energy of CHs in one round is the sum of $E_{T_{BS}}$ data transmission towards BS, E_T dissipated energy of CH while transmission and E_{rcv} dissipated energy of CH while receiving.

E. Node-Mode Decision

As mentioned above, CHs generate a table of remaining energy and D_{BS} of nodes and broadcast it to all CSNs. Nodes decides at the end of round weather to be in sleep-mode or in active-mode after reviewing this table. If a node is elected as active-node in next round AN_{nr} it will turn ON its transceiver in the next round otherwise it will remain OFF until its get elected. CSNs nodes participating in tripling procedure turn OFF their transceiver in consecutive two rounds when these nodes are not active as LSN node. Furthermore, their sleep-mode continue until they elected again as LSN to start their active-mode. Some nodes face remote dispersion due to initial random deployment, so these nodes can miss the initial triplet bonding, thus these nodes remain in active-mode during the whole network operation. Algorithm 3 shows the node-mode decision whether to be in sleep-mode or in active-mode.

```
Input  : Nodes
Output: Either active-mode or sleep-mode
New Round Starts;
if node == triplet CSN then
    if node-mode == active && $AN_{nr}$-flag == 1
     then
        node-mode == active ;
        if node-mode == active && $AN_{nr}$-flag == 0
         then
            node-mode == sleep ;
    else if node-mode == sleep && $AN_{nr}$-flag == 0
     then
        node-mode == sleep ;
        else if node-mode == sleep && other two
         nodes $AN_{nr}$-flag == 0 then
            node-mode == active ;
        end
    end
else if node == triplet CSN && one node == dead
 then
    check status of couple node ;
    if couple node == active then
        node-mode == active ;
    else if couple node == dead then
        node-mode == active ;
    end
end
```

Algorithm 3: Node-Mode Decision Algorithm

Initially nodes will check either they are selected for triplet CSNs or not, if node is selected for CSNs than it will check its flag in AN_{nr} table. If flag is ON in AN_{nr} table, this node will sustain as LSN. Similarly, all nodes check their status in the table, nodes which are in active-mode but their flag is OFF, they will turn OFF their transceiver and receiver and will remain in sleep-mode until their flag turn ON. CSNs will check the status of their triplet nodes in AN_{nr}. If node in sleep mode, it will check the flag of other two nodes, if their flags are OFF, it will turn itself into active-mode, if not it will remain OFF until next round. When a node in sleep mode, it will continuously check the status of their triplet CSNs. If a node in CSNs are in sleep mode or dead, this node will turn itself into active-mode. If one node is dead in triplet CSNs, then the decision will remain between coupled nodes. Both nodes check their flag status, one will be in active-mode and the other one in sleep-mode. If two nodes are dead in the triplet CSNs, last node will remain active throughout its lifetime. Nodes which are not selected for any triplet CSNs, will be in active mode throughout the network.

V. Experiments and Simulation Results

This paper introduces a neighbour-cooperation heterogeneity-aware traffic engineering for WSNs with the concept of tripling the sensor nodes. Following subsection explains the performance measures, methodology and simulation results.

A. Performance Measures

To analyze the performance of proposed protocol, we use the following metrics: leftmargin=*,labelsep=5.8mm

- Network Stability (NS): Total time from start of network operation to the demise of first node.
- Network Instability (NI): Total time form demise of first node to the demise of last node.
- Network Lifetime (NL): Total time from start of network operation to the demise of last node.
- Energy Consumption Ratio (ECR): Total energy consume during the transmission and receiving of data packet.
- Death of First Node (DFN): Network operation from the start of network till the death of first node. It also measures the stability period of routing protocol.
- Death of Last Node (DLN): Network operation from the start of network till the death of last node, it helps to measure the instability period of routing protocol.

B. Experiments Methodology

To evaluate the performance of AINI, we have done simulations in MATLAB environment. Similar to other environments, such as OMNET ++ and ns-2 with the difference that MATLAB provides feature architecture and allows rapid simulations by combining multiple components. To provide the better presentation of experimental results we use OriginLab for deep analysis in few results. Multiple scenarios and various parameters are considered to evaluate the performance.

C. Comparison with State-of-the-Art Algorithms

The proposed AINI routing protocol is compared with state-of-the art routing protocols such as LEACH [17], LEACH-C [18], SEP [19], ESEP [20] and DEEC [21]. First we validate the proposed analysis model for AINI routing protocol then compare the proposed protocol by executing it in different network dimensions. To prove the performance of proposed protocol in large scale network, we took the simulations in two different scenarios. Specifically, simulation scenarios include 100 and 200 sensor nodes which are deployed in $100m \times 100m$ and $200m \times 200m$ network area with initial network energy $1J, 1.5J$ respectively, while the BS is located in the center of the network. To remove the error caused by randomness each simulation runs at least for five times and average is considered as a final result. Major simulation parameters are given in Table I.

TABLE I. SIMULATION PARAMETERS

$Sensor\ Nodes\ Scenario\ 1$	100
$Sensor\ Nodes\ Scenario\ 2$	200
$Network\ dimensions\ Scenario\ 1$	$100m \times 100m$
$Network\ dimensions\ Scenario\ 2$	$200m \times 200m$
$Heterogeneous\ Energy\ Level\ Scenario\ 1$	$1J$
$Heterogeneous\ Energy\ Level\ Scenario\ 2$	$1.5J$
$Required\ CHs\ Per\ Round$	10%
$Transmission\ Energy\ Dissipation$	$50pJ/bitj$
$Data\ Packet\ Size$	$4000\ bit$
$Transmission\ Energy$	$50nJ/bit$
$Receiver\ Energy$	$50nJ/bit$
$Amplifier\ Transmission\ Energy\ Dissipation$	$100pJ/bit/m2$

D. Experiments

It is noticed that, tripling of nodes extends the network lifetime and stability of network. Moreover, multi-hoping improves the lifetime by distributing the load of one CH among multiple CHs. In the simulation results of first scenario, first node dies in proposed protocol after 2432 which shows the stability period is much better than LEACH, LEACH-C, SEP, ESEP and DEEC. Moreover, last node dies in proposed protocol at 5978 rounds, which shows the instability period of proposed protocol. It can be seen clearly in Figure 7 and 8 that proposed protocol produce better performance then state-of-the art routing protocols.

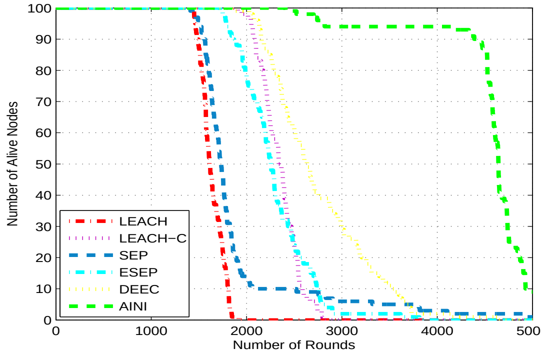


Fig. 7. Network Lifetime Respectively in Network of 100 × 100 with 100 Nodes

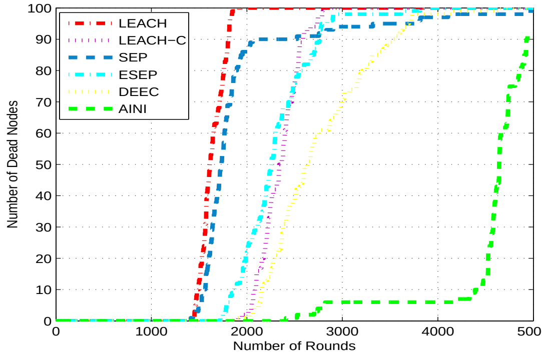


Fig. 8. Network Stability in Network of 100 × 100 with 100 Nodes

Figure 9 and 10 show the number of data packets transmitted towards CH and BS respectively. From these figures, it is examined that the multi-hop increases the number of sent packets to BS, because inter-cluster multi-hoping helps the CHs to transmit data to BS with minimum energy dissipation. Undoubtedly, proposed protocol improves the network lifetime and stability of network. Furthermore, it improves the throughput and reduces the energy dissipation throughout the network. This means, that proposed protocol is more energy efficient than previous well known existing protocols, because propose protocol allows the nodes to work with full functionality by improving the packet delivery ratio and reducing the delay in transmission.

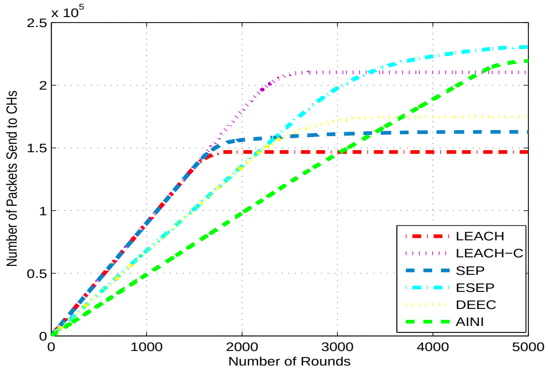


Fig. 9. Successfully Delivered Packets to CHs in Network of 100 × 100 with 100 Nodes

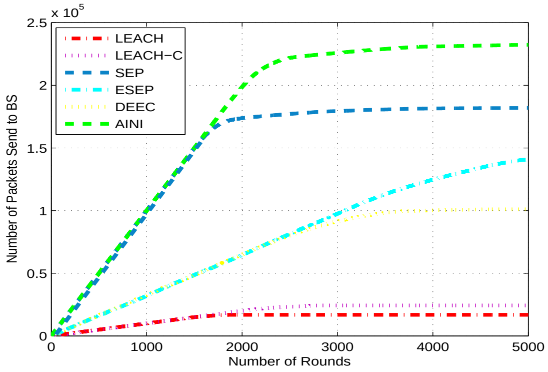


Fig. 10. Successfully Delivered Packets to BS in Network of 100 × 100 with 100 Nodes

Figure 11 and 12 shows the performance of proposed protocol for second scenario, in terms of network lifetime and network stability of proposed protocol with the comparison of other protocols. As per the results, our proposed protocol AINI outperforms in large scale heterogeneous network. The result presented in Figure 11 and 12 also demonstrate the instability period of proposed protocol at large scale which proves that AINI protocol have the minimum instability period.

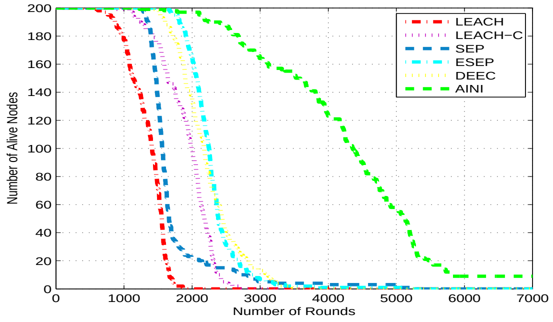


Fig. 11. Network Lifetime in Network of 200 × 200 with 200 Nodes

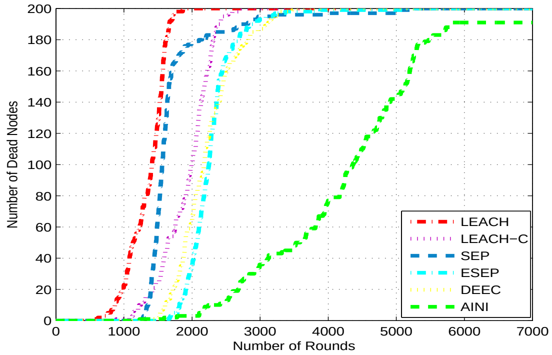


Fig. 12. Network Stability in Network of 200 × 200 with 200 Nodes

Figure 13 and 14 shows the comparison of overall network energy consumption with respect to rounds. It is noticed that proposed routing protocol is utilizing minimum energy in both graphs. First graph shows the network energy consumption of scenario 1 while second graph shows the energy consumption of scenario 2. It can be seen clearly in second graph that all other protocol are dead after 5500 rounds but proposed protocol is still surviving in terms of energy (joule).

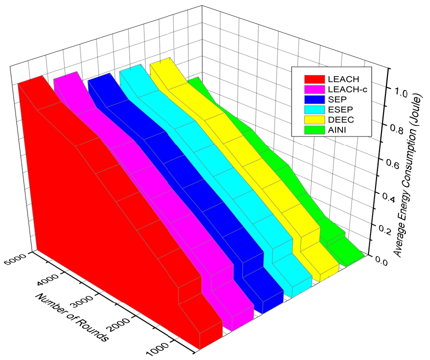


Fig. 13. Average Energy Consumption in Network of $100m \times 100m$

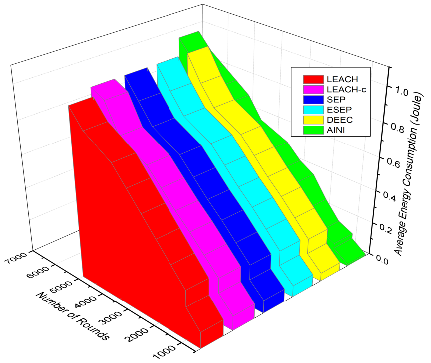


Fig. 14. Average Energy Consumption in Network of $200m \times 200m$

Figure 15 shows the comparative result of proposed protocol with other protocols in terms of heterogeneity-aware. Figure shows the impact of multiple heterogeneity where multiple network scenarios are considered and shows that proposed routing protocol obtains higher stability and longer network lifetime in larger network.

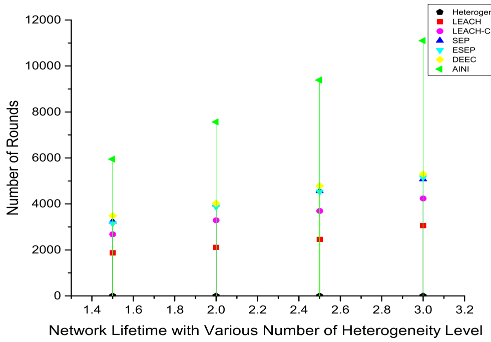


Fig. 15. Impact of Different Heterogeneity Level

VI. Conclusions and Future Work

This paper introduces a new technique Neighbour-Cooperation Heterogeneity-Aware Traffic Engineering for WSNs and proposed a new clustering routing protocol called Adaptive Inter-networking Improved routing protocol (AINI) to improve the stability, network lifetime, CH selection and average energy consumption. Furthermore, this protocol uses multi-hop (intra-cluster communication for transmission between nodes and CHs and inter-cluster communication for transmission between CHs and BS) routing which enhanced the reliability of network. Simulation results proved that proposed protocol is more energy efficient and perform far better than well known existing routing protocols such as, LEACH, LEACH-C, SEP, ESEP and DEEC. Mobility based AINI routing protocol is considered as future work.

Statement of conflict

Authors of this paper: Christopher Mumpe, Da Tang, Muhammad Asad, Muhammad Aslam, Jing Chen, Jinsi Zhu, Luyuan Jin declares that there is no conflict of interest regarding the publication of this research article entitled "Neighbour-Cooperation Heterogeneity-Aware Traffic Engineering for Wireless Sensor Networks".

Notations

The following notations are used in this paper.

CSN	Child Sensor Node
LSN	Leading Sensor Node
PSN	Parent Sensor Node
CH_{pr}	Cluster Heads Per Round
α	Set of Active-Nodes in First Round
AN_{nr}	Active-Nodes for Next Round
R_w	Route Width
R_n	Node Range
R_h	Random Hop
S_n	Source Node
$N_{h_a^b}$	Number Of Hops
δ	Non-Transmitted Data
ε	Intermediate Transmission Result Between Two Nodes
ψ	Final Transmission
ϵ	Forgetting Factor
D_{BS}	Distance to Base Station

References

[1] K. Sohraby, D. Minoli, and T. Znati, *Wireless sensor networks: technology, protocols, and applications*. John Wiley & Sons, 2007.

[2] I. M. El Emary and S. Ramakrishnan, *Wireless sensor networks: from theory to applications*. CRC Press, 2013.

[3] S. Tyagi and N. Kumar, "A systematic review on clustering and routing techniques based upon leach protocol for wireless sensor networks," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 623–645, 2013.

[4] Y. Peng, F. Al-Hazemi, R. Boutaba, F. Tong, I.-S. Hwang, and C.-H. Youn, "Enhancing energy efficiency via cooperative mimo in wireless sensor networks: State of the art and future research directions," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 47–53, 2017.

[5] N. A. Pantazis, S. A. Nikolidakis, and D. D. Vergados, "Energy-efficient routing protocols in wireless sensor networks: A survey," *IEEE Communications surveys & tutorials*, vol. 15, no. 2, pp. 551–591, 2013.

[6] N. A. O. Al-Humidi and G. V. Chowdhary, "Comparative analysis of clustering algorithms for routing protocols in wireless sensor networks," 2017.

[7] M. I. Chidean, E. Morgado, E. del Arco, J. Ramiro-Bargueno, and A. J. Caamaño, "Scalable data-coupled clustering for large scale wsn," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4681–4694, 2015.

[8] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267–279, 2014.

[9] P. Neamatollahi, M. Naghibzadeh, S. Abrishami, and M.-H. Yaghmaee, "Distributed clustering-task scheduling for wireless sensor networks using dynamic hyper round policy," *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 334–347, 2018.

[10] M. C. M. Thein and T. Thein, "An energy efficient cluster-head selection for wireless sensor networks," in *Intelligent systems, modelling and simulation (ISMS), 2010 international conference on*. IEEE, 2010, pp. 287–291.

[11] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Computer communications*, vol. 30, no. 14-15, pp. 2826–2841, 2007.

[12] W.-P. Chen, J. C. Hou, and L. Sha, "Dynamic clustering for acoustic target tracking in wireless sensor networks," *IEEE transactions on mobile computing*, vol. 3, no. 3, pp. 258–271, 2004.

[13] A. Wang, D. Yang, and D. Sun, "A clustering algorithm based on energy information and cluster heads expectation for wireless sensor networks," *Computers & Electrical Engineering*, vol. 38, no. 3, pp. 662–671, 2012.

[14] M. Pramanick, P. Basak, C. Chowdhury, and S. Neogy, "Analysis of energy efficient wireless sensor networks routing schemes," in *Emerging Applications of Information Technology (EAIT), 2014 Fourth International Conference of*. IEEE, 2014, pp. 379–384.

[15] B. Guo and Z. Li, "A dynamic-clustering reactive routing algorithm for wireless sensor networks," *Wireless Networks*, vol. 15, no. 4, pp. 423–430, 2009.

[16] J.-J. Liaw, C.-Y. Dai, and Y.-J. Wang, "The steady clustering scheme for heterogeneous wireless sensor networks," in *Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC'09. Symposia and Workshops on*. IEEE, 2009, pp. 336–341.

[17] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *System sciences, 2000. Proceedings of the 33rd annual Hawaii international conference on*. IEEE, 2000, pp. 10–pp.

[18] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on wireless communications*, vol. 1, no. 4, pp. 660–670, 2002.

[19] G. Smaragdakis, I. Matta, and A. Bestavros, "Sep: A stable election protocol for clustered heterogeneous wireless sensor networks," Boston University Computer Science Department, Tech. Rep., 2004.

[20] F. A. Aderohunmu, J. D. Deng *et al.*, "An enhanced stable election protocol (sep) for clustered heterogeneous wsn," *Department of Information Science, University of Otago, New Zealand*, 2009.

[21] L. Qing, Q. Zhu, and M. Wang, "Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks," *Computer communications*, vol. 29, no. 12, pp. 2230–2237, 2006.

[22] A. Roshini and H. Anandakumar, "Hierarchical cost effective leach for heterogeneous wireless sensor networks," in *Advanced Computing and Communication Systems, 2015 International Conference on*. IEEE, 2015, pp. 1–7.

[23] F. Xiangning and S. Yulin, "Improvement on leach protocol of wireless sensor network," in *Sensor Technologies and Applications, 2007. SensorComm 2007. International Conference on*. IEEE, 2007, pp. 260–264.

[24] A. Manjeshwar and D. P. Agrawal, "Teen: a routing protocol for enhanced efficiency in wireless sensor networks," in *null*. IEEE, 2001, p. 30189a.

[25] A. Garg *et al.*, "Distance adaptive threshold sensitive energy efficient sensor network (dapteen) protocol in wsn," in *Signal Processing, Computing and Control (ISPCC), 2015 International Conference on*. IEEE, 2015, pp. 114–119.

[26] A. Manjeshwar and D. P. Agrawal, "Apteen: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks," in *ipdps*. IEEE, 2002, p. 0195b.

[27] V. Gupta and R. Pandey, "An improved energy aware distributed unequal clustering protocol for heterogeneous wireless sensor networks," *Engineering Science and Technology, an International Journal*, vol. 19, no. 2, pp. 1050–1058, 2016.

[28] S. Khurana and K. R. Rekha, "Energy efficient of inter cluster multihop routing protocol for wireless sensor network," 2016.

[29] S. Chowdhuri, S. Chakraborty, N. Dey, S. S. Chaudhuri, and P. Banerjee, "Propagation analysis of mimo ad hoc network in hybrid propagation model and implement less propagation loss algorithm to find the minimum loss route," *International Journal of Information and Communication Technology*, vol. 10, no. 1, pp. 66–80, 2017.

[30] J. Banerjee, S. K. Mitra, and M. K. Naskar, "Comparative study of radio models for data gathering in wireless sensor network," *International Journal of Computer Applications*, vol. 27, no. 4, 2011.

A Proposed Model of Cloud based e-Learning for Najran University

Ibrahim Abdulrab Ahmed
Department of Information System, Community College
Najran University, KSA

Zakir Hussain
Department of Information System, Community College
Najran University, KSA

***Abstract*—For the time being, the educational institutions are keen to use e-learning in their educational environment. This, in turn, will support their learning process and allow the learners to access any service or learning material or information at any time they need it. With all the pros by the e-learning, it still suffers from many problems that are explained clearly in this paper. In contrast, along with the innovation of cloud computing technology as a new paradigm in the IT world. With the establishment of cloud computing, numerous services for numerous fields (e.g., education, business, and government) have been introduced that have greatly facilitated the e-learning. In this paper, it demonstrates how the inclusion of the cloud computing paradigm in the e-learning environment assist positively. A lot of obstacles that are introduced by e-learning have been remedied. It combines the cloud computing in the e-learning system, thus, the proposed E-learning Embracing Cloud Computing Model (ELECCM) has completely developed and performed with all the essential components that are needed for their architecture. The study presents all the procedures that are run in order by the proposed system. Then, a fully functional e-learning system based on cloud computing; with low cost and low technical barriers, is demonstrated and explained clearly.**

Keywords*—*E-learning; cloud computing; E-learning Embracing Cloud Computing Model (ELECCM); SaaS; PaaS; IaaS

I. Introduction

In the digital world, in which the new technologies are emerging rapidly and radically, the innovative e-learning methods must be facilitated to allow to transfer the more effective knowledge and to enable to participate for lifelong. The fact that traditional e-learning methods are inadequate to meet the needs of advanced e-learning processes, especially in higher education. Higher education highlights the experiences and outcomes of higher learning that require a major transformation in the knowledge and communication community [1]. E-learning is the way to deliver an instruction electronically partially or solely through a web browser, such as Netscape Navigator, via the Internet or multimedia platforms (CD-ROM or DVD) [2]. Due to the high accessibility of network bandwidth, the World Wide Web has been extensively used as a medium for displaying and delivering teaching material, etc. Currently, the events and activities of e-learning are conducted via various electronic media. Educational and technological aspects have become progressively important facets in e-learning content development [3]. The e-module interface design and the usage of interactive multimedia elements are frequently being focused on designing e-learning content [1].

E-learning offers many benefits such as flexibility, diversity, and measurement, although it still faces many difficulties in its implementation [2], [3]. With the experienced, the high initial cost; namely the economic factor, was the main problem they are facing when they are starting e-learning. It is the main concern by the institutions whose interest to implement e-learning; for example, the students in a university with different departments may need to access to a lot of computing resources (hardware and software) [4], [5], [6]. Infrastructure, Maintenance and Human Resources are the main issues that are considered by the initial cost. The access to the learning material is another problem that might occur when implementing e-learning [4]. Any university can develop its own e-learning system. Any conventional e-learning system; Figure 1, may cause a lot of problems. They are such as the time consumed to design the system, the costs needed for infrastructure or for the appointment of professional staff to maintain e-learning and promotion system [6].

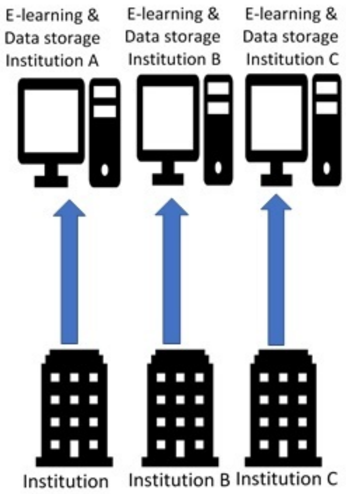


Fig. 1. Conventional E-Learning

In this paper, a proposed e-learning model based on cloud computing services introduced, namely E-learning Embracing Cloud Computing Model (ELECCM). The proposed ELECCM has completely developed and performed with all the essential components that are needed for their architecture. The study presents all the procedures that are run in order by the proposed system.

The rest of the paper is organized as follows. Section II explores an introduction cloud computing. In Section III, the proposed architecture is explained in detail starting by the important procedures, the layers details, and ending by the details of the proposed model of the cloud-based e-Learning. Finally, the paper concludes in Section IV.

II. The Cloud Computing

Cloud computing has been one of the most important advances in information technology in the last decade since then it has become mainstream [7], [8]. Numerous layers of services have been introduced with establishing grid computing, all people from academic fields, business, consumers to governmental officials were enjoying the revolutionary experience of cloud computing. The benefits of cloud computing in which e-learning became easily apparent. First of them, no worries by the developers of e-learning about the balance between hardware and resources, thus reducing their reliance on professional knowledge [9]. Secondly, it is sufficient for the users of e-learning systems to have some lightweight, low-cost devices (e.g., smartphone, and tablet) to access and contact with the e-learning systems via the internet. Finally, all the data used in the e-learning systems (e.g., reports or personal notes) would be easily stored and backup automatically. Thus, the time needed to manage data is reduced and the storage spaces, with improved security [7], [9], [10]. Usually, applying the e-learning on cloud environment will strongly help the educational institutions to utilize the e-learning services which run in the cloud environment; Figure 2. It would be enough for the institution to only rent the cloud computing provider's infrastructure. Similarly, with the maintenance and the human resources for the development stages in which the e-learning environment has been established by the provider of the cloud service. Accordingly, it reduces the costs required by the institutions to implement the e-learning based on cloud computing because it will not be needed by the institutions to pay to buy [4].

Figure 1 and Figure 2 explain the big change in the e-learning era and how to shift from the conventional e-learning to the cloud-based e-learning. The cloud-based approach helps to reduce the cost of implementing e-learning in the educational institution. As shown in Figure 1, the basic elements in the implementation of conventional e-learning are system upgrade, system maintenance and e-learning system development [11]. However, the conventional e-learning is suffering from several problems in terms of scalability flexibility, and accessibility that may affect its performance [12]. The scalability is considered as one of the important features of the cloud e-learning. This feature helps virtualization infrastructure layer provided by the cloud service provider. Virtualization helps solve the problem of the physical barriers that are generally inherent in the lack of resources and infrastructure to automate the management of these resources as if they were a single entity through hypervisor technologies such as virtual machine (VM) [4], [11]. As a result, we believe that the cloud-based e-learning system is considered the best way that combined the education with the information technology fields. Hence, this study report and display the proposed model that emerged by employing the cloud computing technology into e-learning along with the setup procedures [6].

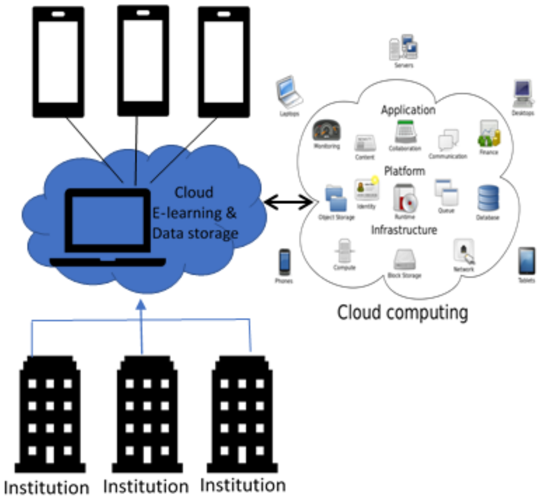


Fig. 2. Cloud-Based E-Learning

III. The Proposed Architecture

A. An Overview

Figure 3 shows the general architecture of the proposed cloud e-learning [3], [13]. The proposed (ELECCM) system components are the cloud partners which connect with a cloud central system through several local servers. In the figure, each PC user has the property of any specific educational institute/university and it works as a cloud partner that provides the needed resources for the cloud system from its available resources. Each local server is also connected with each institute/university. Each local server will monitor the PC user status up to each request from the associated institute/university [13], [14], [15]. In general, the procedure is when requests from the users whom associated with a specific local server are submitted to the cloud through it. Then, the local server collects these requests from the clients in its domain during a certain time period. It verifies as well as forwards the requests. On the other side, different providers with different services for the users are available. These providers should have the agreement with the cloud system [13]. The following will explain each procedure in the proposed system exhaustively.

1) Request Configuration Procedure: As shown in Figure 3 that each user (cloud partner) receives their requests from the cloud sides via the local servers that communicate with. This procedure can be depicted in Figure 4 and the steps are as follows:

1) First, the user does a request. Then, the local server will receive the request that associated with required identification information of the user (e.g. password/ User ID).
2) Each local server does the authentication and verification process for each user. Then, it sends a form with a suitable graphical user interface (GUI). This GUI differs according to the type of the user (e.g. student, lecturer, etc.).
3) The user determines the needed services via the user interfaces. After getting the user requests, the

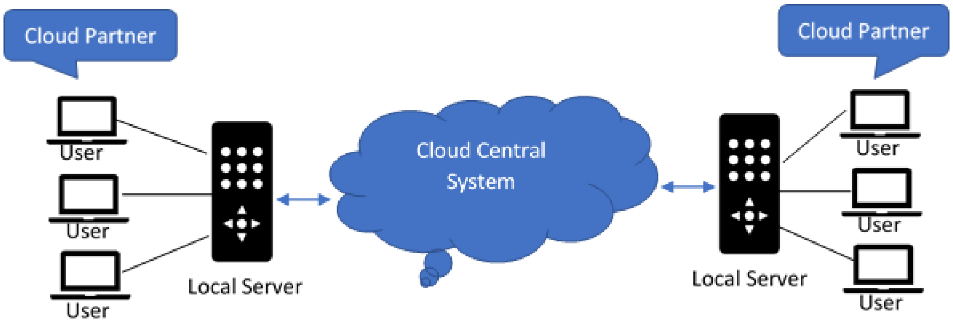


Fig. 3. The General Architecture of the Proposed ELECCM

verification should be done by the local server. This verification includes the instantly available resources, policy to the cloud like pricing policy, encryption system and other data security.

4) If the user has no agreement to receive the requested services or if the pricing policy is mismatched, then the local server directly informs the user for alternatives such as payment through credit card.
5) Otherwise, if the user agrees with the current policy, a reply; an acknowledgement message, will be sent to the local server.
6) Finally, the local server sends the requested resource to users when it receives the cloud system resources.

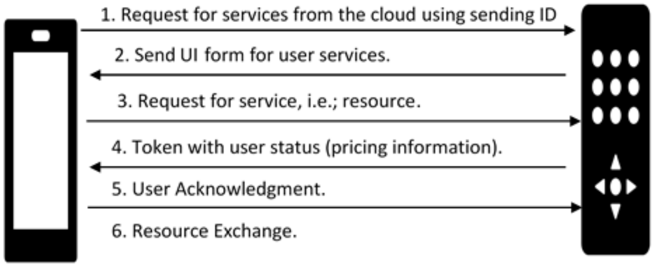

1. Request for services from the cloud using sending ID
2. Send UI form for user services.
3. Request for service, i.e.; resource.
4. Token with user status (pricing information).
5. User Acknowledgment.
6. Resource Exchange.

Fig. 4. The Communication steps between the User and Server

2) Resource Monitoring Procedure: Some resources which are unused, the cloud system sends a periodic message; salutation message, to each institute/ university associated server. This message is sent to search for the status of the respective clients. Each server reproduces various copies of that salutation message and then forwards each message copy to the respective client under its domain. The Resource Information Message from all clients of the server will be received by the server, and then the server will generate a summarized message based on the information that it collects. From the client and send back the message to the cloud system. The resource identification procedure is shown in the figure 5.

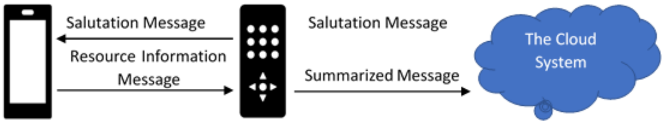


Fig. 5. The Flow Diagram of sharing the Resource Information

3) Resource Allocation Procedure: At a particular time, the server collects the client's requests under its domain. Moreover, summarizes and combines the overall requests based on the individual group of services. Assume, the university server accepts two distinct clients' requests. The first client request with 10GB storage size and a Microsoft office. While the second request for 15 GB storage with two different software such as Antivirus and visual C++. The university server will summarize both requests as 25 GB storage with all software's, i.e., visual C++, Antivirus and Microsoft office. As soon as both requests are received by the cloud system from the server, all the requests (25 GB of storage, Microsoft Office software, one Antivirus software and visual C++ software) are sent to the clients. Generally, the structure of the Cloud Central System consists of two sub-layers; upper and low-sub-layers. The upper sub-layer carries out several processes such as authentication and credit verification before offering any service. Moreover, the upper sub-layer is associated with government central system to control and monitor the cloud system processes. Furthermore, the cloud lower sub-layer can provide more four types of services based on the user requests. These are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) or e-Learning tools. IaaS, PaaS and SaaS represent the layers of any cloud computing. Figure 6 illustrates the architecture of the cloud system.

The Upper Sub-Layer: The security is an important issue in the cloud system. This is because the services in the cloud system are accessed over the internet. Each client can select its own security methods such as the needed encryption process. Furthermore, the cloud system has to agree the all methods with the local server to interpret them. As well as, the users in our educational system are at several levels so the request for services is diverse. The access method will be maintained by identifying the services and user types. The policy among the user and provider will be defined by the sub-layer and will be depended on multiple factors. Examples of these factors are the user level, the latency and the throughput. Based on the policy, different priorities are set by the government for the users. For example, the higher priority users can access the resources with lower latency. The policy also guarantees the provider to run the software smoothly with maximum throughput and highest load balance. Moreover, an authentication and credit verification sub-layer are required in this layer to verify the local server as soon as a request for resources is coming from the server end. It also authenticates and verifies the

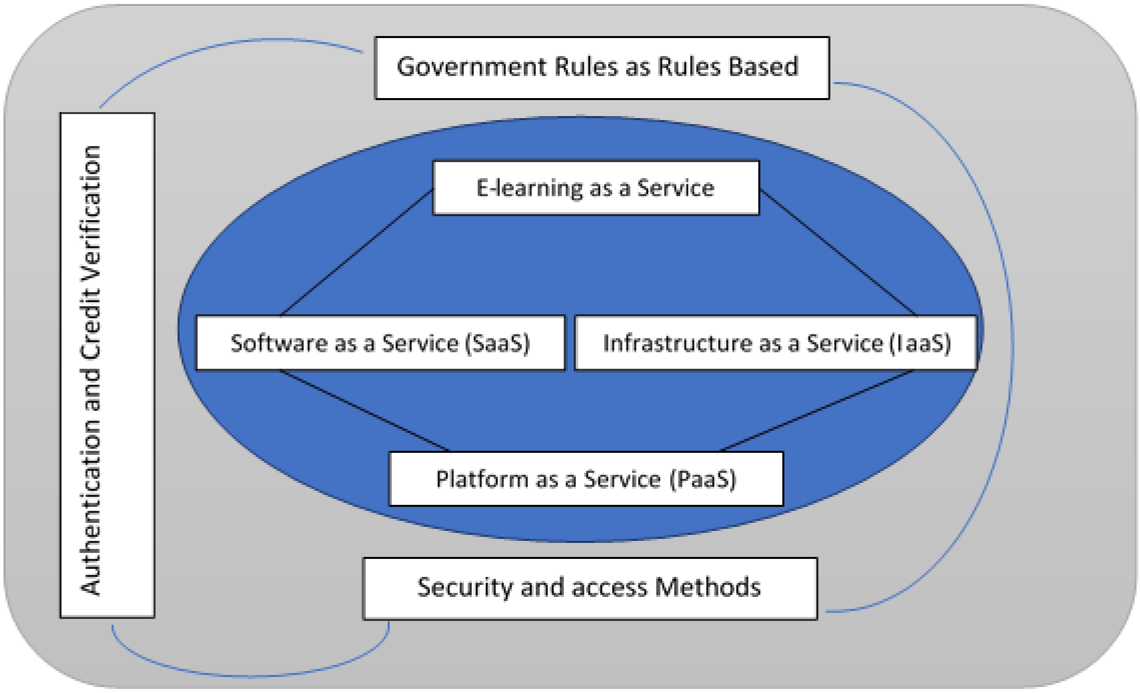

Government Rules as Rules Based
Authentication and Credit Verification
E-learning as a Service
Software as a Service (SaaS)
Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Security and access Methods

Fig. 6. The internal architecture of ELECCM system

user credit information for the requested service; if he has sufficient balances for the requested services it accepts and transform the requests to the lower sub-layer. As soon as the lower sub-layer confirms the request it adjusts the user account after deducting the amount for the requested service. Rules by the Government are set; they named the planning and monitoring committee. For example, the planning committee decides the prices for different types of services based on analysis and agreement with the cloud partners. It also decides the number of funds needed to be allocated to an individual organization. The corruption monitoring committee monitors the daily proceedings of every institute and all objections come from the users' end (e.g. unmatched software).

The Lower Sub-Layer: The lower layer of the cloud architecture allows accessing the private resources that are user request. The lower layer is waiting for the positive acknowledgement that will be sent from the upper layer. Once the lower layer receives the positive acknowledgement, it provides the user requested services. The interaction will be established between the vendors and clients under the responsibility of an instrumental panel in the layer [16]. The layer has an operational panel in which it performs different tasks such as monitors the circumstances, handling the PCs and managing the images. This panel will include a script tool for controlling, monitoring, configuring, and maintaining the clusters. This tool is named as Extreme Administrator Toolkit (xCAT). Each request arrives from the server ends in the form of bare-metal image format is first loaded on xCAT and then process by the virtual cloud system.

B. The Proposed Model of Cloud-Based E-Learning

Actually, the e-learning is just updating the tools, concepts and techniques. The e-learning can't replace the teachers completely. The e-learning will not replace the functions and roles of the teachers. It only provides new content, methods, and concepts for education. Moreover, the teachers will still the backbone of the learning as well as developing and making use of the e-learning based cloud. Recently, many e-learning cloud computing models have been proposed by researchers [17], [18]. According to USA NIST's definition, cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [18]. In this section, the architecture of our proposed ELECCM is presented; in details. The model is consisting of five essential layers, namely as (1) infrastructure layer, (2) platform layer, (3) services layer (4) clients-access layer and (5) user layer. They are depicted in Figure 7. The first layer is the hardware layer. It includes all the hardware, computing and storage capacity for the high-level layer.

The infrastructure layer contains resources and architecture that supporting infrastructure, such as virtual machine, cloud platform. It shares the IT infrastructure resources and connects the system huge system pool together to provide services. The cloud computing enables the hardware and infrastructure layers to work like internet/intranet. Then, the data resources can be accessed in secure as well as scalable way. The third layer is the platform layer; the software resource layer consists of middleware and operating system. Different software resources are integrated by the technology of middleware to develop a unified interface for software developers to develop applications and embed them in the cloud. The operating system in which the e-learning application will be running in this layer. The fourth layer is the service layer; namely SaaS. In SaaS, the cloud computing service is provided to customers. Web Services, Multimedia Applications, Business Applications are examples of the provided services. The client-access layer is the fifth layer of our proposed architecture. The access layer which consists of multi-channel access from multi devices for addressing the access issue to cloud e-learning services which is available on the architecture such as types of access devices and presentation models. The concept of multi-channel access that allows a variety of available services which can be accessed through a variety of devices (e.g., computer, mobile phones, smartphones) and a variety of presentation models (e.g., desktop, mobile applications). The final layer in our proposed ELECCM system is the user layer which consists of the provider, administrator, teacher and student.

IV. Conclusion

In this paper, a complete E-learning Embracing Cloud Computing Model (ELECCM) had been developed with all essential components that are needed. The proposed model considered a lot of the obstacles that may be the e-learning suffered. The study also demonstrated how the including of the cloud computing paradigm in the e-learning environment can be assisted positively. Furthermore, presented all the procedures that are run in order by the proposed system. As a conclusion, a fully functional e-learning system based on cloud computing; with low cost and low technical barriers, is demonstrated and explained clearly.

Acknowledgment

This paper was funded from Scientific Research Deanship at Najran University under No NU/ESCI/14/020.

References

[1] L. W. Y. Kiaw, L. S. Hoe, L. S. Ling, and L. M. Chew, "Proposed object-based e-learning framework embracing cloud computing," in *Proceedings of the International Conference on E-Commerce (ICoEC)*, 2015, pp. 8–13.

[2] H. Al-Samarraie, B. K. Teng, A. I. Alzahrani, and N. Alalwan, "E-learning continuance satisfaction in higher education: a unified perspective from instructors and students," *Studies in Higher Education*, pp. 1–17, 2017.

[3] R. C. Clark and R. E. Mayer, *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. John Wiley & Sons, 2016.

[4] N. Selviandro, M. Suryani, and Z. A. Hasibuan, "Enhancing the implementation of cloud-based open learning with e-learning personalization," in *International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2015, pp. 472–479.

[5] D. Elmatary, S. Abd El Hafeez, W. Awad, and F. Omara, "Sla for e-learning system based on cloud computing," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 10, pp. 189–194, 2015.

[6] N. Selviandro and Z. A. Hasibuan, "Cloud-based e-learning: A proposed model and benefits by using e-learning based on cloud computing for educational institution," in *Information and Communication Technology-EurAsia Conference*. Springer, 2013, pp. 192–201.

[7] U. J. Bora and M. Ahmed, "E-learning using cloud computing," *International Journal of Science and Modern Engineering*, vol. 1, no. 2, pp. 9–12, 2013.

[8] B. Mazhar, R. Jalil, J. Khalid, M. Amir, S. Ali, and B. H. Malik, "Comparison of task scheduling algorithms in cloud environment," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 384–390, 2018.

[9] M. T. Baldassarre, D. Caivano, G. Dimauro, E. Gentile, and G. Visaggio, "Cloud computing for education: A systematic mapping study," *IEEE Transactions on Education*, no. 99, pp. 1–11, 2018.

[10] S. Aldossary and W. Allen, "Data security, privacy, availability and integrity in cloud computing: issues and current solutions," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, pp. 485–498, 2016.

[11] M. Paul and A. Das, "A computational intelligent learning architecture in cloud environment," in *Intelligent Engineering Informatics*. Springer, 2018, pp. 523–531.

[12] P. Neelakantan, "A study on e-learning and cloud computing," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 1534–1539, 2018.

[13] S. Al Noor, G. Mustafa, S. A. Chowdhury, M. Z. Hossain, and F. T. Jaigirdar, "A proposed architecture of cloud computing for education system in bangladesh and the impact on current education system," *IJCSNS International Journal of Computer Science and Network Security*, vol. 10, no. 10, pp. 7–13, 2010.

[14] Ş. A. Rădulescu, "A perspective on e-learning and cloud computing," *Procedia-Social and Behavioral Sciences*, vol. 141, pp. 1084–1088, 2014.

[15] A. Elgelany and W. G. Alghabban, "Cloud computing: Empirical studies in higher education," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 121–127, 2017.

[16] M. Phankokkruad, "Implement of cloud computing for e-learning system," in *International Conference on Computer & Information Science (ICCIS)*, vol. 1. IEEE, 2012, pp. 7–11.

[17] C.-C. Wang, W.-C. Pai, and N. Y. Yen, "A sharable e-learning platform based on cloud computing," in *3rd International Conference on Computer Research and Development (ICCRD)*, vol. 2. IEEE, 2011, pp. 1–5.

[18] P. Mell, T. Grance *et al.*, "The nist definition of cloud computing," *National institute of standards and technology*, vol. 53, no. 6, p. 50, 2009.

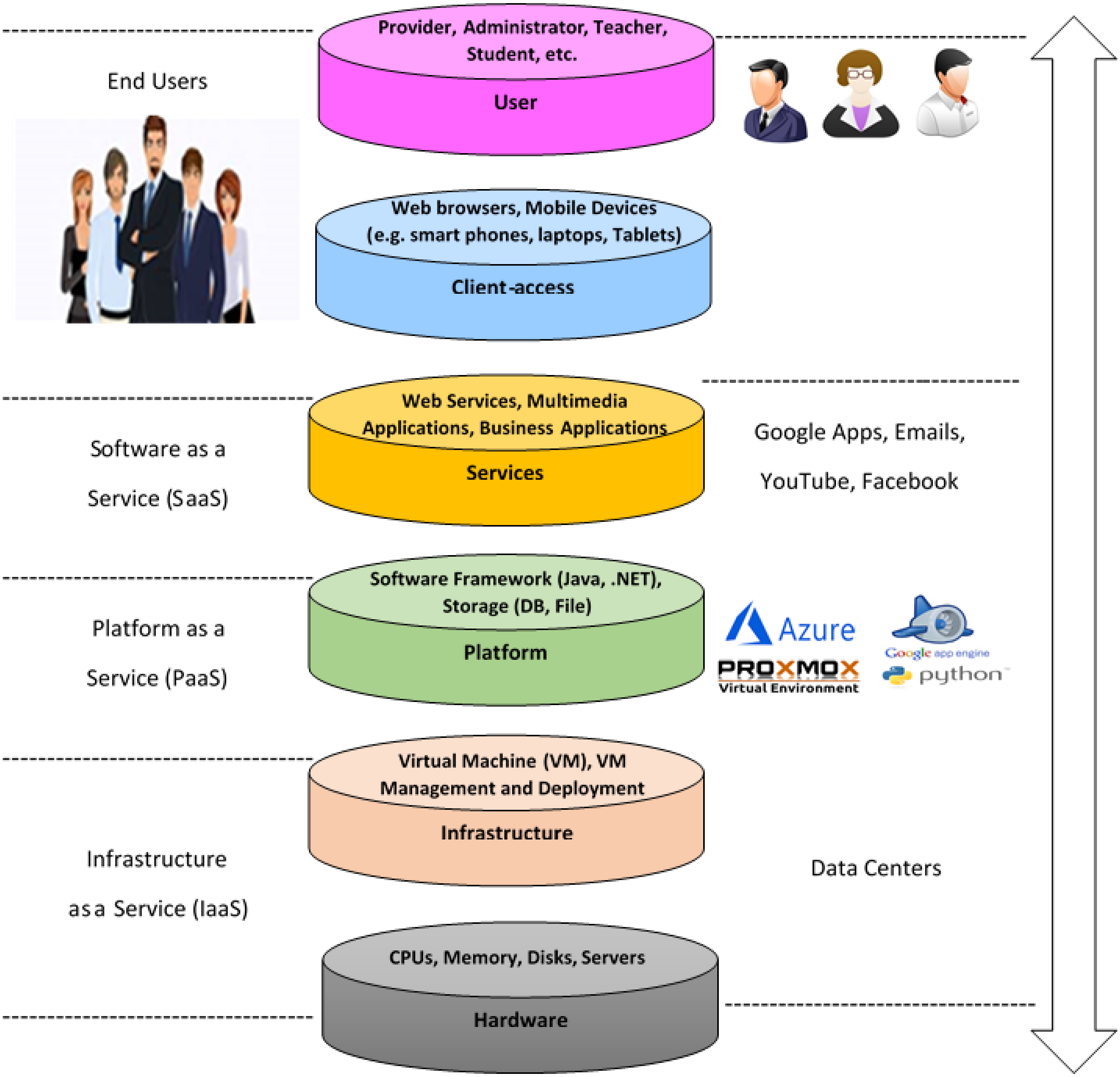

End Users
Provider, Administrator, Teacher, Student, etc.
User
Web browsers, Mobile Devices (e.g. smart phones, laptops, Tablets)
Client-access
Software as a Service (SaaS)
Web Services, Multimedia Applications, Business Applications
Services
Google Apps, Emails, YouTube, Facebook
Platform as a Service (PaaS)
Software Framework (Java, .NET), Storage (DB, File)
Platform
Azure
PROXMOX Virtual Environment
Google app engine
python
Infrastructure as a Service (IaaS)
Virtual Machine (VM), VM Management and Deployment
Infrastructure
Data Centers
CPUs, Memory, Disks, Servers
Hardware

Fig. 7. The architecture of the proposed ELECCM system

Web Assessment of Libyan Government e-Government Services

Mohd Zamri Murah[1], Abdullah Ahmed Ali[2]
Center for Cybersecurity,
Universiti Kebangsaan Malaysia,
Malaysia

Abstract—**Libya has started transferring traditional government services into e-government services. The e-government initiative involves the use of websites to offer various services such as civil registration, financial transaction and private information handling. Currently, there has not been many studies about the security assessment of the Libyan government websites. Therefore, in this paper, we did a web security assessment of 16 Libyan government websites. The main purpose of this study is to determine the security level of these websites. The web security assessment was done in four phases: Reconnaissance, Enumeration and Scanning, Vulnerability assessment (web vulnerabilities and SSL encryption evaluation) and Content Analysis(security and privacy policies). The results showed that 9 websites have high and medium level vulnerabilities. Only 3 websites have *A* SSL rating. Also, only 3 websites have published security and privacy policies. We found 1 *highly unsafe* website, 6 *unsafe* websites, 8 *somewhat safe* websites and, 1 *safe* website. Overall, the study indicated the Libyan government websites are adequately secured without major security issues. Since these Libyan government websites deal with sensitive data, adequate security measures should be implemented to reduce the vulnerabilities and to mitigate future cyber security attacks.**

Keywords*—*Libya; e-Government; web security assessment; information security; website vulnerability; penetration testing

I. Introduction

Internet technology has made a great contribution in changing the global economy. Many governmental and private organization see the opportunity to improve efficiency by providing services online (E-services) through websites or portals[1][2]. The e-services or websites are important to make organization compete and survive in the global economy. Therefore, many governmental and private organization transferred traditional services into e-services which made peoples' lives easier, by getting serve without the constraints of time, location and with less effort and cost[3]. However, the increased usage of websites brought up many new security issues[4]. The websites might have various flaws and weaknesses which they could be exploited by cyber attackers. These security issues are threatening the confidentiality, the integrity of peoples and government information, and threatening availability of the services[5], [6], [7]. According to Edgescan vulnerability statistic report 2018, that both large global organization and governments have faced various breaches. Millions of clients' and employees' records were leaked, and web services are facing various critical and high vulnerabilities.

Libya is one of the countries that have started to transfer traditional government services into e-government such as websites and portals. However, Libya is facing some challenges to implementing these online services. Some main challenges are[8][9]:

1) Lack of studies and researches on the implementations of e-government in Libya.
2) Low trust in e-services from the users.
3) Security and privacy concerns about the websites from the users.

Security of websites is one of the main concerns in Libya today[10][11]. There has been several hacking cases happened in the Libyan government websites due to the lack of security and defensive capabilities[12], [13]. Moreover, not many studies has been done to assess the current security level of Libyan government websites[14]. This is because the Libyan government has only started its e-government in 2013. Therefore, it is important to conduct a study of the current security level of the Libyan government websites. The result of this study might encourage the government to concern more about the importance of web security[15].

II. Related works

Abuzawayda, Y.[16] investigated security issues in Libya by conducting a vulnerability assessment of four Libyan government websites using three vulnerability scanning tools: *N-Stalker*, *Acunetix* and *Nessus*. Also, a survey was carried out to collect data from IT managers. The research results showed that many websites were suffering from various vulnerabilities: critical, high, moderate and low. Ihmouda, R.[14] conducted a web penetration testing on three Libyan government ministries websites using three vulnerability scanning tools: *N-Stalker*, *Acunetix* and, *Nessus*. Moreover, they also interviewed experts to understand the security status of the Libyan government websites in general. The results also showed that many websites have various vulnerabilities and the current security of these websites status need to be improved.

A security assessment was conducted for 51 states government websites in the United States by Zhao(2010)[1]. The assessment was a combination of three methods: web content analysis by searching for security and privacy policies implementation, information security auditing by evaluating SSL encryption, and computer security network mapping using *nmap* scanning tool. The results indicated that many state government websites in the USA were vulnerable to cyber attacks.

Awoleye,W.[17][18] conducted a vulnerability assessment for 64 Nigerian government websites under the domain *gov.ng*.

The assessment carried out using web a web scanner *Acunetix*. The websites were divided into 8 categories which they were evaluated and compared between the categories. The results indicated many Nigerian government websites were open to cyber attacks.

AL-Sanea, M.[3] assessed 150 financial, academic, governmental and commercial websites in Saudi Arabia. The assessment has been done using open-source tools *W3af* and *Skipfish*. Also, they compared between governmental and commercial websites in terms of vulnerabilities numbers. The results indicated some websites are vulnerable to cyber attacks.

We summarized the previous studies with respect to government websites security assessment in Table I. From previous studies, we concluded that many government websites in Saudi Arabia, Nigeria, USA and, Libya are vulnerable to cyber attacks. In this paper, our aim is to determine the current security level of Libyan government websites.

TABLE I. PREVIOUS STUDIES ON SECURITY ASSESSMENT ON GOVERNMENT WEBSITES AND THE TOOLS USED. IN THIS STUDY, WE WILL USE THE SIMILAR TOOLS FOR WEB VULNERABILITIES ASSESSMENT.

Study	Year	Data	Tools
Abuzawayda, Y.	2016	4 Libya government websites	N-Stalker, Acunetix, Nessus
Ihmouda, R.	2013	3 Libya governments websites	N-Stalker, Acunetix, Nessus
Zhao, J	2010	51 United Stated government websites	nmap, SSL, security policy
AL-Sanea, M.	2015	150 financial, academic, governmental and commercial websites is Saudi Arabia	W3af, Skipfish
Awoleye,W.	2012	64 Nigerian government websites	Acunetix

A. Web Application Vulnerabilities

A vulnerability is a security flaws, defects or mistakes in software and system that can be directly exploited by cyber attackers to gain access or to hack the system[19][20]. A good deal of research have found that web applications in general are unsafe[21][22]. There are many types of many types of web vulnerabilities. There are vulnerabilities databases that list all the web vulnerabilities and rank their level of risk. One widely used vulnerabilities' database is CVE and CVSS database[23]. The OWASP Foundation also published top ten web vulnerabilities[24]. Among the OSAWP top vulnerabilities are SQL Injection, Broken Authentication, Sensitive Data exposure, XXL External Entities, Broken Access Control, Security Configuration, Cross-Site Scripting, Insecure Serialization, Using Components with Known Vulnerabilities and, Insufficient Logging and Monitoring[25]. SANS institute also provided 25 top web vulnerabilities[26].

There are many security assessment frameworks to assess websites security. Some methods are manual and others are semi-automated or automated. In recent years, automated web security assessment have become the first choice because its can save time, effort and, covers more security issues. The automated web security assessment consist of three phases: crawl the website and try to list all pages and its links with input vectors, generate specific input values to be submitted to the website and, search for vulnerabilities based on the website responses[27]. Web scanners are different from one another. Some can find more vulnerabilities than others. Therefore, different web scanners will produce slightly different result from one another.

In this study, we use *Acunetix* and *Netsparker* for web scanning. These two tools are considered among the top web scanner available. Other tool that can be used are *AppScan*(IBM), *Arachni*, *Burp Suite*, *WebInspect*(HP) and, *Nessus*[28][29].

B. Secure Socket Layer (SSL)

SSL is a protocol used for securing Internet communication through encryption, decryption and authentication[30]. SSL uses private key to encrypt the transferred data through SSL connection. This allows confidential data such as credit card number, private information and, financial transaction to be transferred through the Internet safely. URLs that uses SSL start with HTTPS, to differentiate its from normal HTTP connection that uses clear text.

SSL protocol establishes secure connection between the website and the user. Its provides authentication between both end points. Also, SSL provides integrity and privacy during the data exchange between the website and the user[31]. Transmitted information between the website and the user is encrypted by SSL, thus ensure high degree of confidentiality.

In general, SSL contains two phases: a hand shake phase and a data transfer phase. During the hand shake phase, a browser will connect to an SSL-based website and request the website to identify itself. In return, the website will send a public key a copy of its SSL certificate. The browser will check the SSL certificate and send an encrypted key back to the website. Finally, the website return an encrypted key with content as a message. The browser will encrypt the message and completes the hand shake phase. After this phase, the browser and the website will continue in a data transfer phase.

It is important for government websites that offer online information and sensitive information transaction to implement secure SSL encryption on their websites. This is important to ensure the security of sensitive information such social security numbers, credit card numbers, private information, health information and, financial information. Sharing information without SSL is a critical risk and may lead to data leakage of sensitive information.

The implementation of SSL can be evaluated using *Qualys SSL* evaluation tool. The tool is an open source software. Other similar tools are *SSL Labs*, *Symantic SSL*, *SSL Analyzer* and, *McAfee SECURE*. This tool check for validity of certificate, protocol version, key exchange, cipher strength and, overall rating.

C. Security and Privacy Policy

The use of government websites involves sharing confidential information between the users and the websites. Users are usually concern with the risk of sharing such information. There are many cases of data leak and data breach where much confidential information is stolen from many websites. Users are less confident in using the government websites if there is no known security policy and privacy policy implemented on the websites[1].

Therefore, in orders to make the users confident in the government websites, the government need to implement and to publish security and privacy policy on the websites. The security policy indicates how secure are the websites and privacy policy indicates how private information is being maintained and used. These policies need to be published and to be make known to all users so that they will trust the websites. If the government websites do not implement or publish their policies, it will be a concern to users to trust the websites and share private information[1].

III. Methodology

The security assessment framework consists of four main phases: Reconnaissance, Enumeration and Scanning, Vulnerability Assessment, and Content Analysis as shown in Fig. 1. We didn't conduct any exploitation or proof-of-concept for any vulnerabilities. The security assessment is based on passive penetration testing[32].

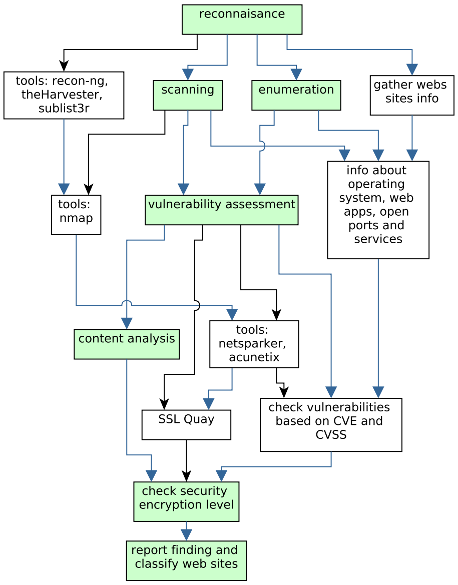

reconnaisance
tools: recon-ng, theHarvester, sublist3r
scanning
enumeration
gather webs sites info
info about operating system, web apps, open ports and services
tools: nmap
vulnerability assessment
content analysis
tools: netsparker, acunetix
SSL Quay
check vulnerabilities based on CVE and CVSS
check security encryption level
report finding and classify web sites

Fig. 1. Security assessment procedures. This procedures consists of four major phases; Reconnaissance, Enumeration and Scanning, Vulnerability Assessment, Content Analysis. The first three phases are the standard procedures. The fourth phase is an additional phase that we proposed.

A. Reconnaissance

Reconnaissance is the stage when we collect as much as possible information about the target from Internet, DNS, and, other public available information. There are two types of reconnaissance: passive and active. Passive reconnaissance is to gather information from search engines and other Internet tools while active reconnaissance is to gather information from a direct contact with the target through social engineering.

This study used passive information gathering by searching for all the Libyan government websites that under the domain *gov.ly*. There are many open source reconnaissance tools that can be used such as *recon-ng*, *sublist3r*, *discovery* and, *theHarvester*[33]. In this study, *TheHarvester* tool have been used to find the sub- domains of *gov.ly* domain. *TheHarvester* is a tool that uses publicly available resources from Internet search engines like Google, Bing, and, Shodan to search for subdomains, hosts, emails and other information. The command that used in this study was:

```
theharvester -d gov.ly -b all > out.txt
```

The command will save the result in the text file `out.txt`.

B. Enumeration

Enumeration is the phase when we gather information about the network and information technology infrastructure such as open ports, operating systems, running services, IPs, status of firewall and, routers[34]. There are also two distinct types of enumeration: active and passive. Passive enumeration is the utilization of the received packets from a website host and it does not require any packets to be sent. Active enumeration is very noisy which requires packets to be sent and waiting for a reply from the website host.

When an active enumeration is conducted, the website or the firewall at the website would detect any attempts from the Internet. Therefore, any active enumerations are logged by the system, thus would alert the website owner of a possible cyber attack. For this reason, enumeration normally done in stealth mode to avoid detection by the websites being scanned.

This study used active enumeration by using *nmap*[35] to scan for open ports, operating systems, running services, and main IPs from the Libyan government websites. The command that has been used in *nmap* scanning is:

```
nmap -iL -F -Pn -sV -A -oX
```

The enumeration phase took a lot of time to be conducted. In this study, we took two working days to enumerate and to scan all 16 websites using *nmap*. One possible issue was the speed of network to reach Libyan websites, and another was probably the websites implemented some security measures to avoid active scanning and enumeration such as throttling the web traffics.

C. Vulnerability Assessment

Vulnerability assessment[36] is the phase where we search for vulnerabilities and flaws in the website's network architecture, operating systems, web applications, content management system and, infrastructures. There are two types of vulnerability scanning: Manual and Automatic. Manual vulnerability scanning requiring advanced skills, experience and may takes a long time. This normally done by experienced hacker and black hats[37][38]. Automatic or semi-automatic vulnerability scanning is much faster than manual. It can be done using open source vulnerabilities scanners like *Vega* and commercial vulnerabilities scanners like *Acunetix* or *Netsparker*. In this study, we used an automatic method to scan for web applications vulnerabilities using *Acunetix* and *Netsparker*.

Different web vulnerabilities scanners would produce different results from one another. This is because each scanner uses different algorithms to detect and to identify vulnerabilities. For example, some vulnerabilities are discovered using web vulnerabilities A but not by web vulnerabilities B, and vice versa. A web vulnerabilities scanning requires a lot of time because of the software need to crawl the website and to verify each vulnerability found. A typical web vulnerability scanning for a typical website will take about 8-15 hour.

In this phase, we also evaluated SSL(Secure Socket Layer) encryption implementation as part of vulnerability assessment using *Qualys SSL.*

1) Web Vulnerability Scanning: In this phase, we used *Acunetix* and *Netsparker* to scan for web vulnerabilities of the 16 Libyan government websites. The time required to scan the websites depends on the websites' security, firewall protections, web application firewall, network speed and network protection. A typical website scanning takes between 2 and 8 hours. The reports from *Acunetix* are very comprehensive, depending on the websites complexity. The reports included vulnerabilities types such as high, medium, low or informational and their CVE/CVSS rating. The reports doesn't indicate any counter-measures for the vulnerabilities. A website scanning using *Netsparker* also takes from 1 to 12 hours. The reports generated were very extensive and includes CVE/CVSS rating. However, *Netsparker* reports give some counter-measures for each vulnerabilities.

This active web vulnerability scan is very noisy and will be logged into the system log file. The scan could also trigger the firewall or (Intrusion Detection System) IDS alarm about a possible cyber attack to the website. Some websites have implemented a counter-measure where it would block connections from Internet that appears to be an active scan[39].

2) Secure Socket Layer (SSL) Encryption Evaluation: During this phase, we used *Qualys SSL* to evaluate SSL encryption implementation at each government website. We assume that it is essential for a government website to implement a secure SSL to protect the data security on the website. If a government website does not implement SSL for data transaction, the data will be at risk. Many government websites deals with highly sensitive and crucial data, and SSL implementation is an important requirement.

The tool *Qualys SSL* checks for SSL validation, certificate expiration, cipher strength and, protocol version of SSL implementation. The results give a detailed and an overall rating for SSL encryption implementation at the websites. The rating levels are: **A**, **A+** for secure encryption, **B**, **C**, **D**, **E**, and **F** means need some updates or improvements, **T** is not trusted, usually because of certificate expiration. If there is no SSL implementation, the evaluation will indicate as such. The SSL evaluation takes between 5 to 10 minutes for each website.

D. Content Analysis

In this phase, we searched for security and privacy policies listed on the 16 Libyan government websites. The idea is that, if the website is serious about security and privacy issues, the website will published the policies on websites for the users. This will indicates that the websites follow the current standard in security and privacy issues. This practice also increase the user trust on the websites. The search was done manually by opening each website and searched for security and privacy policies links in all main page sides. We also checked for the availability of the links provided. Usually, in Arabic websites, security and privacy policies are named by their Arabic links.

E. Safety Level Classification

Based on our experimental previous results, we propose a new safety classification model based on all three factors: vulnerabilities analysis, content analysis and SSL encryption assessment. We called this new classification a website safety level. There are four levels of safety: highly unsafe(A), somewhat unsafe(B), unsafe(C) and safe(D). The basic idea is to combine all security assessment results and to come up with a safety status by combining all three important factors as judged by security experts. The safety classification model is shown in Table II.

TABLE II. A PROPOSED SAFETY CLASSIFICATION MODEL FOR A WEBSITE BASED ON SECURITY ASSESSMENTS, CONTENT ANALYSIS AND SSL ENCRYPTION EVALUATION. IN THIS MODEL, THERE ARE FOUR LEVELS OF *safe*: A(*highly unsafe*), B(*unsafe*),C(*somewhat unsafe*) AND D(*safe*).

		SSL rating			
vuln	**data**	**A**	**B**	**T**	**no SSL**
critical	unencrypted	A	A	A	A
	encrypted	B	B	A	A
high	unencrypted	B	B	B	B
	encrypted	C	C	B	B
medium	unencrypted	C	C	B	B
	encrypted	C	C	B	B
low	unencrypted	C	C	B	B
	encrypted	D	D	C	B
info	unencrypted	D	D	C	C
	encrypted	D	D	D	C

IV. RESULTS AND ANALYSIS

A. Results from Reconnaissance Phase

We found 742 hosts in the domain and subdomain. The 742 hosts have been copied into Excel file and classified to get only the main domains of the government websites. We also check for live websites, and eliminated dead links and duplicate websites. We identified 37 Libyan government websites under the domain *gov.ly* from our analysis.

We verified the 37 government websites manually using a web browser. This verification was important because there were still some available websites related to the previous government. However, these websites are not used any more due to Libyan government transformation. The verification was conducted by accessing each website to check for availability of the websites. We also checked whether the websites were under the control of the current government "Government of National Accord". The verification process resulted in 16 available websites under the current government. The 16 government websites have been changed and indicated by the letter (w) followed by a number to protect the confidentiality of the websites and to avoid abusing the sensitive information that the experiment might revealed.

B. Results from Enumeration and Scanning phase

In the Enumeration and Scanning phase, we used *nmap*. From *nmap*, we discovered the type of operating system used by the websites, how many open ports were available, the type of running services and the websites IPs number. The information is summarized in Table III.

TABLE III. THE RESULTS FROM *nmap* SCANNING ON 16 LIBYAN GOVERNMENT WEBSITES. WE GET INFORMATION ABOUT OPERATING SYSTEM (YES/NO), THE NUMBER OF OPEN PORTS(COUNT), TYPE OF RUNNING SERVICES(YES/NO) AND, IP NUMBER(YES/NO). THE WEBSITES HAVE BEEN CHANGED USING NAME $w_1 \ldots w_{16}$.

websites	operating system	services	open ports	IP
w_1	yes	yes	10	yes
w_2	no	no	2	no
w_3	yes	yes	10	yes
w_4	yes	yes	10	yes
w_5	yes	no	2	no
w_6	yes	yes	13	yes
w_7	yes	yes	11	yes
w_8	yes	yes	1	yes
w_9	yes	yes	9	yes
w_{10}	no	yes	9	yes
w_{11}	no	yes	10	yes
w_{12}	yes	yes	11	yes
w_{13}	yes	yes	11	yes
w_{14}	yes	yes	13	yes
w_{15}	no	no	3	yes
w_{16}	yes	yes	9	yes

From the results, we obtained information about operating system from 12 websites. Many of these websites used outdated Window XP and Window Server or Linux 2.0 series operating system. The use of outdated operating system could cause a serious risk for the websites. The operating information could be used by cyber attackers to launch cyber attacks based on the outdated operating system vulnerabilities. We discovered 13 websites which disclosed their running services. The information about running services such as *NTP* and *telnet* could be used to launch another type of Internet attacks. We found 11 websites that have the number of open ports larger than 3. Typically, a website only needs to open three required ports (HTTP, HTTP, ssh), and close other ports to reduce attack vectors.

C. Results from Vulnerability Assessment

In this phase, we used web vulnerabilities scanner *Acunetix* and *Netsparker*. These two web scanners are the standard tools in the industry for web scanning. Each tool have its strengths and drawbacks. The results from the web scanning is shown in Table IV.

From the web vulnerabilities scanning, we obtained 1 critical vulnerability, 24 high vulnerabilities, 139 medium, 129 low and, 230 informational from 16 websites. We have 1 website with 1 critical vulnerability, 7 websites with high vulnerabilities, 15 websites with medium and low vulnerabilities. Many of the common vulnerabilities are shown in Table V.

TABLE IV. RESULTS FROM WEB VULNERABILITIES SCANNING USING *Acunetix* AND *Netsparker* THAT INDICATE THE NUMBER OF VULNERABILITIES FOR EACH CATEGORY HIGH(H), MEDIUM(M), LOW(L) AND INFO(I).

	Acunetix				Netsparker			
	H	M	L	I	H	M	L	I
w_1	4	17	45	17	1	2	6	12
w_2	0	1	0	0	0	0	4	14
w_3	0	0	1	1	1	1	2	7
w_4	0	1	3	0	0	0	1	6
w_5	0	0	0	0	0	2	1	8
w_6	0	0	1	0	1	3	11	24
w_7	2	91	4	18	2	2	7	13
w_8	0	0	1	0	0	4	3	10
w_9	0	1	0	0	0	0	2	5
w_{10}	0	0	1	1	1	2	10	24
w_{11}	0	0	0	1	0	1	4	6
w_{12}	0	2	0	1	0	1	0	9
w_{13}	0	1	1	0	0	0	1	7
w_{14}	9	1	2	2	2	3	11	28
w_{15}	0	0	1	0	0	0	0	8
w_{16}	0	1	2	0	1	1	4	8
mean	0.94	7.25	3.86	2.56	0.56	1.38	4.19	11.81
max	9	91	45	18	2	4	11	28

TABLE V. THE COMMON VULNERABILITIES, THEIR RISK LEVEL AND THEIR SECURITY IMPACTS.

vulnerability	**count**	**risk**	**impact**
application error message	78	medium	allow hackers to determine web apps being used
out of date version of JQuery	12	medium	security not patched
out of date wordpress	5	high	security not patches
cross site scripting	5	high	allow hackers to hijack websites
CSRF protection	4	medium	allow hackers to hijack websites
PHP disclosure	3	medium	allow hackers to attack PHP
PHP DOS vulnerability	3	medium	allow hackers to attack PHP
out of date PHP	2	high	security not patches
data in clear text	2	medium	allow hackers to sniff sensitive info

The second vulnerability assessment is SSL encryption evaluation. This step is to determine how a website handle sensitive data such as user data, login information, privacy and, financial transaction. A secure website would use an SSL protocol to encrypt data transaction between the website and the user to ensure the data security. The results is shown in Table VI.

In the Content Analysis phase, we manually searched for security and privacy policies on the 16 websites. Security policy concerns with how secure is the data being used on the website transaction such as login information, personal information, financial data and, sensitive information. Privacy policy concern with whether the website tracks the users that use the websites by extracting IP number, Geo location, time of transaction and, personal information. It was to determine whether the websites informed the users about the security and

TABLE VI. RESULTS FROM SSL ENCRYPTION EVALUATION. THE WEBSITES WERE RANKED BASED ON THEIR LEVEL OF SSL IMPLEMENTATION. THE RANKING ARE A(SECURE COMMUNICATION), BNEED UPDATES AND IMPROVEMENTS, T(NOT TRUSTED, CERTIFICATE EXPIRED) AND, NO SSL.

	SSL features				
	certificate	**protocol support**	**key exchange**	**cipher strength**	**overall rating**
w_1,w_7 w_3,w_{11} w_{16}		*no secure protocol*			
w_2, w_5 w_{12}	100%	95%	90%	90%	A
w_4, w_6 w_9, w_{10} w_{13},w_{14} w_{15}	100%	95%	70%	90%	B
w_8	-	90%	70%	90%	T

privacy policies implemented at the websites. From our sample 16 websites, only 3 websites specifically mentioned security and privacy issues. The others didn't have any links to security or privacy issues.

D. Safety Level Classification

The classification criteria in Table II has been used to classify the websites into four main safety categories: highly unsafe, safe, somewhat unsafe, safe. The criteria are based on security assessment, content analysis and SSL encryption evaluation. Each website would be manually assessed by a panel of security experts to determine the safety category of each website. Table VII showed the safety category and previous security assessments. The current approach for safety classification is a heuristic approach. It is derived based on security assessments and, security experts experiences.

V. DISCUSSION

Based on the security assessments results, SSL encryption evaluations and content analysis results, we found that many Libyan government websites have some vulnerabilities issues and do not have good security implementations. Enumeration and Scanning phase has detected a good deal of information about operating systems, open ports, services, and main IPs about the websites. These information might be used by the attacker to find vulnerabilities of the websites. For example, some websites wre detected using operating outdated systems are Windows XP or old versions of Linux (version 2.0 series). These two operating systems have many old vulnerabilities that could be exploited by hackers.

Open ports are another attack vector for cyber hackers. A hardened website only needs 3 essential open ports for data transmission: *HTTP* at port 80, *HTTPS* at port 443 and *SSH* at port 22. All other ports are not important for a website. Port numbers below 100 are reserved for system services such as *ftp*, *telnet* and, *NTP*. We found twelve of the 16 websites have 4 or more open ports and only 4 websites have 3 open ports or fewer. We opined that if a website have more open ports available, the higher the risk of a cyber attack to a website. Services from a website are served using open ports. Cyber hackers can exploit these available services to gain access to the website. Websites should only employed important services such as *HTTP*, *HTTPS* and, *SSH* only to reduce the risk of illegal access to a website. Basic services such as *ftp* and *telnet* could be exploited by cyber attackers if these services are not properly configured.

TABLE VII. SUMMARY OF SECURITY ASSESSMENT, SSL ENCRYPTION EVALUATION AND CONTENT ANALYSIS. FOR *vulns* COLUMN, THE NUMBER INDICATE THE TOTAL NUMBER OF HIGH(H) AND MEDIUM(M) VULNERABILITIES FROM WEB SCANNING.CONTENT ANALYSIS COLUMN INDICATE WHETHER THE WEBSITE HAS SECURITY POLICY AND HAS PRIVACY POLICY(1,1), NO SECURITY POLICY BUT HAS PRIVACY POLICY(0,1), HAS SECURITY POLICY BUT NO PRIVACY POLICY(1,0), NO SECURITY POLICY AND NO PRIVACY POLICY(0,0) . SAFETY CATEGORY COLUMN INDICATE THE SAFETY CATEGORY: 1(HIGHLY UNSAFE), 2(UNSAFE), 3(SOMEWHAT UNSAFE), 4(SAFE)

website	**vulns**	**SSL**	**Content Analysis**	**Safety**
w_1	21	-	(0,0)	2
w_2	1	A	(1,0)	3
w_3	2	-	(1,1)	2
w_4	1	B	(0,0)	3
w_5	2	A	(0,0)	3
w_6	4	B	(0,0)	2
w_7	99	-	(0,0)	2
w_8	4	T	(0,0)	3
w_9	1	B	(0,0)	3
w_{10}	3	B	(0,1)	2
w_{11}	1	-	(0,0)	3
w_{12}	2	A	(0,0)	3
w_{13}	1	B	(0,0)	3
w_{14}	10	B	(0,0)	1
w_{15}	0	B	(0,0)	4
w_{16}	2	-	(0,0)	2

Using *Acunetix* and *Netsparker*, we found that 9 of the 16 websites have high or medium vulnerabilities as shown in Table IV and, 7 of the 16 websites do not have high or medium vulnerabilities. That security implementation at the 16 Libyan government websites are good but need improvement. There is only 1 website with a critical vulnerability.

The results from *Acunetix* and *Netsparker* are different. This is to be expected since both of them use different algorithms to detect vulnerabilities. Some vulnerabilities were found by *Acunetix* and not by *Netsparker* and, vice versa. *Acunetix* discovered more high and medium vulnerabilities than *Netsparker*. *Netsparker* discovered more low and informational vulnerabilities than *Acunetix*. Therefore, it is good practice to use both web scanners.

The websites with high and medium vulnerabilities might be compromised or might be open to future cyber attacks. Therefore, these high and medium vulnerabilities need to be fixed urgently. Many of these vulnerabilities involve outdated operating systems and web applications. However, The websites with low and informational vulnerabilities are also at risk. Attackers might use this information to find more critical vulnerabilities, to conduct social engineering and, to launch phishing attacks. Many of the web vulnerabilities found at the websites are included in OWASP 2017 top 10 web vulnerabilities: Cross-Site Scripting(XS), Sensitive Data Exposure (lack of SSL, send sensitive data in clear text), Using Components with Known Vulnerabilities (outdated programming language or content management system), Broken Authentication (expired SSL certificate, outdated SSL)and Security Configuration (application error logs).

In SSL encryption evaluations, we found only 3 websites with rating **A**, which indicate the websites implement the latest SSL security measures for data security. We found 7 websites rated **B**, because they have weak SSL exchange keys. The weak keys would allow attackers to do a Man-In-The-Middle (MITM) attack and to access the data communication channel. There are 5 websites with no SSL implementations. These websites would need to implement SSL encryption protocol since government websites normally involves in transferring users credential, private data or sensitive data through the Internet. We found 1 website with an expired SSL certificate. Hackers may take advantage of this website by using MITM attacks or POODLE attacks.

Based our security assessment and content analysis study, we propose a safe classification model that would categorize the websites into 4 *safe* categories: *highly unsafe*, *unsafe*, *somewhat safe*, *safe*. The criteria are based on the web security assessment, the SSL encryption evaluation and the content analysis of security and privacy policies. Based on these criteria, we have 1 *highly unsafe* website because this website involves in handling Libyan citizen private information and have low SSL rating. We have 6 *unsafe* websites with a high number of vulnerabilities and low SSL rating. The websites need to fix their vulnerabilities and improve their SSL implementation. We also have 8 *somewhat unsafe* websites with low numbers of vulnerabilities and low SSL rating. These websites need to improve their SSL implementations. We have 1 *safe* website where its has 0 vulnerabilities and low SSL rating. This website might need some improvement in SSL implementation.

Based on our proposed *safe* classification model, we have 1 *highly unsafe* website, 6 *unsafe* websites, 8 *somewhat unsafe* and 1 *safe* website. For a website to have a safe category, we propose for a website to follow these guidelines;

1) Eliminate critical and high level vulnerabilities. This can be done with regular web vulnerabilities assessment. Also, the website needs to update operating system regularly, to patch securities holes, to updating web apps and, to harden the system.
2) Implement secure SSL and avoid expired certificate. This will make sure the data communication channel is safe.
3) Published policy on privacy and security. This will install confident for users to use the website for sensitive data transactions.

VI. Conclusion

In this paper, we have studied security assessment of 16 Libyan government websites using a four phases framework: Renaissance, Enumeration and Scanning, Vulnerability Assessment and Content Analysis. Our study found 3 websites with more than 9 vulnerabilities and, 12 websites have between 0 and 8 vulnerabilities. Twelve websites implement SSL encryption at different level of implementation with only 3 websites have *A* rating. Only 3 websites have published security and privacy policies on their websites. Based on our *safe* category model, we have 1 *highly unsafe* website due to its SSL implementation and the nature of its operation. It is highly recommended for the websites to keep up-to-date with securities issues, system patches and, cyber attacks vectors. The websites also need to develop security and privacy policies for their users so the users would trust the websites.

Overall, we considered the Libyan government websites are adequately secured without major security issues.We encourage the websites to improve their security implementation by fixing the vulnerabilities, updating security patches, updating system configurations and, improving SSL implementations. For future work, this study can be extended to cover all Libyan government and educational websites. Also, the same study can be conducted on commercial websites. Then, a comparison can be made between these websites from a security assessment perspective.

Acknowledgment

The authors would like to thank Center for Cybersecurity at Universiti Kebangsaan Malaysia for their assistance.

References

[1] J. J. Zhao and S. Y. Zhao, “Opportunities and threats: A security assessment of state e-government websites,” *Government Information Quarterly*, vol. 27, no. 1, pp. 49–56, 2010.

[2] R. Ismail ova, “Web site accessibility, usability and security: a survey of government web sites in Kirghiz republic,” *Universal Access in the Information Society*, vol. 16, no. 1, pp. 257–264, 2017.

[3] C. G. Red dick and M. Turner, “Channel choice and public service delivery in Canada: Comparing e-government to traditional service delivery,” *Government Information Quarterly*, vol. 29, no. 1, pp. 1–11, 2012.

[4] M. Felderer, M. Büchler, M. Johns, A. D. Brucker, R. Breu, and A. Pretschner, “Security testing: A survey,” in *Advances in Computers*. Elsevier, 2016, vol. 101, pp. 1–51.

[5] M. S. Al-Sanea and A. A. Al-Daraiseh, “Security evaluation of Saudi Arabia’s websites using open source tools,” in *2015 First International Conference on Anti-Cybercrime (ICACC)*. IEEE, 2015, pp. 1–5.

[6] M. M. Yusof and A. Y. A. Yusuff, “Evaluating e-government system effectiveness using an integrated socio-technical and fit approach,” *Information Technology Journal*, vol. 12, no. 5, pp. 894–906, 2013.

[7] H. Kasimin, A. Aman, and Z. M. Noor, “Using evaluation to support organizational learning in e-government system: A case of Malaysia government,” *International Journal of Electronic Government Research (IJEGR)*, vol. 9, no. 1, pp. 45–64, 2013.

[8] A. A. Ahmed, S. Dalbir, and M. Ibrahim, “Potential e-commerce adoption strategies for Libyan organization,” *International Journal of Information and Communication Technology Research*, 2011.

[9] O. Elaswad and C. D. Jensen, “Identity management for e-government Libya as a case study,” in *Information Security for South Africa (ISSA), 2016*. IEEE, 2016, pp. 106–113.

[10] T. R. Gebba and M. R. Zakaria, “E-government in Egypt: An analysis of practices and challenges,” *International Journal of Business Research and Development*, vol. 4, no. 2, 2012.

[11] M. M. Elmansori, H. Atan, and A. Ali, “Factors affecting e-government adoption by citizens in Libya: A conceptual framework,” *i-Manager’s Journal on Information Technology*, vol. 6, no. 4, p. 1, 2017.

[12] P. M. Tehrani, N. A. Manap, and H. Taji, “Cyber terrorism challenges: The need for a global response to a multi-jurisdictional crime,” *Computer Law & Security Review*, vol. 29, no. 3, pp. 207–215, 2013.

[13] A. R. M. Yusof, M. F. Sukimi, S. B. Ismail, and Z. B. Othman, “The cyber space and information, communication and technology: A tool for westernization or orientalism or both,” *Journal of Computer Science*, vol. 7, no. 12, pp. 1784–1792, 2011.

[14] R. Ihmouda, N. H. Alwi Mohd *et al.*, "Penetration testing for Libyan government website," in *International Conference on Computing and Informatics*. Universiti Utara Malaysia, 2013.

[15] Y. B. Forti and M. G. Wynn, "A new model for e-government in local level administrations in Libya," in *The Proceedings of 17th European Conference on Digital Government ECDG 2017*, 2017, p. 315.

[16] Y. I. Abuzawayda, "Security issues on Libya's e-government," *Imperial Journal of Interdisciplinary Research*, vol. 3, no. 1, 2016.

[17] O. M. Awoleye, B. Ojuloge, and W. O. Siyanbola, "Technological assessment of e-government web presence in Nigeria," in *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance*. ACM, 2012, pp. 236–242.

[18] O. M. Awoleye, B. Ojuloge, and M. O. Ilori, "Web application vulnerability assessment and policy direction towards a secure smart government," *Government Information Quarterly*, vol. 31, pp. S118–S125, 2014.

[19] N. F. Awang, A. A. Manaf, and W. S. Zainudin, "A survey on conducting vulnerability assessment in web-based application," in *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 2014, pp. 459–471.

[20] S. M. Srinivasan and R. S. Sangwan, "Web app security: A comparison and categorization of testing frameworks," *IEEE Software*, no. 1, pp. 99–102, 2017.

[21] I. Alsmadi and E. Abu-Shanab, "E-government website security concerns and citizens' adoption," *Electronic Government, an International Journal*, vol. 12, no. 3, pp. 243–255, 2016.

[22] N. Antunes and M. Vieira, "Penetration testing for web services," *Computer*, vol. 47, no. 2, pp. 30–36, 2014.

[23] P. Mell, K. Scarfone, and S. Romanosky, "Common vulnerability scoring system," *IEEE Security & Privacy*, vol. 4, no. 6, 2006.

[24] E. Bertino and N. Islam, "Botnets and internet of things security," *Computer*, no. 2, pp. 76–79, 2017.

[25] T. F. OWASP, "Application security risks-2017. open web application security project (OWASP)," 2017.

[26] T. Scholte, D. Balzarotti, and E. Kirda, "Have things changed now? an empirical study on input validation vulnerabilities in web applications," *Computers & Security*, vol. 31, no. 3, pp. 344–356, 2012.

[27] F. R. Munoz and L. G. Villalba, "Methods to test web applications scanners," in *Proceedings of the 6th International Conference on Information Technology*, 2013.

[28] Y. Makino and V. Klyuev, "Evaluation of web vulnerability scanners," in *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, vol. 1. IEEE, 2015, pp. 399–402.

[29] G. Roldán-Molina, M. Almache-Cueva, C. Silva-Rabadão, I. Yevseyeva, and V. Basto-Fernandes, "A comparison of Cybersecurity risk analysis tools," *Procedia Computer Science*, vol. 121, pp. 568–575, 2017.

[30] A. Freier, P. Karlton, and P. Kocher, "The secure sockets layer (SSL) protocol version 3.0," Tech. Rep., 2011.

[31] M. A. Alnatheer, "Secure socket layer (SSL) impact on web server performance," *Journal of Advances in Computer Networks*, vol. 2, no. 3, pp. 211–217, 2014.

[32] G. Weidman, *Penetration testing: a hands-on introduction to hacking*. No Starch Press, 2014.

[33] C. Buchanan and V. Ramachandran, *Kali Linux Wireless Penetration Testing Beginner's Guide: Master wireless testing techniques to survey and attack wireless networks with Kali Linux, including the CRACK attack*. Packt Publishing Ltd, 2017.

[34] S. Mehta, G. Raj, and D. Singh, "Penetration testing as a test phase in web service testing a black box pen testing approach," in *Smart Computing and Informatics*. Springer, 2018, pp. 623–635.

[35] G. F. Lyon, *Nmap network scanning: The official Nmap project guide to network discovery and security scanning*. Insecure, 2009.

[36] S. Shah and B. M. Mehtre, "An overview of vulnerability assessment and penetration testing techniques," *Journal of Computer Virology and Hacking Techniques*, vol. 11, no. 1, pp. 27–49, 2015.

[37] M. A. Mahmood, M. Siponen, D. Straub, H. R. Rao, and T. Raghu, "Moving toward black hat research in information systems security: an editorial introduction to the special issue," *MIS quarterly*, vol. 34, no. 3, pp. 431–433, 2010.

[38] I. Idris, M. U. Majigi, S. Abdulhamid, M. Olalere, and S. I. Rambo, "Vulnerability assessment of some key Nigeria government websites," *International Journal of Digital Information and Wireless Communications*, vol. 7, no. 3, pp. 143–153, 2017.

[39] D. Stuttard and M. Pinto, *The web application hacker's handbook: Finding and exploiting security flaws*. John Wiley & Sons, 2011.

Swarm Eye: A Distributed Autonomous Surveillance System

Faisal Khan[1], Tarapong Sreenuch[3]
Integrated Vehicle Health Management Centre
Cranfield University, Bedford, UK

Jörn Mehnen[2]
Manufacturing and Materials Department
University of Strathclyde
Glasgow, UK

Syed Alam[4, #], Paul Townsend[5, $]
[#]Senior Software Developer
Clockwork IT Limited,
Coventry, UK
[$]Lockheed Martin, UK

***Abstract*—Conventional means such as Global Positioning System (GPS) and satellite imaging are important information sources but provide only limited and static information. In tactical situations rich 3D images and dynamically self-adapting information are needed to overcome this restriction; this information should be collected where it is available. Swarms are sets of interconnected units that can be arranged and coordinated in any flexible way to execute a specific task in a distributed manner. This paper introduces Swarm Eye – a concept that provides a platform for combining the powerful techniques of swarm intelligence, emergent behaviour and computer graphics in one system. It allows the testing of new image processing concepts for a better and well informed decision making process. By using advanced collaboratively acting eye units, the system can observe, gather and process images in parallel to provide high value information. To capture visual data from an autonomous UAV unit, the unit has to be in the right position in order to get the best visual sight. The developed system also provides autonomous adoption of formations for UAVs in an autonomous and distributed manner in accordance with the tactical situation.**

Keywords—Swarm intelligence; distributed surveillance system; bio-inspired algorithm; cooperative UAVs

I. Introduction

Nature has been a source of inspiration for today's various modern and complex technology and intriguing systems. Many inventions have been done by adopting the natural system into either mechanical or electric manner. Swarms are schools of fish, colonies of ants and termites, flocks of birds, herds of land animals. Insects such as mosquitoes, spiders and dragon flies have a compound eye which provides them with a large view angle. This helps them to quickly detect fast movements and compensate for the physical lacks of the single units by 'intelligent' combination of information, [1], [2].

A major problem with a centralized system is that if the centre of the system becomes dysfunctional the whole system may stop working. To make a system more robust, adaptable systems such as autonomous swarms look very promising [3]. The formation of pulling chains by weaver ants or the surrounding of a prey like a target by a group of predators can be seen as examples or multi-pattern formation of swarms in nature. A first engineering adoption of this idea can be seen in the swarm intelligence based multi-robot pattern formation [4]. This model argues that the pattern to be formed is controlled by the surroundings, and the final shape and the size are partially decided by the task or objective of each agent during the coordination.

In the approach used in [5], the focus is on collective intelligence. The rules used to organize their swarm raids has broad application for advancement in programming of multi-agent autonomous systems, the design and for providing insight into the concepts of collective self-organization and intelligence.

Insect colonies distribute resources and tasks to each unit in order to collectively solve difficult problems instead of using any centralized control. The behaviour comes from the local interactions with no global knowledge and it uses a simple set of rules [5]. These rules provide robustness to change, keeping their effectiveness in extremely diverse environment or recourse distribution over a broad range of tasks [5].

The usage of Global Positioning System (GPS) in the formations of autonomous airships has been discussed in [6]. Experiments were carried out using two blimps, both equipped with multi carrier-phase GPS (CPGPS) receivers allowing the establishment of attitude and location of the craft. The experiment identified formations with attitude errors. The errors were in the range of less than five degrees with further positional errors of less than a foot. Formation control is one approach to comprehend cooperative coordination proposed in [7]. Multi-UAV formation flight combines the research of both UAV and coordination, so it has increase interest from both unmanned system and control communities [8]. Cooperative coordination allows that a group of UAVs can follow a predefined trajectory for flight missions while using their on-board sensors to acquire useful information while maintaining a specified formation pattern. Because formation flights of a UAV fleet can significantly increase the universal efficiency of the entire system, it can benefit most of the applications which are handled by a single UAV.

The functionality of compound eyes in animals has been described extensively in literature. The Compound eye is excellent in motion detection [9]. Different facets of the eye are progressively stimulated when objects move across the visual field of the insect. This is call the "flicker effect", which makes the insect respond far better to moving objects

This article contains material from M. Res. Thesis, Cranfield University, England, UK.

than to unmoving ones. [1] Has shown that compound eyes like that of a moth can also show a very high sensitivity to light. This makes some insects, such as the praying mantis (*archimantis latistyla*), a good hunter even during the early and late hours of the day. It has also been shown that some insects such as the drone fly (*eristalis tenax*) can see ultraviolet as well as natural colour light [2]. This helps drone flies to forage together with the honeybees they are mimicking.

The compound eyes of most insects have many hundreds of lens facets. This provides these insects with a very wide field of view. [10] Shows that composite lens eyes of some insects having no more than 50 facets can still show excellent performance. The issue of seeing the world multiple times with different view angles is compensated by the cortex of these insects. [11] Describes a design of a composite eye for computer vision for detecting moving objects in a closed environment. They designed a small unmoving array of seven off-the-shelf cameras and show that it is able to track a moving object with this kind of system. The 3D position of the object can also be determined in the overlap areas. In [12], a compound eye is modeled using a spherical field of view, overlapping Gaussian-shaped receptive fields, a singular viewpoint, and a space-variant receptor distribution. The algorithm creates low resolution spherical images from multiple static perspectives. For representing spherical images, the 3D images are projected onto cubes because current raster graphics technology is optimised to construct planar, perspective images.

After motion detection, the most important part is to identify the moving object. Identification of an object autonomously is a complex task as most real environments have many variations which can affect the image and visibility of the scene. [13] Proposed a method of autonomous real-time vehicle detection from a medium-level. The detection of moving vehicles is vital in tactical decision making for military. The general problem of aerial vehicle detection is also made significantly more demanding due to the non-uniformity of vehicle colour, localised shape characteristics and overall dimension. To determine the relative position within an unknown terrain is a challenging task. The DGPS /AGPS receivers make a fine selection for UAV positioning main sensors, because they are able to reach accuracy up to few centimetres using carrier-phase measurements. DGPS /AGPS receivers are the only position sensors that are generally used for UAV positioning [14]. The reliability of their measurements is critical for UAV missions [15].

The deployment of multiple eyes/cameras on key location could offer a high sensitive and effective motion detection solution. Real-time multiple images can be exploited and analysed. More cameras in the right locations mean fewer blind spots. Most of the motion detection algorithms are based on comparing the frames with each other and some also compare the frames with each other with the dimensions of time to extract the moving object's speed.

This paper exploits the concept of compound eyes in a multi-UAV (i.e. multiple cameras) surveillance application. This, a swarm formation algorithm for optimal automatic positioning of multi-angle 3D visual information sources and implementations will be the contribution of this paper.

The outline of this paper is as follows: Section 2 describes the biological background related to insect compound eye and its relation to Swarm Eye. Sections 3 explains the visibility factor index concept and proposed formation algorithm for dynamically positioning multiple UAVs. A software implementation of the Swarm Eye algorithm and test results are described in Section 4. Finally, conclusions are drawn in Section 5.

II. Biological Background

A. Insect Compound Eye

A compound eye can consist of hundreds of individual photoreceptors. The image perceived is a combination of inputs from the numerous ommatidia (individual 'eye units'), which are located on a convex surface, thus pointing in slightly different directions, see Fig. 1. Each ommatidia consists of a lens, transparent crystalline cone, light sensitive visual cells and pigment cells.

To compare with simple eyes, compound eyes possess a very large view angle and can detect fast movement and, in some cases, the polarization of light. Because the individual lenses are so small, the effects of diffraction impose a limit on the possible resolution that can be obtained. This can be countered by increasing lens size and number. To observe with a resolution comparable to our eyes, humans would require compound eyes which would reach the size of their head [16].

The compound eye is excellent at detecting motion. As an object moves across the visual field, ommatidia are progressively turned on and off. Because of the resulting 'flicker effect', insects respond far better to moving objects than stationary ones. Honeybees, for example, will visit wind-blown flowers more readily than still ones.

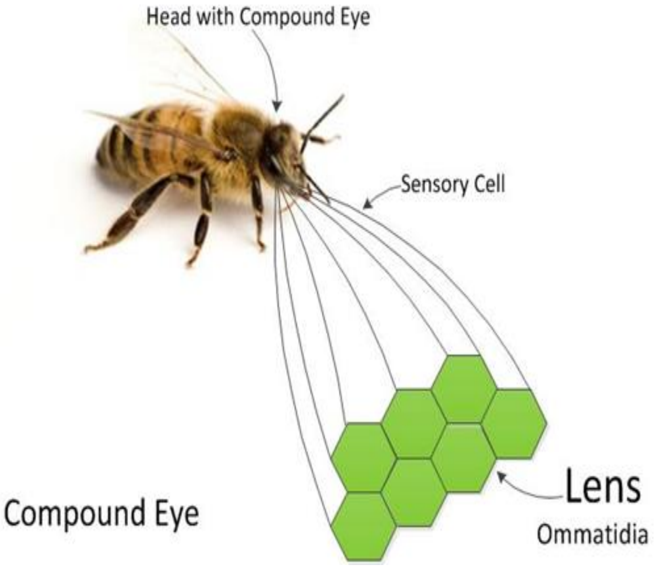


Fig. 1. Basic Structure of Compund Eyes.

B. *Swarm Eye*

The Swarm Eye philosophy follows the idea that a set of individuals can form a very powerful system when joined and self-organized in a swarm. The individuals in a swarm can communicate and interact with each other directly or indirectly. The Swarm Eye System consists of several independently running units that can communicate in any specified way. The units can also take images while navigation through a simulated environment. The images can be combined in any fashion for generating e.g. high quality panoramic images, 3D image generation, and automatic position determination in case of GPS malfunctions in a tactical situation or rapid movement detection in a scene. The intelligent combination of multiple units gathering images and processing them at the same time leads to techniques that can make use of the full power of parallel images processing.

Swarm Eye follows the idea of using method motivated by nature to shape swarms. Swarms are structures caused by the emergent behaviour of individuals. Natural swarms prove to be very effective while being robust against variations in the environment or even loss of individuals from the swarm. Swarms follow rules that can lead to extremely complex behaviour [17]. The rules are often surprisingly simple and the components of the swarm are very basic [17]. However, the emergent behaviour leads to very powerful and effective results. The Swarm Eye system is a distributed system therefore the loss of a few individuals may compromise the performance, but the system will be in working order.

The Swarm Eye is a tactical decision support system that mainly uses optical information. It can be airborne or land based. The swarm consists of a scalable number of rather simple units that collaborate in a self-organized way. The Swarm Eye concept is designed for medium to large scale swarms. The swarm unit numbers can range from a couple to several hundred units. The Swarm Eye is designed to operate in the field, where several sources of information are needed to make well informed tactical decisions.

In nature, the compound eye is excellent at detecting motion. The concepts that can be adapted to Swarm Eye applications are:

- To have many eyes /camera
- Observe the object from many different angles

In Swarm Eye, the camera locations can be in different places, whilst in compound eye each ommatidia is fixed in pace next to each other. The eye units are the basic components in both the compound eyes and Swarm Eye.

Fig. 2. Decentralized System.

III. Distributed Surveillance Algorithm

A. *Decentralized Communication*

In a centralised communication approach, every unit in the swarm communicates with one centralised hub and the central hub sends commands/messages to all other units. The drawback of this approach is that if the centralised hub fails then the whole system will eventually fail.

There are different schemes to implement a decentralised communication model. In general, decentralised communication requires each unit to broadcast the message to every unit, see Fig 2. These systems are more robust as they do not rely on any central hub of the system. They are also fault tolerant as the failure of a communicating unit does not adversely affect the functionality of the whole system. In the development of Swarm Eye, the adopted communication approach is distributed decentralized. This is due to the requirements of autonomous operations and fault tolerance.

B. *Swarm Eye Formation*

The identification and implementation of the proper formation of swarms is a vital element for efficient swarm application. These formations typically occur during the various phases of the swarm lifecycle.

The initial formation refers to the state of the swarm upon its creation; however this initial formation can be changed accordingly based on the swarm deriving parameters such as visibility. This adoption of a new formation overcomes any drawbacks associated with the static formation limitations. Within the context of aerial surveillance this dynamic formation is a desirable behaviour as it aids in overcoming obstacle avoidance and any static restrictions. It also allows for maximum coverage of any area of surveillance from the sky.

The formation is derived by an algorithm that is applied on the swarm members to achieve a particular orientation of these members within the swarm. To aid this process, this paper proposes a v*isibility factor index driven* algorithm to apply to a suitable swarm formation. This algorithm comprises of several steps as depicted in Fig. 3.

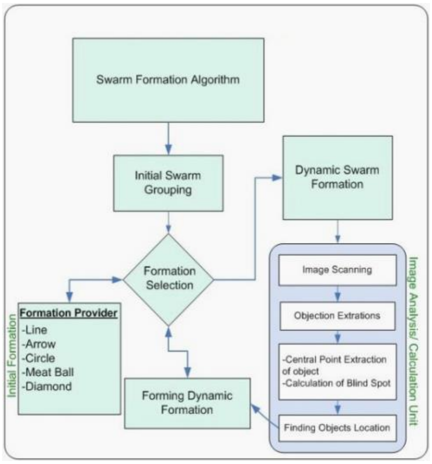


Fig. 3. Architecture of a Decentralized Swarm Eye Flying Formation Algorithm.

These steps aid in achieving a suitable formation for the swarm. The algorithm aims to apply a suitable formation to the swarm so that maximum visibility can be achieved over an area of surveillance.

C. Visibility Factor Index

The visibility factor will be decided depending on how much of the required object is visible to the individual eye. Each eye is assigned a rank. The eye that will cover more area of the required object that eye will have higher Visibility Factor Index (VFI), i.e. rank. VFI ranking is shown in more detailed in Fig 4.

Fig. 4 shows four swarm units; each unit with its own point of view. The Unit 1 can see more of the required object (the car behind the tree) - therefore in the ranking table it is ranked higher than the others. The VFI ranking is dependent on the visibility, therefore, higher visibility yields higher VFI values and vice versa. The VFI aids the swarm system to establish swarm units with higher visibility at their present location. On the other hand, VFI allows for selection of low VFI swarm units for relocation in order to improve their VFI. The swarm formation algorithm also utilises low VFI swarm units as candidates for location transformation and swarm formation.

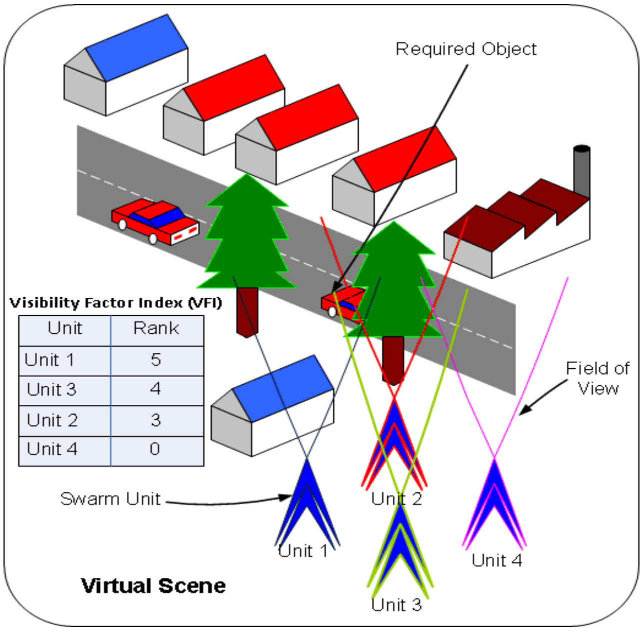


Fig. 4. Virtual Scene Demonstration of Visibility Index Execution.

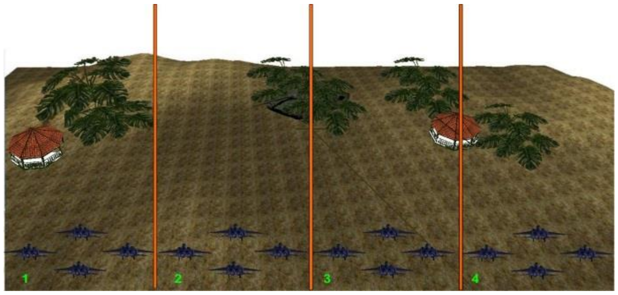


Fig. 5. Example of Full Swarm Dividing into Four Small Sub-Groups with Four UAV Each.

D. Formations

The swarm formation can be divided into 1) Forming Initial Formation and 2) Forming Dynamic Formations. The initial formation refers to the formation which will be adopted by the swarm at the start up. As soon as the unit will be airborne and get registered into the swarm they will form a formation to fly, which could be arrow, diamond, semi-circle, circle formations etc. This initial formation can be decided based on a few factors, for example if a swarm unit intents to save fuel then they will adopt the arrow formation [18], [19], and [20]. In the case of poor visibility, the formation will switch into a dynamic formation offering maximum visibility. The dynamic and initial formation algorithms are explained later in this paper.

The dynamic formation will be adopted by the swarm if the required objects are not visible due to blind spots in the current formation. The proposed formation algorithm presented in this section aims to derive a suitable formation based on the Visibility Factor Index. The objective of the algorithm is to apply a suitable formation to the swarm so that maximum collaborated visibility is achieved. The algorithm divides a swarm into several sub-swarms applying suitable formations at both, the sub-swarm and the overall swarm level, see Fig. 5. This approach allows for individual sub-swarms within a main-swarm to have their own formation. With the main swarm divided into multiple individual sub-swarms each can have its own unique formation. This allows for a mix of formations within a main swarm that can be changed accordingly as desired.

The algorithms attempt to create sub-swarms of similar number of units by applying a square root function to the number of units in the swarm. For example, if the total numbers of units are 16 then applying $\sqrt{16}$ will yield 4. The formation will make 4 groups, each group containing 4 units (in other words 4 × 4 = 16 total swarm units) as shown in the above Figure 5. In the case of any leftover unit, these units will join the nearest group. Each group can be used for different purposes as well as share the same task and each group can perform a certain job or task. The square root function approach has been selected due to the simplicity of implementation and it also allows for creation of equal number of groups with similar number of units. A fair distribution of tasks to individual units is also guaranteed by this approach. Various other approaches such as clustering can also be applied as a replacement to this approach. However the intent of this function is allocate units to swarm and if required it can be easily replaced by more efficient similar function.

Once all the units are assigned a group, each group will search its own allocated task. If the group faces an object/terrain with blind spots it will scan all the eyes visual index numbers. The eye which has the best ranked visual index of the required object will scan the required object's image and calculate the centre of the object (ObjC) from its own view. After locating the centre point of the object it will calculate the distance (dBS) between the ObjC and blind spot (BS) and after calculating this distance (dBS) the system will scan the visual index again. The eye which has the lowest visual index and is closest to the destination point will shift its position to a surveillance point which will allow it to cover the previously hidden area of the object. This process will iterate itself until the blind spot of the required object is optimally covered. The overview of the algorithm is shown in the Fig. 6.

The algorithm has been partially implemented and is not fully optimised at this stage. Specifically there may be a case of repeated iterations to cater for the dynamicity of the scene. These iterations might need further optimization but this issue has been left for future work. However, it allows for the exploration of this novel approach within tactical decision making scenarios. Moreover, there are yet features which have to be incorporated into the algorithm, such as obstacle avoidance, keeping the formation, collision avoidance, etc. The swarm formations are designed to have units from 16 to 250; however, the general design idea of Swam Eye is not limited to any number of individuals.

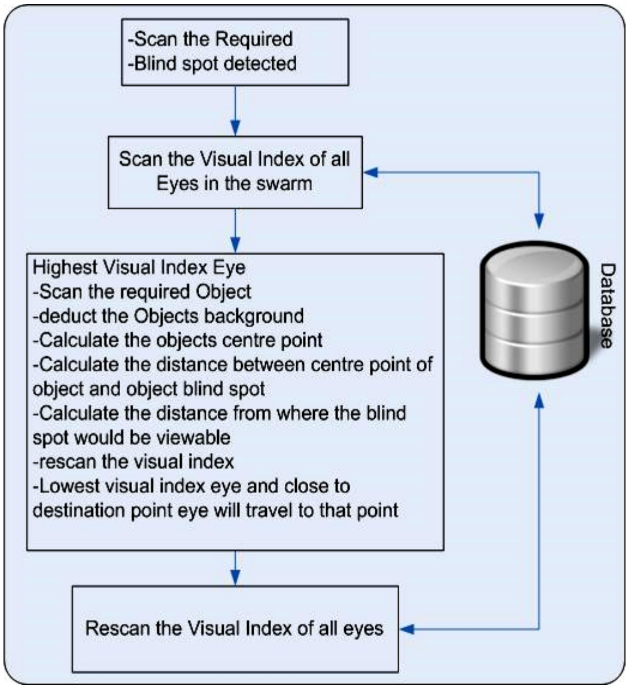

-Scan the Required
-Blind spot detected
Scan the Visual Index of all Eyes in the swarm
Highest Visual Index Eye
-Scan the required Object
-deduct the Objects background
-Calculate the objects centre point
-Calculate the distance between centre point of object and object blind spot
-Calculate the distance from where the blind spot would be viewable
-rescan the visual index
-Lowest visual index eye and close to destination point eye will travel to that point
Database
Rescan the Visual Index of all eyes

Fig. 6. Decentralized Dynamic Formation Algorithm.

IV. Implementation and Results

A. Prototype Software

The Swarm Eye software prototype comes with a simulation environment for testing various application setups. A safe simulation environment has been chosen to prepare for real-world tests with UAVs. The basic principle is to provide a set of autonomous eye units which can analyse a tactical scene in parallel, running efficient image analysis in a distributed manner and, thus, providing richer information to the tactical decision maker in the field. The first steps into this direction have been made by providing a parallel communication platform. The software also delivers a parallel scene simulation and analysis. After that all the essential steps are made to start with the first distributed image analysis processes. The software system consists of

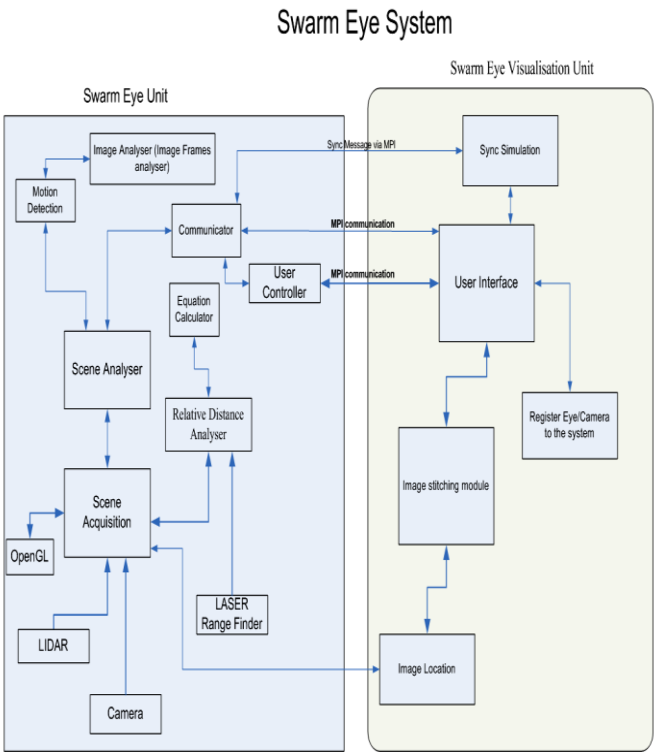


Fig. 7. Swarm Eye Software Structure.

- warm environment
- Swarm Eye units
- Parallel image processing capabilities
- Parallel communication
- User interface

Fig. 7 shows the structural layout of the Swarm Eye software. Although the current implementation consists of several individual modules, the system can be divided into two major parts: the Swarm Eye Units and the Swarm Eye User Interface (Visualisation Unit).

The visualization unit is a separate object in the Swarm Eye system. It connects to the individual Swarm Eye units via MPI (Data exchange interface). The major purpose of the visualisation unit is to link the user with the swarm. Through the registration module each eye is registered at the Visualisation Unit. This way each eye can be addressed and manipulated independently by the user. The environment is simulated in each Swarm Eye unit independently. This allows for highly efficient communication between the eye units as only small changes happening in each eye-environment simulation need to be communicated. A panoramic image generation module which uses stitching techniques is implemented in the visualisation unit. The individual images from the activated Swarm Eye units are collected by the visualisation unit and processed centrally. However, it is not necessary to have only one visualisation unit. The unit can log on to any eye in the swarm, thus avoiding a vulnerable hub-spoke configuration where taking out the hub causes a failure of the whole system.

The Swarm Eye units are the eyes/cameras of the system. They are programmed as objects so they can be copied and independently used as often as desired by the user. This way it is easy to build a swarm of many similar eye units that can interact with the environment, the user and with each other. In each Swarm Eye unit a scene analyser interacts with the scene acquisition module which interacts with the OpenGL [21] environment, a LIDAR (Light Detection And Ranging) data reader, a camera (for real-world applications), and a relative distance analyser. The motion detection module interacts with the scene analyser. Each eye has a fast communication interface (Communicator) to the visualisation unit and the other eyes. The user controller interprets direct commands from the user for e.g. activating a feature or changing the orientation or position of an eye unit. The eye unit can interpret high level commands and execute them independently. This simplifies the control of the swarm and avoids direct low level (protocol level) interaction between user and the single swarm unit.

The Swarm Eye consists of a set of autonomously performing eye units. In order to perform the task in a group, each individual needs to have a fully duplex communication protocol in real time (without significant delay). The messages between the units would be different types of data i.e. images, simple x-y-z-coordinates and messages /commands. Individuals in swarms interact either directly by communication or indirectly through joint behaviour. In Swarm Eye each eye unit can communicate with each other and the visualisation unit via the quasi-standard internet peer-to-peer protocol MPICH2 [22]. All processes in Swarm Eye are running in parallel and can communicate asynchronously. The communication unit also provides a platform for launching new eye units, communication between the objects and to delete the present eye/camera if needed. After each eye is released, it acts as an independent process which runs independently from the initialisation unit. This mimics the independent implementation of each eye unit on an e.g. airborne UAV.

The virtual environment simulation has been created to simulate the camera (eye). This is a 3D simulation which will allow placing a camera or eye at any desired location in the scene. A simulation is cost and time effective as compared to a real camera. There are a few simulations which have been developed and have been adopted. The simulation has been plugged into the GUI as a module and can be replaced by different simulations or by real cameras. There are few simulations that have been created from the start of the project which are all developed in C++/C and by using OpenGL. The C++/C offers to be very effective in processing consumption which is very crucial for this project.

The image rendering module is a key module in Swarm Eye. It supplies the eye units with images from the simulated environment. Significant time was spent in this work in developing an image rendering package that provides simulations that looks as realistic as possible. Until now, the image rendering module is the principle platform for experiments in Swarm Eye. All following experiments on collective image representation, position determination and movement detection have been implemented using this module. The image rendering engine used in Swarm Eye uses the MD2 file format and the OpenGL library for image rendering. During the project the image rendering module in Swarm Eye was significantly modified and improved from a very basic triangulation viewer to a sophisticated 3D rendering engine.

B. *Collective Image*

In Swarm Eye, the problem of generating panoramic images has been solved by image stitching. Image stitching is the process of combining several images together to form one single new image showing the content of the individual images in one single picture. The technology of image stitching is rather advanced today [12], [23]. Swarm Eye makes explicit use of these advanced technologies and incorporated them into the system. Fig. 8 shows the general structure of the panoramic image generation process.

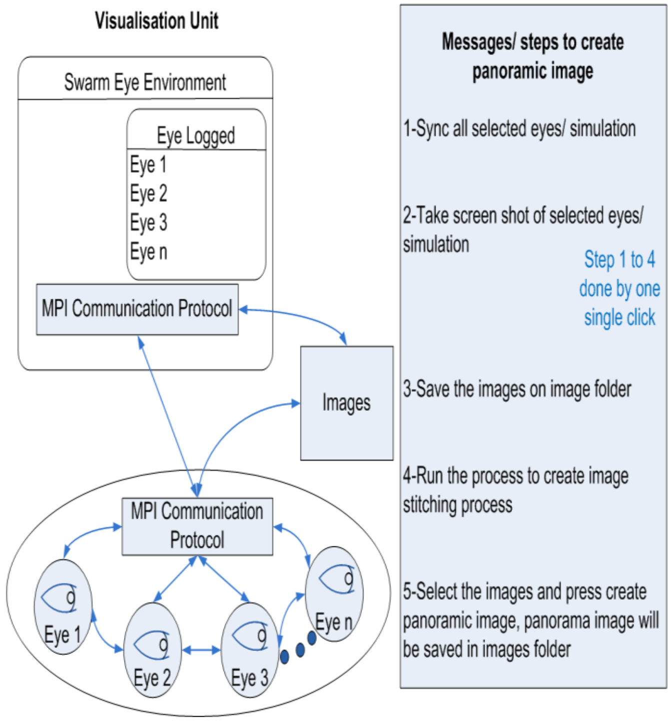

Visualisation Unit
Swarm Eye Environment
Eye Logged
Eye 1
Eye 2
Eye 3
Eye n
MPI Communication Protocol
Images
MPI Communication Protocol
Eye 1
Eye 2
Eye 3
Eye n
Messages/ steps to create panoramic image
1-Sync all selected eyes/ simulation
2-Take screen shot of selected eyes/ simulation
Step 1 to 4 done by one single click
3-Save the images on image folder
4-Run the process to create image stitching process
5-Select the images and press create panoramic image, panorama image will be saved in images folder

Fig. 8. Panoramic View Generation in Swarm Eye.

The eye units contributing to the panoramic image generation are first selected in the visualisation unit by the user. The scene is synchronised via MPICH2 messages so that every eye has the same time frame. In real-world scenarios the synchronisation procedure is, of course, not necessary. Each eye takes a snapshot of the scene and sends it to the Visualisation Unit. In order to perform image stitching in Swarm Eye, the individual views of the eye units need to be synced and offer overlapping coverage of the scene. In practice any orientation of the eyes can be used for this purpose as long as there is a certain amount of overlap between the individual images. An overlap of ca 20% between each image is generally recommended to allow the stitching software to identify unique markers within the overlap area to determine stitching points. The marking procedure is performed by the Scale-Invariant Feature Transform algorithm

SIFT [24]. The RANSAC (RANdom SAmple Consensus) [25] algorithm is used to compare similarities between the point sets identified by SIFT in the individual overlaps. The images are adjusted and transformed so that similar point clusters overlap with the highest probability. The adjusted images are then rendered and combined into one single image. In Swarm Eye, the images are collected and passed to the image stitching software to produce panoramic view [13], [26].

In Fig 9 the views of three eye units are displayed showing a landscape. One can see that each eye is pointing in a slightly different direction covering the major area of a scene.

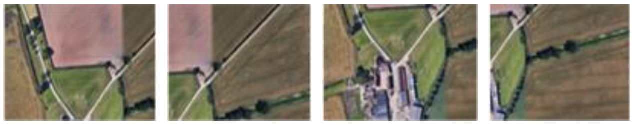


Fig. 9. View of Different Eye units.

Fig 10 shows the compiled panoramic view. After activation of image stitching, this image can be seen. All four images are sent by the eye unit to a shared common port. From there the stitching software automatically collects the images and joins them to a single panoramic image. The image is displayed automatically by the Swarm Eye to a user on the ground who logs into this viewing unit.

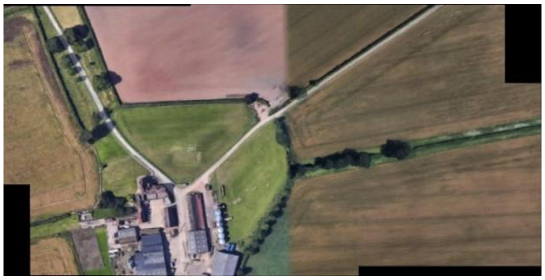


Fig. 10. Panoramic View Generation in Swarm Eye

C. Swarm Formation

The formation algorithm has two parts in the software development. The first part is to form the initial formation and the second part is to form a dynamic formation when a blind spot is being detected or the required object is not visible anymore. This approach can be based on particle swarm optimisation [27]. The visibility factors, object of interest depth perception, assisted technologies such as LIDAR usage also contribute towards format adjustments, see Fig. 11.

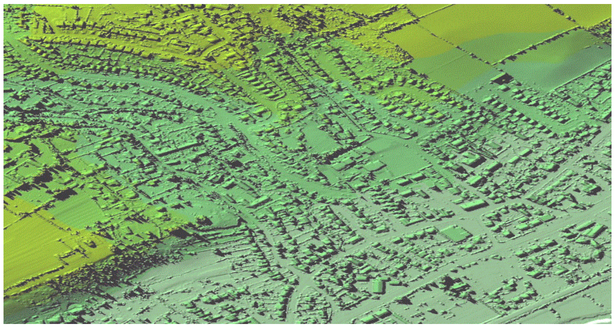


Fig. 11. Aerial LiDAR View of a Landscape.

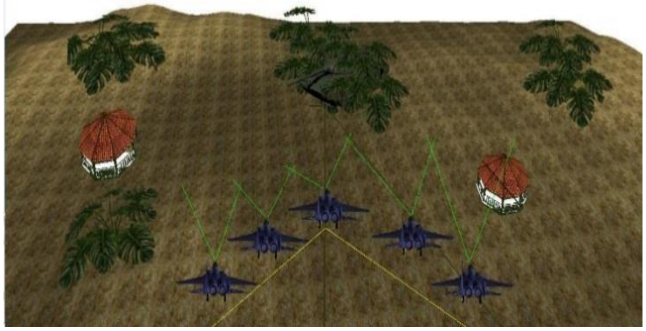


Fig. 12. Arrow Formation of Swarm units.

The selected formation is applied at the initial stage of the swarm lifecycle. If during the simulation a blind spot is encountered by a swarm unit then this event triggers a dynamic reformation of the swarm. Individual swarm units have a limited area of observation. However, when multiple swarm units are placed in an appropriate formation, this yields a larger area of coverage. To maintain the optimum collaborated coverage of the area of surveillance, the individual swarm units maintain a sight distance between each other. Fig 12 presents this scenario where five swarm units are arranged in an arrow formation with the whole swarm coverage area. This feature allows for wide range panoramic views of the scene of interest to be generated.

However, the generation of 3D images from 2D captured images requires suitable multi-angled 2D images. This is where dynamic formation plays a vital role upon identifying a suitable object of interest. The formation will consider the number of units in the swarm, the requirement to cover the object of interest from various angles and then apply a suitable formation to the swarm. This scenario is presented in the Fig 13, where the hidden tank (in the middle of the scene) acts as the object of interest and has been covered by the various units in the swarms from different angles.

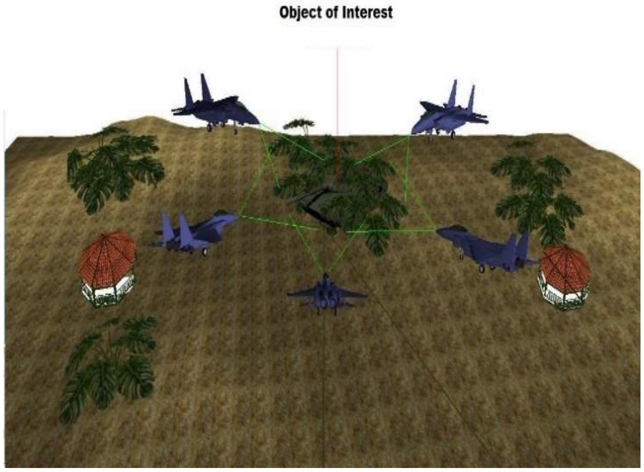


Fig. 13. Demonstration of Dynamic Formation.

These dynamic formations are essentially objective driven. The objective may vary from multi-angle surveillance of a specific object of interest to acquisition of sufficient multi angle data to construct a 3D representation of a tactical scene.

The dynamic transformation is governed by the following selected steps of the algorithm

Step 1: The required object within the tactical scene is scanned by each individual eye. This will generate the visual index of the individual eyes.

Step 2: The Unit with the highest visual index and closest to the required object will subtract the background from the object and then detect the centre point of the visible required object.

Step 3: The centre point of the required visible object is calculated and the distance from the centre point to a blind spot of the object is established.

Step 4: The calculations above will allow for the selection of swarm units for repositioning in order to increase their visual index factor and the overall collaborated visual coverage of the area.

These steps are summarised in the block diagram shown Fig 14. They have been implemented as communication between various classes within the system.

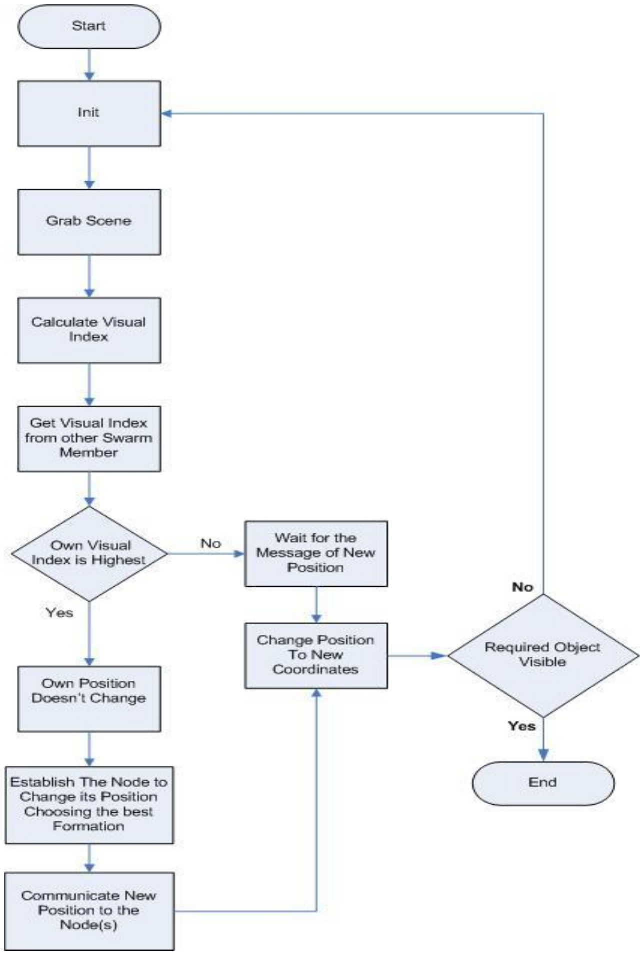

Start
Init
Grab Scene
Calculate Visual Index
Get Visual Index from other Swarm Member
Own Visual Index is Highest
No
Yes
Wait for the Message of New Position
Change Position To New Coordinates
Own Position Doesn't Change
Establish The Node to Change its Position Choosing the best Formation
Communicate New Position to the Node(s)
No
Required Object Visible
Yes
End

Fig. 14. Block Diagram of Dynamic Formation Algorithm.

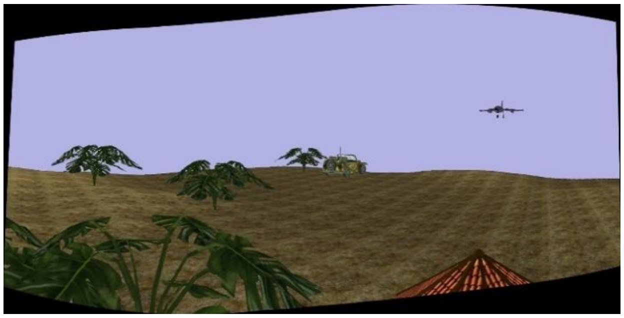


Fig. 15. Panoramic View of Swarm Curve Formation.

A swarm unit dynamically operates in a particular formation. The position and orientation of an individual swarm member within a formation affects its individual view and the overall stitched panorama. Different swarm formation coverage of a particular area of interest may yield a different type of panoramic view. This section presents some interesting panoramas created by stitching individual views of Swarm Eye units in different formations.

The panorama in Fig 15 is created by stitching individual views of swarm units in a curve formation. This panoramic image is based on individual views of 15 swarm units. The formation and orientation of the individual swarm is presented in Fig 16.

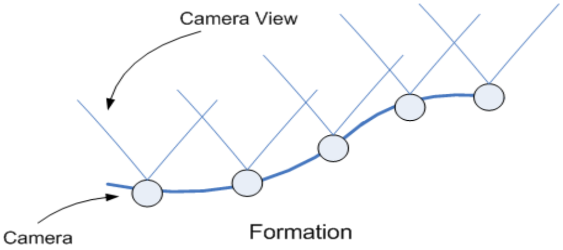


Fig. 16. Orientation of Swarm units in Curve Formation.

The collaborative viewing capability of the swarm allows for coverage of wide areas for an aerial surveillance perspective. A downward camera orientation within an arrow formation can be used to generate the collaborated view of an even wider area. An example of this case is shown in Fig 17. The panoramic image was generated by the swarm in an arrow formation shown in Fig 18 with a downward camera orientation towards the ground. This picture provides the vital information of the ground beneath the aircrafts.

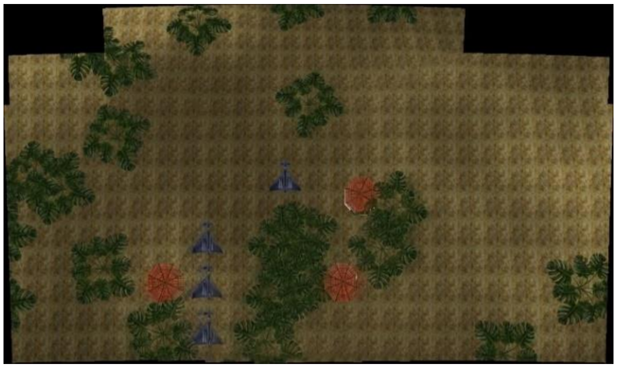


Fig. 17. Panoramic View of Swarm Arrow Formation.

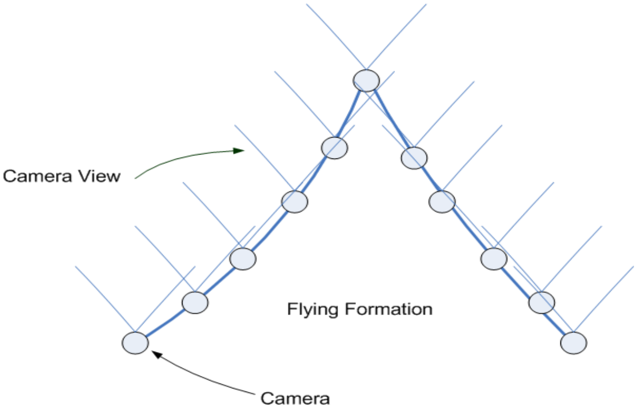


Fig. 18. Orientation of Swarm Units in Arrow Formation.

Several types of UAV can hover remaining still in the air allowing them to form formations that are usually not possible with wing based UAV. For instance helicopters can rotate in the air and this allows them to capture 360º degree coverage of the area of surveillance.

Furthermore, the flying capability allows them to form a circular formation. It is also possible to change the camera orientation within a circular formation to focus towards the centre of the circle. This allows for circular angular coverage of an area of surveillance and stitching individual views can generate spherical panoramas. The semi-circle formation shown in Fig. 19 allows for multi-angle coverage of the surveillance area. A full circle formation can also aid in acquisition of 3D information about the object of interest within a scene, see Fig. 20. The quality of the generated panorama is dependent on several factors such as overlap, view orientation and the algorithm used to stitch images together. In certain instances a particular algorithm may fail to generate a panorama if any of the required parameters of the algorithm are not met.

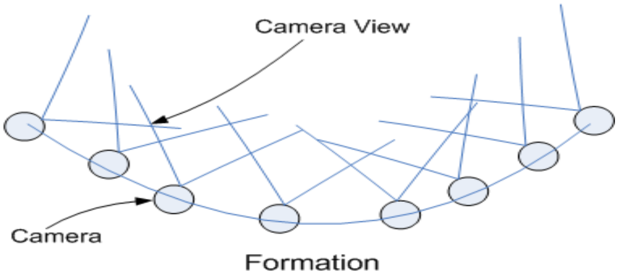


Fig. 19. Orientation of Swarm Units in Arrow Formation.

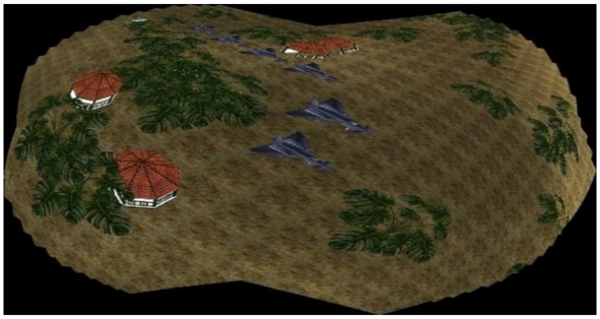


Fig. 20. Panoramic View of Circular Formation.

V. Conclusions

Distributed ground teams or swarms of UAVs can provide different and dynamic views of tactical scenes. Swarms are sets of interconnected units that can be arranged and coordinated in any flexible way to execute a specific task in a distributed manner. Swarm Eye provides a simulation environment to explore the potential of swarms and its merger with various advanced technologies to aid in surveillance and reconnaissance missions.

The identification and implementation of the proper formation of a swarm is a vital element for efficient swarm applications. The formation is derived by a suitable algorithm that is applied on the swarm members to achieve a particular orientation of these members within the swarm. The initial formation refers to the formation which will be adopted by the swarm at the start up. As soon as the unit will be airborne and get registered into the swarm they will form a formation to fly. The dynamic formation will be adopted by the swarm if the required objects are not visible due to the blind spots in the current formation.

In this paper, a formation algorithm has been proposed which demonstrates that blind spots can be minimised to get a better view of the target object or terrain. This formation algorithm aims to derive a suitable formation based on a visibility factor index. This index refers to a quantifiable unit that can measure visibility with regards to a particular object of interest within an area of surveillance. A higher visibility of the object of interest yields a higher visibility factor index. This quantified visibility unit is then used in an algorithm to apply a suitable formation to the swarm units.

The objective of the algorithm is to apply a suitable formation to the swarm so that maximum collaborated visibility is achieved. The algorithm divides a swarm into several sub-swarms applying suitable formations at both the sub-swarm and the overall swarm level. This approach allows for individual sub-swarms within a main-swarm to have their own formation. The division of a whole swarm into sub-swarms is very effective; by having sub-swarms, the user can allocate different jobs to different groups and individual sub-swarms are easier to manage as compared to managing whole swarms with hundreds of units.

The Swarm Eye consists of a set of independently performing eye units. In order to perform the task in a group each individual needs to have a suitable communication protocol to allow for communication between the units. The messages between the units would be different types of data i.e. images, simple 3D coordinate and messages /commands. Individuals in swarms interact either directly by communication or indirectly through joint behaviour. This coordinated behaviour allows for deployment and experimentation with the swarm of variable sizes. The proposed simulation platform provides an extensible framework for further exploration and experimentation with the autonomous swarm operations and formations with regards to tactical decision making. The self-formation capability using visual factor index allows for autonomous swarm formation and aids surveillance with reference to a specific object of interest.

VI. Future Works

The article has presented a novel approach for the swarm formation based on VFI. This approach allows dynamic swarm formations with the focus of maximum coverage of an object of interest. This approach can aid in tactical decision making and allows for autonomous swarm behaviour. This research study was constrained by time and budge and further research development can be carried on in the following points.

- The initial formation algorithm has been implement ed but the dynamic algorithm implementation is to be undertaken.
- The optimised implementation of the swarm formation algorithm with the integration of other collaboration algorithms such as (collision avoidance, obstacle avoidance, target seeking and formation keeping) at present the Swarm Eye system does not fully implement these methods.
- Implementation of image analysis by swarm unit to help tracking, detecting the target and form dynamic form ation by applying the swarm formation algorithm.
- The analysis of real time constraints and time delay of the Swarm Eye formation building and processin g of the visual data by the swarm units could be imple mented and tested on real units rather then simulation.
- The UAV physical manoeuvring constraints to be catere d by the algorithm in more detail.

The swarm formation is a considerable new topic and minute research has been performed. The area of swarm formation, coordination of multiple autonomous flying vehicle offers diverse edges of Swarm and still need further exploring and development.

The Swarm Eye project is open to many applications; this project provides the foundation of the Swarm Eye concept and could also benefit different industries.

Acknowledgment

This work is part of a research degree program at Cranfield University. The authors would like to thank an industrial sponsor, Lockheed Martin Corporation of this project for financial and technical support during the course of this study.

References

[1] G. Struwe, "Spectral Sensitivity of the Compound Eye in a Moth. Intra and Extra Cellular Recordingds," Acta Physiologica Scandinavica, vol. 87, no. 1, pp. 63-68, 2008.

[2] L. Bishop, "An ultraviolet photoreceptor on a Dipteran compound eye," J.comp. Physiol, vol. 91, pp. 267-275, 1974.

[3] "Online," [Online]. Available: http://news.bbc.co.uk/1/hi/world/americas/4808342.stm..

[4] H. Xu, H. Guan, A. Liang and X. Yan, "A Multi-Robot Pattern Formation Algorithm Based on Distribution Swarm Intelligence," in Second International Conference on Computer Engineering and Applications, 2010.

[5] M. J. Mataric, "Designing emergent behaviours: from local interactions to collective intelligence," in Proceedings of the second international conference on From anomals to animats 2: simulation of adaptive behaviour, Cambridge, MA, USA, 1993.

[6] E. A. Oslen, C. W. Park and J. P. How, "3D Formation Flight Using Differential Carrier-phase GPS Sensors," in Institution of Navigation GPS Meeting, Nashville, TN, 1998.

[7] A. Ryan, M. Zennaro, A. Howell, R. Sengupta and J. K. Hedrick, "An Overview of Emerging Results in Cooprative UAV Control," in Proceeding of the 43rd IEEE Conference on Decision and Control, 2004.

[8] P. K. Wang and F. Hadaegh, "Coordination and control of multiple micro spacecraft moving in formation," The Journal of the Astronautical Sciences , vol. 44, no. 3, pp. 315-355, 1996.

[9] S. Miller, All Kind of Eyes, Marshall Cavendish, 2007.

[10] E. Buschbeck, B. Ehmer and R. Hoy, "Chunk versus Point Sampling: Visual Imaging in a Small Insect," Science, vol. 286, no. 5442, pp. 1178-1180, 1999.

[11] R. Koshy, R. Munasinghe and A. Davari, "A Design of a Composite Eye for Computer Vision," System Theory , pp. 284-288, 2008.

[12] T. Neumann, "Modeling Insect Compound Eyes: Space-Variant Spheric Vision," Lecture Notes in Computer Science, pp. 360-367, 2002.

[13] T. P. Breckon, S. E. Barnes, M. L. Eicher and K. Wahren, "Autonomous Real-time Vehicle Detection from a Medium-Level UAV," in in Proc 24th International Unmanned Air Vehicle Systems, pp. 29.1-29.9, 2009.

[14] J. Tisdale, Z. Kim and J. Hedrick, "Autonomous UAV path planning and estimation," IEEE Robotics and Automation Magazine, vol. 16, no. 2, p. 35, 2009.

[15] G. Heredia, F. Caballero and I. Maza, "Multi-UAV Cooperative Fault Detection Employing Vision-Based Relative Position Estimation," in 17th World Congress, The International Federation of Automotic Control, Seoul, 2008.

[16] R. Fernald, "Casting a genetic light on the evolution of eyes," Science, vol. 313, no. 5795, pp. 1914-1918, 2006.

[17] M. J. Krieger and J. B. Billeter, "The call of duty: Self-organised tast allocation in a population of upto twelve mobile robots," Robotics and Autonomous Systems, pp. 65-84, 2000.

[18] B. R. Cobleigh and J. L. Hansen, "Induced Moment Effects of Formation Flight Using Two F/A18 Aircraft," in NASA Dryden Flight Research Centre, California.

[19] P. Binetti, K. B. Ariyur, F. Bernelli and M. Kristic, "Formation Flight Optimisation Using Extremum Seeking Feedback," Journal of Guidance Control and Dynamics, vol. 6, no. 1, 2003.

[20] A. Gopalarathnam, Aerodynamic Benefit of Aircraft Formation Flight, Encyclopaedia of Aerospace Engineering, 2010.

[21] D. Shreiner, M. Woo, J. Neider and T. Davis, OpenGL(R) Programming Guide: Version 2, 5th Edition, Addison-Wesley Professionals, 2005.

[22] P. Pacheco, Parallel Programming with MPI, Morgan Kaufmann, 1988.

[23] P. Burelli, L. Di Gaspero, A. Ermetici and R. Ranon, "Virtual Camera Composition with Particle Swarm Optimisation," Lecture Notes in Computer Science, vol. 5166, pp. 130-141, 2008.

[24] Q. Zhang and H. Li, "MOEA/D: A Multi-objective Evolutionary Algorithm Based on Decomposition," Evolotionary Computation, IEEE Transactions, vol. 11, no. 6, pp. pp. 712-731, 2008.

[25] R. Roy and J. Mehnen, "Technology Transfer In Studies in Computational Intelligence," Evolutionary Computation in Practice, vol. 88, pp. pp. 263-281, 2008.

[26] J. Sokalski, T. P. Breckon and I. Cowling, "Automatic Salient Object Detection in UAV Imagery," in in 25th Bristol International UAV System.

[27] M. Duong, P. Cong, H. Quach, T. H. Dinh and Q. Ha, "Enhanced Discrete Particle Swarm Optimization Path," Automation in Construction, vol. 81, pp. pp 25-33, 2017.

A Review of Data Synchronization and Consistency Frameworks for Mobile Cloud Applications

Yunus Parvej Faniband[1], Iskandar Ishak[2], Fatimah Sidi[3], Marzanah A. Jabar[4]
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM, Serdang, Selangor Darul Ehsan, Malaysia

***Abstract*—Mobile devices are rapidly becoming the predominant means of accessing the Internet due to advances in wireless communication techniques. The development of Mobile applications (“apps”) for various platforms is on the rise due to growth in the number of connected devices. Numerous apps rely on cloud infrastructure for data storage and sharing. Apart from advances in wireless communication and device technology, there is a lot of research on special data management techniques that addressed the limitations of mobile wireless computing to make the data appear seamless for accessing and retrieval. This paper is an effort to survey the frameworks that support data consistency and synchronization for mobile devices. These frameworks offer a solution for the unreliable connection problem with customized synchronization and replication processes and hence helps in synchronizing with multiple clients. The frameworks are compared for the parameters of consistency and data models (table, objects or both) support along with techniques of synchronization protocol and conflict resolution. The review paper has produced interesting results from the selected studies in areas such as data consistency, handling offline data, data replication, synchronization strategy. The paper is focused on client-centric data consistency and the offline data synchronization feature of various frameworks.**

Keywords*—Mobile cloud computing; data consistency; mobile back-end as a service; distributed systems; mobile apps

I. Introduction

The model of mobile cloud computing utilizes the services of cloud computing. The mobile cloud environment consists of portable computing devices, mobile Web and location-based services, supported by wireless communication infrastructure, to provide mobile devices online access to large storage space and unlimited computing power [1]. A wireless network with mobile clients is fundamentally a distributed system but suffer from the primary challenges such as limited computational capacity and storage of mobile devices, intermittent loss of connectivity and battery power restrictions. The transmission bandwidth of the mobile device is likely to be lesser than the transmission bandwidth of the mobile support stations (MSSs) and this leads to the phenomenon of communication asymmetry. The effective management of data in systems with the mobile client is affected by these limitations. The environment of frequent disconnections and limited bandwidth, impact the data and transaction management as well as the data consistency guarantees.

To provide the illusion of uninterrupted data access, the data management must hide the constraints of mobile wireless computing. The technique of replicating data locally on the mobile device enables the user to carry offline data without the need to always be connected to the data server. The ability to disconnect with the network, do local changes, and then reintegrating (synchronize) these changes back into the system makes the mobile gadget an essential extension to modern distributed databases and collaborative tools [1].

Data synchronization is an empowering process that eliminates the critical requirement of having steady connectivity and permits users to run data-centric mobile applications while being offline. Hence data synchronization allows users to carry out operations of additions, deletions, and updates on the offline data, while disconnected from the network.

Generally, the mobile applications (which we generally refer to as ‘Apps’) are developed according to the different application programming interfaces (API) abstractions supported by the underlying mobile middleware. The middleware may provide a simple file-based API (possibly extended with replication-specific methods). It may also support complex abstraction such as objects, tuples, relational entities or an object which may contain pointers to other interdependent objects. Middleware with database replication primarily provides query oriented CRUD APIs(Create, Read, Update and Delete) to application developers for typical operations on data with declaratively defined by SQL queries for the update, creation/insertion, and deletion of records.

This section covers the introduction to replication, data-centric and client-centric consistency models. Section II and III classify and describe various consistency and synchronization frameworks in mobile cloud computing. In Section IV we discuss research findings and recommendations followed by related work (Section V) and conclusion with future work (Section VI). Table I shows the list of acronyms used in the paper.

A. Mobile Computing Environment and Limitations

Generally, a mobile cloud computing environment has two unique sets of entities namely Fixed Hosts (FH) and Mobile Hosts (MH) [1]. FHs are machines (Works stations and Servers) with efficient computation power and reliable storage of data and run large databases. FHs that are connected through the fixed network. MHs with limited processing and storage power(cellular phone, palmtops, laptops, notebooks) are not continually communicating with the fixed network. They may be disconnected for various reasons such as due to the battery power saving measures and also due to disconnections during frequent relocation between different cells. Additional dedicated fixed hosts called mobile support stations (MSSs) acts

TABLE I. LIST OF ACRONYMS.

Symbol	Description
C	Causal Consistency
CMP	Consistency scheme with PRACTI property [2]
CRDT	Conflict-Free Replicated Data Types
HU	Hoarding Unit
E	Eventual Consistency
FSC	Fork-sequential consistency
MRC	Monotonic read consistency
MWC	Monotonic write consistency
PRACTI	PR - Partial Replication, AC - Arbitrary Consistency, TI -Topology Independence
RAWC	Read after writes consistency
RYWC	Read your writes consistency
S	Strong Consistency
SC	Sequential Consistency
T , O	Table , Objects
WFRC	Write Follows Read Consistency

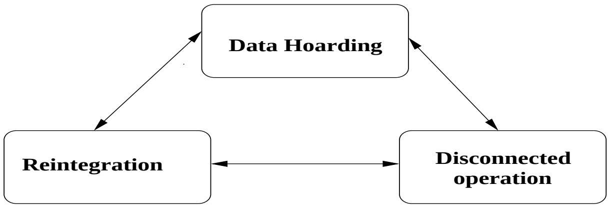


Fig. 1. States of Disconnected Operation.

as the channel between the FH and MH through wireless LAN (local area network) connections, cells or connections to the network with standard modems.

When the network connectivity becomes unavailable or unacceptable, the MH enters the disconnected state. Disconnected operation (see Fig. 1) is a three-stage changeover between the following states [3]:

1) Data hoarding: This is the process of preloading or prefetching the data in anticipation of a foreseeable disconnection. Before going to offline mode (disconnection), the data structures necessary for operation during disconnection are either replicated (catched) or moved (partitioned) at the MH.
2) Disconnected operation: When the MH is offline (disconnected from the network), data might be changed, added or even removed at either the MH or the FH.
3) Synchronization or Reintegration: When the connection is reestablished, each operation executed at the MH should be synchronized (reintegrated) with appropriate updates executed at other sites in order to attain seamless consistency.

For a given distributed system, the complexity of operations in each of the above three states is determined by the interdependence of data operated on. The issues pertaining to three states is summarized in Table II.

B. Replication

Replication is a basic strategy to support fault resilience, high data availability and quick response for universally available services. Replication process creates many instances of the identical object in different machines, over a distributed or local network ([3], [4], [5], [6], [7], [8]). The copies of these multiple objects (replicas) are persistently maintained over time in order to allow the workload to be divided over the possible number of replicas. The replication strategy involved in a different distributed system depends on the requirements of data freshness tolerance. The use cases in some applications need only read operations, while others high ratio of writes(updates) compared to read. Banking systems require that data is always consistent over time and some social networks may tolerate stale data.

TABLE II. ISSUES IN DISCONNECTED OPERATION.(SOURCE [1])

State	Problem	Resolution
Hoarding	Unit of caching/hoarding	System dependent (e.g. a file or a database fragment)
	Which items to cache (hoard)?	Application dependent , based on purpose of the system Defined distinctly by the user Generate from the knowledge of past operations
	When to execute hoarding ?	Based on regular intervals Before disconnection
	Call for locally unavailable data	Add requests to queue for future service Raise an exception/error
Disconnection	What to log?	Timestamps Data Values Operations
	When to optimize the log?	Before synchronization Incrementally
	How to optimize the log?	System dependent
	How to synchronize?	Re-execute an operational log
Reintegration or Synchronization	How to resolve conflicts?	Automatic resolution Use application-semantics Provide utility to aid the user

C. Consistency Models

The literature [9] [10] describes data-centric and client-centric consistency as the two principle viewpoints on consistency. The data-centric consistency manages the internal state details by guaranteeing that all the replicas are same and ensures system maintains consistency for updates. Data-centric consistency is important to system developers. Client-centric consistency deals with only observing data updates as a black box and hence application developers focus on client-centric guarantees. Ordering and Staleness are the two criteria for measuring guarantees of both data-centric and client-centric consistency. Staleness is measured in the unit of time (t-perceivability) or versions (k-staleness), calculated based on how much a given copy is falling behind [11] [12].

D. Data-Centric Models (Server-Side Consistency Models)

1) Strong consistency - A system adopting a strong consistency model is in a consistent state all times. The strong consistency is a single-copy consistency model that is not suitable for mobile applications dealing with cloud data due to the availability and performance issues as mandated by CAP theorem [13].
2) Sequential consistency - This is a slightly weaker form of strong consistency with the condition that, same order of execution is maintained for all the sequentially related requests. Subsequently, the clients observe the same order and sequence of updates.

3) Causal consistency - In a system adopting the Causal Consistency (CC), the same sequence of execution is maintained on all replicas, for all the causally related requests. The non-related requests are followed in random order.
4) Grouping operations - This model deals with handling the cases of, series of reading and write operations. The Grouping Operation model allows raising the level of granularity to span multiple reads and writes, into an atomically executed unit.

E. Client-Centric Models (Client-Side Consistency Models)

1) Weak consistency - A weak consistency model does not guarantee that subsequent accesses will return the updated value. The term 'inconsistency window' [10] attribute to the time between the update and the instant when it is guaranteed that any observer will always see the refreshed value.
2) Eventual consistency - Eventual consistency is considered as another model of Weak consistency with an added guarantee that when no new updates are made to an object, eventually all replicas will see the last. Eventual consistency provides the following four main ordering guarantees [14]:
    a) Monotonic Read Consistency (MRC) - In this model after reading a version 'n' of an object, the same client will never access a version less than 'n' on a subsequent read.
    b) Read Your Writes Consistency (RYWC) - In this model, after writing version 'n', the same client will never again read an older version less than 'n'. This is a unique case of the causal consistency model [9] [10].
    c) Monotonic Write Consistency (MWC) - In this model, all writes by the same client guaranteed to be serialized in the order of time of update It guarantees that a write operation is always ended prior to any subsequent write operation on the same data item [9] [10]
    d) Write Follows Read Consistency (WFRC) - It guarantees that an update succeeding a read of version 'x' will never be carried out on replicas that are prior to version 'x' [9].

The studies [15] [16] conclude that it is mandatory to guarantee all four client-centric models (MRC, RYWC, MWC, and WFRC) for the system to achieve client-centric consistency.

II. Classification of Consistency, Synchronization and Replication Systems in Mobile Computing

This section classifies the current efforts into different types such as systems for weakly connected clients, sync services and systems supporting geo-replication. The studies are also classified based on three PRACTI properties [2].

A. Systems for Weakly-Connected Clients

Many previous attempts have dealt with data replication and management in systems where mobile clients intermittently connected either to servers or to peers ([3], [4],[17], [2], [18], [19]). Systems like Coda [3] and Ficus [19] address the issues in handling disconnected operations and replicate files providing high availability at the cost of consistency. Bayou [4] is a distributed relational database system that provides eventual data consistency, under offline mode. These systems differ on their procedures to handle conflicts. For instance, Bayou performs application-level conflict resolution, while Coda and Ficus allow system level resolution of conflicts. Some Systems (like Simba [20]) are aimed to provide more control for mobile applications to select suitable consistency abstractions for data synchronization services.

There are several studies which explicitly focus on the efficiency of data management systems for weakly connected clients ([21], [22], [23], [24]). In compliance to different requirements of apps, Odyssey [23] system give OS support for applications to modify the fidelity of their data to accommodate resource changes, such as wireless network bandwidth fluctuations and battery conditions. Cedar [24] increase the query processing capability by identifying the commonality between client and server query results and hence provides productive mobile database access. In LBFS [21] (low-bandwidth network file system), the content-based chunking technique prevents redundant transfer of files and also detect inter-file similarities.

B. Geo-Replication

There are several studies which focus on the tradeoff between consistency, availability, and performance for geo-distributed services. These system handle data replications within and across and data centers. Some systems primarily aimed at providing low-latency causal consistency at scale (e.g. , COPS [25] and Eiger [26]) and others (e.g., Red Blue consistency [27], Walter [8], Transaction Chains [28], and) focus to reduce the latency involved in supporting other forms of stronger-than-eventual consistency, including serializability under limited conditions. Arbitrary consistency selection systems (e.g., Pileus [29] and SPANStore [30]) attempt to provide more control for applications to choose suitable consistency across data centers, to meet SLAs or to minimize operating costs.

C. PRACTI Paradigm

In a distributed system, an optimal replication system should support all the three PRACTI [2] properties. 1) PR-system (Partial Replication) allow any node to store a subset of data and metadata. 2) AC-system (Arbitrary Consistency) provide flexibility of selection of consistency semantics (different types of configurable consistency guarantees like both strong and weak consistency) for applications. 3) The TI-systems (Topology Independence) permit all nodes to send updates to all other nodes (TI).

Applying PRACTI taxonomy to the current studies, the existing replication systems fall into the following four protocol groups. Each system compliant to most two of the PRACTI paradigm properties.

1) Server replication: Some systems use the log-based peer-to-peer update exchange protocol for server-side replication. This protocol follows full replication mechanism and allow all nodes to store complete data from any volume and also all nodes collect

all updates. This protocol helps to achieve topology independence (TI) in some systems (e.g., Bayou [4] and Replicated Dictionary [31]), where any node to send updates to any other node. Some Systems like in TACT [32] and Lazy Replication [33]) use this protocol to provide more control to select suitable consistency guarantees for data synchronization (AC). Since the protocol does not support efficient network usage due to full replication, these systems may be not suitable for devices with limited resources.

2) Some systems with client-server architecture (e.g., Coda [3] and Sprite [34]) and hierarchical caching systems (e.g., hierarchical AFS [35]) implement a protocol to selectively replicate/cache arbitrary subsets of content (PR). Apart from supporting a group of consistency policy by the system, a supplementary extension of consistency guarantees are provided by changing the basic architecture (AC). In order to support consistency, the partial replication protocols need intercommunication between a child and its parent and also serialize control messages at the central server node [36]. Due to these communication complexities, the performance, availability and data sharing features may be paralyzed in such systems.
3) In the Distributed hash table (DHT)-based storage systems (e.g., PAST [37], CFS [38] and BH [39]), the scalability is achieved by load balancing the server across various nodes, on a per-object or per-block basis. For high availability, the data is also replicated to multiple nodes and such architecture becomes challenging for providing the consistency guarantees.
4) Object replication systems (e.g., WinFS [40], Ficus [41] and Pangaea [42]) permit nodes for selective replication/caching of arbitrary subsets of data (PR) and communication with every other peer (TI).These protocols lack consistency guarantees since they do not mandate ordering constraints on updates across multiple objects.

D. Synchronization Service Frameworks

The existing services mainly offer sync services into three categories: (1) File-only, (2) Table-only and combination of (3) Table and Object. Izzy [43] and Mobius [44] sync services provide a platform for structured data like tables only, to expedite the development and deployment of data-centric mobile apps. Mobius guarantee that all clients observe write operations in the same order, maintaining the flexibility of local client views to diverge (fork-sequential consistency [45]).Dropbox sync service provides dedicated API for tables and do not store files and tables together. Many Mobile apps are developed using the file sync services of Google Drive [46], iCloud [47], Dropbox[48] [49] and Box Sync[50]. StackSync [51] is an open-source Personal Cloud framework that provides scalable file synchronization and sharing. QuickSync [52] is a system that focuses on improving the synchronization performance of cloud storage, in wireless networks depending on network conditions.

Sapphire [53] is a cloud-enabled distributed programming platform for mobile and cloud applications. Sapphire makes a smooth application execution using the techniques of code-offloading, caching, and fault-tolerance. Sapphire lacks in data management services but provides smooth application execution. The work on Pebbles[54] revealed that apps massively depend on structured data (table) to manage unstructured objects (files). Simba [20] extended the table interface of Izzy [43] to provide a unified abstraction for both table and object, the benefits of which are explored the context of local systems in these studies [[55], [56]].

CouchDB [57] is a schema-free “document store” supporting eventual consistency and provide “document” sync with coordination from its client TouchDB [58].

SwiftCloud [59] and Cloud types [60] provide cloud-enabled programming interface to facilitate the mobile apps for storing local replicas of data on the devices and subsequently sync with the cloud servers. The programmer needs to handle synchronization in SwiftCloud and Cloud types, while Simba permits automatic synchronization.

Mobile operating systems provide some kinds of data storage abstractions to developers. Apple expanded its iCloud [47] service with CloudKit [61], a new means for applications to store and access data stored in iCloud [47]. There are some open source mobile back-end-as-a-service offering, such as Parse Server platform [62] and StackSync [51].

Many commercial services provide back-end cloud storage services to link mobile and web applications to the cloud, such as IBM Bluemix Mobile Cloud Service [63] [64]. Services of Firebase [65] and Kinvey [66], also aid app developers to connect their apps to cloud backend.

III. Discussion on Literature of Consistency and Synchronization Frameworks

The existing literature from the database community and distributed systems community focus on consistency models, their implementations and their measurement. This paper focuses on the reference implementations helping the mobile clients for end-to-end data consistency and data synchronization service utilizing the cloud resources. The literature has case studies investigating the difficulties related to consistent replication across mobile devices with intermittent network connectivity and bandwidth constraints. Some studies in the literature address the frameworks designed to handle the current constraints in Mobile app development.

Coda [3], was one of the initial client-server architecture systems, to emphasize the difficulties in addressing the offline operations. BlueFS [67] is another system that focuses on energy efficiency in resource-constrained mobile devices. Bayou [4] is based on client-server architecture and supports a disconnected system and provides a programming interface to application-specific conflict detection and resolution to handle optimistic updates (eventual consistency). Odyssey [23] support application-aware adaptation based on type-specific operations. The Rover [68] toolkit is a client-server, mobile applications development platform that relocatable dynamic object (RDO) and queued remote procedure call (QRPC) for data communication.

Simba [20] provides end-to-end data consistency framework with a data abstraction for a combination of tabular and object data models. Additionally, the applications written to this abstraction are allowed to select from a set of distributed consistency schemes and sync data with the cloud. Simba Server implementation of data Storage use OpenStack Swift [69] for object data and Cassandra [6] to store tabular data. Simba configure OpenStack Swift and Cassandra to utilize three-way replication, in order to achieve high availability. Also, the framework mentioned in [70] support using Cassandra as a backend datastore. Our work [71] proposes to extend Simba with support for large data objects.

Mobius [44] is designed as a cloud-enabled data replication and messaging platform for the mobile applications. It provides table consistency and uses PNUTS [7] as the back-end store.

A middleware framework for a mobile network that performs reliable and real-time data synchronization is proposed by Xue [72]. Izzy [43] and Mobius [44] frameworks provide a platform for structured data like tables only, to expedite the development and deployment of data-centric mobile apps. Simba [20] is built upon the sync framework of Izzy and supports cloud-based data synchronization service, which reduces development complexity of mobile apps.

Cimbiosys [17] is a peer-to-peer system platform (clients share updates directly with each other) that enables various apps to manage cloud-based data on personal computers and mobile devices. Perspective [73] is another platform like Cimbiosys that use filters for selective replication of data on mobile devices. PRACTI [2] is a unique replication system that supports all the three ideal PRACTI properties of partial replication, arbitrary consistency and topology-independence.

Currently, researchers are proposing new principles to deal with weak consistency. Strong correctness guarantees can be achieved without the use of costly global synchronization when all operations in a program are purely monotonic. Built on this monotonic principle, some data structures like sets and sequences can be correctly replicated without the need of synchronization.

The Conflict-Free Replicated Data Types (CRDTs) ([74] [75] [76]) are asynchronous data types that do not need synchronization for updates. They comply Strong Eventual Consistency Model [75] and can be utilized to build other data models, required by applications. Asynchronous quality of CRDTs makes it more qualified for replication in eventual consistency environments.

More recently researchers utilize these special data types (CRDTs) to build the frameworks using Key-Value stores. Riak [77] distributed database is used as a back-end store by systems like SwiftCloud [59] to implement a Key-CRDT. In order to support strong eventual consistency, the SwiftCloud middleware, convert a Key-Value store in a Key-CRDT store, into a data-model that utilize properties of CRDT. The system allows clients to execute updates concurrently without synchronization. By executing automatic conflict resolution specified in CRDTs, the systems guarantees the clients with zero conflict for simultaneous updates. Walter [8] and Gemini [27] are other systems that use CRDT for providing eventual consistency. Indigo [78] enhance SwiftCloud, wherein an application specifies the invariants, or consistency rules, that the system must maintain.

Consistency As Logical Monotonicity (CALM) is another technique used in built consistency frameworks. According to the CALM theorem, logically monotonic programs are guaranteed to be eventually consistent without the requirement of any coordination protocols (distributed locks, two-phase commit, paxos, etc.). Hence CALM approach ensures eventual consistency by necessitating a monotonic logic [79]. In logic languages (e.g. Bloom[80]) CALM analysis helps to analyze whether the code flow is sufficient towards consistency without the integrating co-ordination protocols [79].

The study claimed [81] that the use of revision diagrams along with special abstract Cloud Types is a useful technique for eventually consistent distributed programs. Revisions diagrams are semi lattices designed for the context of multiple versions and eventual consistency and work same as the version control systems. In this approach, the distributed state is stored using special cloud abstract data types. These Cloud types expose interface with well defines update and query operations [60]. Cloud types provide eventually consistent storage and hide the complex backend implementation details of network and coordination protocols. They offer the functionality to perform the optimized fork and join implementations and storing of updates in the form of logs [60]. The prototype implementation of this technique is available in TouchDevelop language and as a library in C# [82]. While the CRDTs help to carry out only commutative operations, the cloud types support non-commutative operations still accomplishing eventual consistency.

Open Data Kit (ODK) 2.0 [83] support to build Android-based application-specific information modules for offline operations. StoArranger [84] is another system framework that aid the programmers to manage cloud data storage on mobile devices by addressing issues of rearranging, and coordinating cloud storage communications. BlueMountain [85] is a modern mobile data management platform supporting solutions for file and database management, which allow to achieve wider deployability and help app developers to spend more efforts on app logic. Unidrive [86] is a client-side middleware system which can integrate multi-cloud capabilities to mobile apps. CacheKeeper [87] allows caching of browser data on mobile devices using system-wide, kernel level caching support for mobile applications.

Parse [62] is a back-end as a service platform that uses MongoDB as the back-end datastore. Parse platform allow the developer to create loosely or strongly typed objects and easily save, update, query, and delete these in a backend data store.

Mobile apps are developed using the file sync services of Google Drive [46], iCloud [47], Dropbox [48] [49] and Box Sync[50]. StackSync [51] is an open-source Personal Cloud framework that provides scalable file synchronization and sharing. QuickSync [52] is a system that optimizes cloud storage synchronization performance in wireless networks based on network conditions. IBM Bluemix Mobile Cloud Service [63] [64] provides back-end cloud storage services to link mobile and web applications to the cloud. Other commercial platforms such as Kinvey [66] and Firebase [65], help app developers to connect the apps to cloud backend.

TABLE III. COMPARISON OF THREE REFERENCE IMPLEMENTATIONS.

Reference Design	Strength	Weaknesses
Simba [20]	- Allow apps for the programmatic delay tolerant data transfer - Uses a single persistent TCP connection to the cloud data , resulting in bandwidth saving	Since multiple apps access the same instance of client, certain poorly written apps may adversely affect other Simba apps
Mobius [44]	- All in one solution with a combination of messaging and data platform - Linear scalability for number of applications, users and size of data	Can be improved in the area of cross-app synchronization, optimization strategies and caching
SwiftCloud [59]	- Allow execution of transactions in the client side as well as at the data centers - Efficient use of caching methods, executing both reads and updates at the client	- Lack support for combined weak and strong consistency, and for object composition - DC implementation is not modular

IV. RESEARCH FINDINGS, DISCUSSION AND RECOMMENDATIONS

This section discusses the selected case studies, based on the criteria to understand the different technologies used for building the frameworks. SwiftCloud uses the CRDT with the Riak key store. Mobius uses PNUTS distributed database and supports P2P communication model. Simba supports configuring different consistency levels using Cassandra and OpenStack Swift object storage. TouchDevelop library utilizes the Cloudtypes using the Revision diagrams. BloomL library covers the BloomL language supported framework. Due to space constraints, we are only covering the three frameworks Swiftcloud, Simba and Mobius. These reference solutions are aimed at providing data replication, synchronization, and offline services to ease the development complexity of mobile apps. The solutions use the client side caching technique to offer offline services. The solutions are backed by the cloud storage to store the data. Table III summarizes the strength and weaknesses of the studied three reference implementations. Table IV (See Appendix) summarizes the consistency and data models (table, object or both) support in the various reference implementations. Table IV also lists PRACTI property supported by each framework along with mechanism of synchronization protocol and conflict resolution.

A. Synchronization Services

Some of the solutions provide the sync services for structured data like table only (Mobius and SwiftCloud). Simba supports both tabular and objects data models. Synchronization operation execution required to be handled by the programmers in case of Mobius and Swiftcloud. In contrast, Simba supports automatic synchronization process in the background.

B. Consistency Support

In order to satisfy the diverse consistency needs the frameworks should support different types of data and independently define their consistency. Mobius provides per-record sequential [88] and fork-sequential [45] consistency through the exclusive type of read operations. Simba provides three consistency semantics, resembling strong, causal and eventual consistency. The extent of consistency specification permit may be per-row or per-request, per-table.

Caching Policy and Offline support: The strategies of caching (replication) data at the client side enable higher availability and improve latency. Caching policies need to take care of the consistency semantic (ordering, updates and fetching of fresh updates). Solutions provide options to access data from client-side storage or remotely. Cache policies can be determined by the server-side back-end. The server-generated policies can be context-aware, globally configurable and dynamic. These policies are created based on run-time usage or access patterns of all users collected from each application. Efficient write caching capabilities group possible numbers of write operations in a one network message to reduce bandwidth. A Prototype of Mobius clients uses the trained decision tree model (policy selector) to determine whether to fetch locally or remotely. Mobius uses cost-sensitive decision tree classifiers to write batch updates. In SwiftCloud the clients can access the causally- consistent view of the stable version of data (cached at multiple servers). In Mobius, MUD tables are partitioned across mobile nodes and one or more server backends. Data access during offline is from the local tables. The write updates are stored locally and forwarded to the backend on reconciliation of client. During offline, reads are delivered from the local scout in SwiftCloud. Scout cache handles the write updates. On network availability, finally, they will be committed at its DC.

C. Limitations of Reference Frameworks

Even though incredible researches have been done in providing end-to-end data consistency solutions, many challenges still remain. This section points out some of the challenges that are needed to be addressed in various reference frameworks.
For app developers, currently, Mobius [44] provide higher level APIs (blocking or asynchronous) abstracted around the basic MUD APIs. The researchers propose the opportunity to support richer interfaces with the declarative query language. Mobius can be improved in the area of cross-app synchronization, optimization strategies and caching. Mobius can be improved in cache operations such as dynamic caching strategies, clearance policies and push-based cache maintenance. For bandwidth consumption and improving access, the outstanding updates stored locally should be compressed. There should be a smooth deterioration of response quality during disconnected operation. For the scalability improvements, the authors of Mobius [44] propose to improve partitioning schemes by adapting their earlier efforts on automatic and fine-grained partitioning ([89], [90]).

Simba's [20] sync protocol does not support streaming APIs to handle big size objects (e.g. Media file like Videos). Simba proposes to handle atomic multi-row transactions as prospective enhancement and currently support only atomic transactions on individual rows.

SwiftCloud [59] can be enhanced with a better caching mechanism and support for transaction migration. Also, better data encapsulation across software stack through API level, to address efficient data access.

V. Related Work

The work of [91] conducts an analysis of concepts of mobile client-server computing and mobile data access with a detailed review of early research prototypes (Bayou [4], Odyssey [23] and Rover [68]) for mobile data management. Our work extends this work by analyzing the consistency support for the latest frameworks. The survey of contributions on data dissemination and support for data consistency techniques for mobiles devices is discussed here [92]. The paper [93] compares and analyze the several contributions to models for mobile transaction. A survey of literature work on synchronization between the mobile device and server-side databases can be found here [94]. A survey of academic work on mobile/cloud computing can be found here [95]. The paper [96] conducts a comprehensive review of the data replication techniques in the cloud environments. Recent review article deals with the comparison of different categories of data synchronization algorithms based on scalability, consistency, accuracy parameters in ubiquitous network [97].

VI. Conclusion and Future work

In this paper, we presented a review of data consistency and synchronization frameworks in Mobile Cloud Computing for Mobile Apps. We considered the latest studies done from 2010 to 2017, and the advantages and disadvantages of three reference implementations in the literature have been presented. Then, the approaches to handle consistency support, sync services, conflict handling and offline operations in reference solutions have been discussed. Furthermore, out of the review, several findings and potential future works have been identified. We believe that this is an important research area, that will attract more contributions from the research community.

The Conflict free replicated data type, logically monotonic programs (CALM approach) and Revisions diagrams as semi lattices are some of the techniques used in these frameworks. Frameworks make use of the backend stores implemented using these technologies to support the data consistency features.

Out of the three frameworks explored, *Simba* is a superior framework ensuring three types of consistency guarantees (strong, causal and eventual consistency) for both table and objects data models. Simba reduces programmer's efforts as it supports automatic synchronization process in the background. Simba lacks multi-row transactions and streaming APIs to access to large objects.

Swiftcloud uses a client-assisted failover solution with CRDT store to support both mergeable and strongly consistent transactions. Programmers need to manage the synchronization process. It utilizes properties of CRDTs to support automatic conflict resolution. SwiftCloud needs to improve in providing efficient data access through APIs.

Mobius provides table consistency and uses PNUTS as the back-end store to support cloud-enabled data replication and messaging platform. Mobius provides per-record sequential and fork-sequential consistency through the exclusive type of read operations. Programmers need to manage the synchronization process. Mobius uses cost-sensitive decision tree classifiers to write the batch update. Mobius needs improvements in the area of caching and optimization strategies with richer client interfaces. It has to be noted that the literature review is limited by sources and keywords, terminologies used in the search, and the search date. Hence it is possible to include more relevant papers while replicating this study in the future. Our final outputs of this research are limited to the current availability of frameworks that address the data consistency, synchronization , and other features. While the current study did not deal with the full details of measurements of numerical deviation, order deviation and staleness (latency) of each framework, we intend to conduct detailed research with simulations on the comparison of these performance parameters for each platform.

Acknowledgment

Authors acknowledge Universiti Putra Malaysia and Faculty of Computer Science and Information Technology, Malaysia, for all the support. This research is funded by Ministry of Education of Malaysia under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2015/ICT04/UPM/02/07).

References

[1] E. Pitoura and G. Samaras, *Data management for mobile computing*. Springer Science & Business Media, 2012, vol. 10.

[2] N. M. Belaramani, M. Dahlin, L. Gao, A. Nayate, A. Venkataramani, P. Yalagandula, and J. Zheng, "Practi replication." in *NSDI*, vol. 6, 2006, pp. 5–5.

[3] J. J. Kistler and M. Satyanarayanan, "Disconnected operation in the coda file system," *ACM Transactions on Computer Systems (TOCS)*, vol. 10, no. 1, pp. 3–25, 1992.

[4] D. B. Terry, M. M. Theimer, K. Petersen, A. J. Demers, M. J. Spreitzer, and C. H. Hauser, "Managing update conflicts in bayou, a weakly connected replicated storage system," in *ACM SIGOPS Operating Systems Review*, vol. 29. ACM, 1995, pp. 172–182.

[5] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," in *ACM SIGOPS operating systems review*, vol. 41. ACM, 2007, pp. 205–220.

[6] A. Lakshman and P. Malik, "Cassandra: structured storage system on a p2p network," in *Proceedings of the 28th ACM symposium on Principles of distributed computing*. ACM, 2009, pp. 5–5.

[7] B. F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D. Weaver, and R. Yerneni, "Pnuts: Yahoo!'s hosted data serving platform," *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1277–1288, 2008.

[8] Y. Sovran, R. Power, M. K. Aguilera, and J. Li, "Transactional storage for geo-replicated systems," in *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*. ACM, 2011, pp. 385–400.

[9] A. S. Tanenbaum and M. Van Steen, *Distributed systems: principles and paradigms*. Prentice-Hall, 2007.

[10] W. Vogels, "Eventually consistent," *Communications of the ACM*, vol. 52, no. 1, pp. 40–44, 2009.

[11] M. Klems, D. Bermbach, and R. Weinert, "A runtime quality measurement framework for cloud database service systems," in *Quality of Information and Communications Technology (QUATIC), 2012 Eighth International Conference on the*. IEEE, 2012, pp. 38–46.

[12] D. Bermbach and S. Tai, "Eventual consistency: How soon is eventual? an evaluation of amazon s3's consistency behavior," in *Proceedings of the 6th Workshop on Middleware for Service Oriented Computing*. ACM, 2011, p. 1.

[13] A. Bradic, "The cap theorem : Brewer ' s conjecture and the feasibility of consistent , available , partition-tolerant web services," no. August, 2010.

[14] D. B. Terry, A. J. Demers, K. Petersen, M. J. Spreitzer, M. M. Theimer, and B. B. Welch, "Session guarantees for weakly consistent replicated data," in *Parallel and Distributed Information Systems, 1994., Proceedings of the Third International Conference on*. IEEE, 1994, pp. 140–149.

[15] J. Brzezinski, C. Sobaniec, and D. Wawrzyniak, "From session causality to causal consistency." in *PDP*, 2004, pp. 152–158.

[16] ——, "Session guarantees to achieve pram consistency of replicated shared objects," in *International Conference on Parallel Processing and Applied Mathematics*. Springer, 2003, pp. 1–8.

[17] V. Ramasubramanian, T. L. Rodeheffer, D. B. Terry, M. Walraed-Sullivan, T. Wobber, C. C. Marshall, and A. Vahdat, "Cimbiosys: A platform for content-based partial replication," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, 2009, pp. 261–276.

[18] A. Muthitacharoen, R. Morris, T. M. Gil, and B. Chen, "Ivy: A read/write peer-to-peer file system," *ACM SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 31–44, 2002.

[19] R. G. Guy, J. S. Heidemann, and T. W. Page Jr, "The ficus replicated file system," *ACM SIGOPS Operating Systems Review*, vol. 26, no. 2, p. 26, 1992.

[20] D. Perkins, N. Agrawal, A. Aranya, C. Yu, Y. Go, H. V. Madhyastha, and C. Ungureanu, "Simba: Tunable end-to-end data consistency for mobile apps," in *Proceedings of the Tenth European Conference on Computer Systems*. ACM, 2015, p. 7. [Online]. Available: https://github.com/SimbaService/Simba

[21] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in *ACM SIGOPS Operating Systems Review*, vol. 35. ACM, 2001, pp. 174–187.

[22] N. Tolia, J. Harkes, M. Kozuch, and M. Satyanarayanan, "Integrating portable and distributed storage." in *FAST*, vol. 4, 2004, pp. 227–238.

[23] B. D. Noble, M. Satyanarayanan, D. Narayanan, J. E. Tilton, J. Flinn, and K. R. Walker, "Agile application-aware adaptation for mobility," in *ACM SIGOPS Operating Systems Review*, vol. 31. ACM, 1997, pp. 276–287.

[24] N. Tolia, M. Satyanarayanan, and A. Wolbach, "Improving mobile database access over wide-area networks without degrading consistency," in *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 2007, pp. 71–84.

[25] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen, "Don't settle for eventual: scalable causal consistency for wide-area storage with cops," in *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*. ACM, 2011, pp. 401–416.

[26] ——, "Stronger semantics for low-latency geo-replicated storage." in *NSDI*, vol. 13, 2013, pp. 313–328.

[27] C. Li, D. Porto, A. Clement, J. Gehrke, N. M. Preguiça, and R. Rodrigues, "Making geo-replicated systems fast as possible, consistent when necessary." in *OSDI*, vol. 12, 2012, pp. 265–278.

[28] Y. Zhang, R. Power, S. Zhou, Y. Sovran, M. K. Aguilera, and J. Li, "Transaction chains: achieving serializability with low latency in geo-distributed storage systems," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013, pp. 276–291.

[29] D. B. Terry, V. Prabhakaran, R. Kotla, M. Balakrishnan, M. K. Aguilera, and H. Abu-Libdeh, "Consistency-based service level agreements for cloud storage," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013, pp. 309–324.

[30] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "Spanstore: Cost-effective geo-replicated storage spanning multiple cloud services," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013, pp. 292–308.

[31] G. T. Wuu and A. J. Bernstein, "Efficient solutions to the replicated log and dictionary problems," in *Proceedings of the third annual ACM symposium on Principles of distributed computing*. ACM, 1984, pp. 233–242.

[32] H. Yu and A. Vahdat, "Design and evaluation of a conit-based continuous consistency model for replicated services," *ACM Transactions on Computer Systems (TOCS)*, vol. 20, no. 3, pp. 239–282, 2002.

[33] R. Ladin, B. Liskov, L. Shrira, and S. Ghemawat, "Providing high availability using lazy replication," *ACM Transactions on Computer Systems (TOCS)*, vol. 10, no. 4, pp. 360–391, 1992.

[34] M. Nelson, B. Welch, and J. Ousterhout, *Caching in the Sprite network file system*. ACM, 1987, vol. 21, no. 5.

[35] D. Muntz and P. Honeyman, "Multi-level caching in distributed file systems," Center for Information Technology Integration, Tech. Rep., 1991.

[36] S. Chandra, M. Dahlin, B. Richards, R. Y. Wang, T. E. Anderson, and J. R. Larus, "Experience with a language for writing coherence protocols," in *Proceedings of the Conference on Domain-Specific Languages on Conference on Domain-Specific Languages (DSL), 1997*. Usenix Association, 1997, pp. 5–5.

[37] A. Rowstron and P. Druschel, "Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility," in *ACM SIGOPS Operating Systems Review*, vol. 35. ACM, 2001, pp. 188–201.

[38] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica, "Wide-area cooperative storage with cfs," in *ACM SIGOPS Operating Systems Review*, vol. 35. ACM, 2001, pp. 202–215.

[39] R. Tewari, M. Dahlin, H. M. Vin, and J. S. Kay, "Design considerations for distributed caching on the internet," in *Distributed Computing Systems, 1999. Proceedings. 19th IEEE International Conference on*. IEEE, 1999, pp. 273–284.

[40] D. Malkhi and D. Terry, "Concise version vectors in winfs," in *International Symposium on Distributed Computing*. Springer, 2005, pp. 339–353.

[41] R. G. Guy, J. S. Heidemann, W.-K. Mak, T. W. Page Jr, G. J. Popek, D. Rothmeier *et al.*, "Implementation of the ficus replicated file system." in *USENIX Summer*, 1990, pp. 63–72.

[42] Y. Saito, C. Karamanolis, M. Karlsson, and M. Mahalingam, "Taming aggressive replication in the pangaea wide-area file system," *ACM SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 15–30, 2002.

[43] S. Hao, N. Agrawal, A. Aranya, and C. Ungureanu, "Building a delay-tolerant cloud for mobile data," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 1. IEEE, 2013, pp. 293–300.

[44] B.-G. Chun, C. Curino, R. Sears, A. Shraer, S. Madden, and R. Ramakrishnan, "Mobius: unified messaging and data serving for mobile apps," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 2012, pp. 141–154.

[45] A. Oprea and M. K. Reiter, "On consistency of encrypted files," in *International Symposium on Distributed Computing*. Springer, 2006, pp. 254–268.

[46] G. Drive, "Google drive," 2016, https://developers.google.com/drive/.

[47] A. Inc, "icloud for developers," 2016.

[48] Dropbox, "Build your app on the dropbox platform," 2016, https://www.dropbox.com/developers.

[49] I. Drago, M. Mellia, M. M Munafo, A. Sperotto, R. Sadre, and A. Pras, "Inside dropbox: understanding personal cloud storage services," in *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM, 2012, pp. 481–494.

[50] B. Inc, "Box sync app," 2016, "http://box.com".

[51] P. Garcia-Lopez, M. Sanchez-Artigas, C. Cotes, G. Guerrero, A. Moreno, and S. Toda, "Stacksync: architecturing the personal cloud to be in sync."

[52] Y. Cui, Z. Lai, X. Wang, and N. Dai, "Quicksync: Improving synchronization efficiency for mobile cloud storage services," *IEEE Transactions on Mobile Computing*, vol. 16, no. 12, pp. 3513–3526, 2017.

[53] I. Zhang, A. Szekeres, D. Van Aken, I. Ackerman, S. D. Gribble, A. Krishnamurthy, and H. M. Levy, "Customizable and extensible deployment for mobile/cloud applications," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 97–112.

[54] R. Spahn, J. Bell, M. Lee, S. Bhamidipati, R. Geambasu, and G. Kaiser, "Pebbles: Fine-grained data management abstractions for modern operating systems," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 113–129.

[55] P. Shetty, R. P. Spillane, R. Malpani, B. Andrews, J. Seyster, and

E. Zadok, “Building workload-independent storage with vt-trees.” in *FAST*, 2013, pp. 17–30.

[56] K. Ren and G. Gibson, “Tablefs: Embedding a nosql database inside the local file system,” in *APMRC, 2012 Digest*. IEEE, 2012, pp. 1–6.

[57] “Apache couchdb,” 2018, http://couchdb.apache.org.

[58] “Touchdb,” 2018, http://tinyurl.com/touchdb.

[59] N. Preguiça, M. Zawirski, A. Bieniusa, S. Duarte, V. Balegas, C. Baquero, and M. Shapiro, “Swiftcloud: Fault-tolerant geo-replication integrated all the way to the client machine,” in *2014 IEEE 33rd International Symposium on Reliable Distributed Systems Workshops (SRDSW)*. IEEE, 2014, pp. 30–33.

[60] S. Burckhardt, M. Fähndrich, D. Leijen, and B. P. Wood, “Cloud types for eventual consistency,” in *European Conference on Object-Oriented Programming*. Springer, 2012, pp. 283–307.

[61] A. Shraer, A. Aybes, B. Davis, C. Chrysafis, D. Browning, E. Krugler, E. Stone, H. Chandler, J. Farkas, J. Quinn *et al.*, “Cloudkit: structured storage for mobile applications,” *Proceedings of the VLDB Endowment*, vol. 11, no. 5, pp. 540–552, 2018.

[62] P. Platform, “Parse platform,” 2016, https://parseplatform.github.io/.

[63] A. Gheith, R. Rajamony, P. Bohrer, K. Agarwal, M. Kistler, B. W. Eagle, C. Hambridge, J. Carter, and T. Kaplinger, “Ibm bluemix mobile cloud services,” *IBM Journal of Research and Development*, vol. 60, no. 2-3, pp. 7–1, 2016.

[64] A. Popov, A. Proletarsky, S. Belov, and A. Sorokin, “Fast prototyping of the internet of things solutions with ibm bluemix,” in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[65] Firebase, “Firebase,” 2017, https://firebase.google.com/.

[66] Kinvey, “Kinvey baas,” 2016, https://www.kinvey.com/.

[67] E. B. Nightingale and J. Flinn, “Energy-efficiency and storage flexibility in the blue file system.” in *OSDI*, vol. 4, 2004, pp. 363–378.

[68] A. D. Joseph, A. F. de Lespinasse, J. A. Tauber, D. K. Gifford, and M. F. Kaashoek, “Rover: A toolkit for mobile information access,” in *ACM SIGOPS Operating Systems Review*, vol. 29. ACM, 1995, pp. 156–171.

[69] O. Swift, “Openstack swift object storage service,” 2018, http.//swift.openstack.org.

[70] D. Bermbach, J. Kuhlenkamp, B. Derre, M. Klems, and S. Tai, “A middleware guaranteeing client-centric consistency on top of eventually consistent datastores.” in *IC2E*, 2013, pp. 114–123.

[71] Y. P. Faniband, I. Ishak, F. Sidi, and M. A. Jabar, “Enhancing mobile backend as a service framework to support synchronization of large object,” in *Proceedings of the 2017 International Conference on Information Technology*. ACM, 2017, pp. 383–387.

[72] Y. Xue, “The research on data synchronization of distributed real-time mobile network,” in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 3. IEEE, 2008, pp. 1104–1107.

[73] B. Salmon, S. W. Schlosser, L. F. Cranor, and G. R. Ganger, “Perspective: Semantic data management for the home.” in *FAST*, vol. 9, 2009, pp. 167–182.

[74] M. Shapiro, N. Preguiça, C. Baquero, and M. Zawirski, “A comprehensive study of convergent and commutative replicated data types,” Ph.D. dissertation, Inria–Centre Paris-Rocquencourt; INRIA, 2011.

[75] ——, “Conflict-free replicated data types,” in *Symposium on Self-Stabilizing Systems*. Springer, 2011, pp. 386–400.

[76] S. Burckhardt, A. Gotsman, H. Yang, and M. Zawirski, “Replicated data types: specification, verification, optimality,” in *ACM SIGPLAN Notices*, vol. 49. ACM, 2014, pp. 271–284.

[77] R. Klophaus, “Riak core: Building distributed applications without shared state,” in *ACM SIGPLAN Commercial Users of Functional Programming*. ACM, 2010, p. 14.

[78] V. Balegas, S. Duarte, C. Ferreira, R. Rodrigues, N. Preguiça, M. Najafzadeh, and M. Shapiro, “Putting consistency back into eventual consistency,” in *Proceedings of the Tenth European Conference on Computer Systems*. ACM, 2015, p. 6.

[79] P. Alvaro, N. Conway, J. M. Hellerstein, and W. R. Marczak, “Consistency analysis in bloom: a calm and collected approach.” in *CIDR*. Citeseer, 2011, pp. 249–260.

[80] N. Conway, W. R. Marczak, P. Alvaro, J. M. Hellerstein, and D. Maier, “Logic and lattices for distributed programming,” in *Proceedings of the Third ACM Symposium on Cloud Computing*. ACM, 2012, p. 1.

[81] S. Burckhardt, D. Leijen, M. Fähndrich, and M. Sagiv, “Eventually consistent transactions,” in *European Symposium on Programming*. Springer, 2012, pp. 67–86.

[82] S. Burckhardt, “Bringing touchdevelop to the cloud,” 2013, https://www.microsoft.com/en-us/research/blog/bringing-touchdevelop-to-the-cloud/.

[83] W. Brunette, S. Sudar, M. Sundt, C. Larson, J. Beorse, and R. Anderson, “Open data kit 2.0: A services-based application framework for disconnected data management,” in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 440–452.

[84] Y. Bai and Y. Zhang, “Stoarranger: Enabling efficient usage of cloud storage services on mobile devices,” in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 1476–1487.

[85] S. Chandrashekhara, T. Ki, K. Jeon, K. Dantu, and S. Y. Ko, “Bluemountain: An architecture for customized data management on mobile systems,” in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 2017, pp. 396–408.

[86] H. Tang, F. Liu, G. Shen, Y. Jin, and C. Guo, “Unidrive: Synergize multiple consumer cloud storage services,” in *Proceedings of the 16th Annual Middleware Conference*. ACM, 2015, pp. 137–148.

[87] Y. Zhang, C. Tan, and L. Qun, “Cachekeeper: a system-wide web caching service for smartphones,” in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 265–274.

[88] L. Lamport, “How to make a multiprocessor computer that correctly executes multiprocess programs,” *IEEE transactions on computers*, vol. 100, no. 9, pp. 690–691, 1979.

[89] C. Curino, E. Jones, Y. Zhang, and S. Madden, “Schism: a workload-driven approach to database replication and partitioning,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 48–57, 2010.

[90] A. L. Tatarowicz, C. Curino, E. P. Jones, and S. Madden, “Lookup tables: Fine-grained partitioning for distributed databases,” in *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 102–113.

[91] J. Jing, A. S. Helal, and A. Elmagarmid, “Client-server computing in mobile environments,” *ACM computing surveys (CSUR)*, vol. 31, no. 2, pp. 117–157, 1999.

[92] D. Barbará, “Mobile computing and databases-a survey,” *IEEE transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 108–117, 1999.

[93] P. Serrano-Alvarado, C. Roncancio, and M. Adiba, “A survey of mobile transactions,” *Distributed and Parallel databases*, vol. 16, no. 2, pp. 193–230, 2004.

[94] A. A. Imam, S. Basri, and R. Ahmad, “Data synchronization between mobile devices and server-side databases: a systematic literature review,” *Journal of Theoretical and Applied Information Technology*, vol. 81, no. 2, p. 364, 2015.

[95] T. Soyata, H. Ba, W. Heinzelman, M. Kwon, and J. Shi, “Accelerating mobile-cloud computing: A survey,” in *Cloud Technology: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2015, pp. 1933–1955.

[96] B. A. Milani and N. J. Navimipour, “A comprehensive review of the data replication techniques in the cloud environments: Major trends and future directions,” *Journal of Network and Computer Applications*, vol. 64, pp. 229–238, 2016.

[97] L. B. Bhajantri and V. V. Ayyannavar, “Cognitive agent based data synchronization in ubiquitous networks: A survey,” *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)*, vol. 10, no. 2, pp. 1–17, 2018.

Appendix

TABLE IV: Summary of reference designs

Framework	CMP	HU	Sync Protocol	Conflict Resolution
Coda [3] ∗	E ,PR	O	Use callbacks based on RPC	System level conflict resolution
Bayou [4]	E , TI	T, O	Two way Log exchange protocol	Application level resolution
Rover [68]	- , AC	O	↕ Two way using relocatable dynamic objects (RDOs) and queued remote procedure call (QRPC)	Object consistency is provided by application-level locking or by using application-specific algorithms
Perspective [73] ↔	E , PR, TI	O	↕ Modified update log to limit the exchanges to only the needed information,	Use two phase conflict resolver(Pre-resolver and a full resolver).
Cimbiosys [17] ↔	EF, PR,TI	O	↕ Use Eventual filter consistency and technique of Eventual knowledge singularity ,a compact synchronization-specific state for making economical use of bandwidth and system resource	Any device whose filter selects both conflicting versions may detect the conflict and either resolve it automatically or store both versions pending manual resolution
PRACTI replication [2] ↔	S, C, PR, AC , TI	O	Two types of communication 1.causally ordered Streams of Invalidations and 2. unordered Body messages. The protocol for sending streams of invalidations use log exchange protocol	Interface for detecting and resolving write- write conflicts according to application-specific semantics
SwiftCloud [59]	E , C, RB , PR	CRDTs	CRDTs do not require consensus on replica reconciliation	CRDTs can be updated without synchronization
Indigo [78]	EC , C, S, RB , PR , AC	CRDTs	CRDTs do not require consensus on replica reconciliation	CRDTs can be updated without synchronization
Izzy [43]	E , PR	T	Provide preferences to applications for setting on when and how to transfer data	Supports per-row versions for synchronization and conflict resolution
Simba [20] ∗	S, C , E , PR , AC	T + O	Versioning scheme which uses compact version numbers instead of full version vectors.3 consistency models also rely on this versioning scheme.Objects are stored and synced as a collection of fixed-size chunks for efficient network transfer.	Conflicts are stored in a separate conflict table until explicitly resolved by the user.
Sapphire [53]	S, E , PR	O	Cross Sapphire Objects (SOs) management and cross-SO transactions are NOT supported.Doc not clear about conflict resolution and handling	One of the deployment manager (DM) for the provide support for Dynamic allocation of load-balanced M-S replicas w/ eventual consistency. The extension OptimisticTransactions provide transactions with optimistic concurrency control, abort on conflict.
Mobius [44]	FSC , PR , AC	T	use a publish subscribe mechanism (Yahoo Message Broker (YMB))for many features such as asynchronous replication.	Supply the client with three kind of information to resolve conflict and can discard its local changes or issue a new update.
Key-Vaue Store with BloomL [80]	E	Built-in lattices	State of two replicas synchronized by exchanging their 'kv-store' maps.	The 'lmap' merge function automatically resolves conflicting updates made to the same key.
TouchDevelop [82] ∗ [60]	E	O (Cloud Types)	Special cloud types data is automatically shared between all devices, and is automatically persisted both on local storage and in cloud storage.	Automatic conflict resolution and no special code needed to handle merging
Middleware for client-centric consistency [70]	MRC , RYWC , E , PR	O	Protocol reads data from the storage system and adds a copy of that datum to its local cache, if the cache does not already contain that datum in that exact version (identified by its vector clock).	The application must take care of the data violation-handling task by reloading data
Open Data Kit 2.0 [83] *	E	T	Single database row is the base unit and use a small granularity of change-tracking to enable smaller data transmission.	User must resolve the conflict by either taking the server's change, the local change, or mix of the local change and the server change
StoArranger [84]	E , PR	T	Delay and batch cloud backup requests from apps to minimize the impact of transmission promotion/tail energy	Conflict detection is based on the folder meta data with folder sync APIs.
Unidrive [86]	S, E, PR	O	Metadata of content is stored in cloud and synced to all clients using a quorum based distributed locking protocol and chunking technique	It support multiple Consumer cloud storage in which synchronization logic is purely implemented at client devices and all communication is conveyed through file upload and download operations
QuickSync [52] ∗	E	O	Sync efficiency is improved by using three key components, the Network-aware Chunker, the Redundancy Eliminator , and the Batched Syncer	Do not address consistency or conflict handling but improves Sync efficiency.
Parse Server [62] ∗	S, E, PR	T	The Parse Server SDKs provides a local datastore which can be used to store and retrieve the PFObjects by 'pinning' process	Server side conflicts are handled using generic 'beforeSave' function in the cloud. Applications are responsible to handle conflicts.
StackSync [51]	S, E, PR	O	Sync protocol is based on RPCs or method calls by ObjectMQ middleware	A copy of the conflicted document is created and user needs to decide about this.

Continued on next page

TABLE IV – continued from previous page

Framework	CMP	HU	Sync Protocol	Conflict Resolution
Dropbox [48]	RAW	O	The basic object in the system is a chunk of data with size of up to 4MB. Files larger than that are split into several chunks. Reduces the amount of exchanged data by using delta encoding when transmitting .	File meta data has a unique revision identifier used to detect changes and avoid conflicts
iCloud with CloudKit [47]	S, E	O	CloudKit APIs support to fetch for only the changes since the last time it updated.	iCloud server returns the error code with objects that contains the different versions of the conflicting record and user can perform whatever resolution logic is needed to resolve the conflict
Amazon Dynamo [5]	E, PR	O	Implements an anti-entropy (replica synchronization) protocol and Merkle trees for faster and to minimal the amount of data transfer.	It makes use of object versioning and application-assisted conflict resolution
Bluemix Mobile Cloud Service [63]	E, PR	O	Applications use Cloudant Sync (an Apache CouchDB^{TM} replication-protocol-compatible) to store, index and query local JSON data on a device	It is the application’s responsibility to detect the conflicts and resolve them based on conflict details
Firebase [65]	E, PR	O	Proprietary sync protocol and lack of documentation	Timestamp based conflict resolution
Kinvey [66]	S, E, PR, AC	O	Delta Sync allows only sync of new and updated entities only.	The libraries and backend implement a default mechanism of ”client wins”, which implies that the data in the backend reflects the last client that performed a write. Custom conflict management policies can be implemented with Business Logic.

Note: See Table-I for the list of symbols used.
*: open-source

Performance Evaluation WPAN of RN-42 Bluetooth based (802.15.1) for Sending the Multi-Sensor LM35 Data Temperature and RaspBerry pi 3 Model B for the Database and Internet Gateway

Puput Dani Prasetyo Adi
Micro - Electronics Research Laboratory
Kanazawa University
Kanazawa, Ishikawa, Japan

Akio Kitagawa
Micro - Electronics Research Laboratory
Kanazawa University
Kanazawa, Ishikawa, Japan

Abstract—**This research will be a test of a multi-sensor data transmission using the Wireless Sensor Network based on Bluetooth RN-42. Accordingly this research, LM35 is a type of Temperature Sensor, furthermore, this research will be used two LM35 sensors installed on the Arduino board and to be processed by Arduino Integrated Development of Environment (IDE) with C++ language. Arduino will be sending of all sensor data from LM35 temperature sensor by Slave RN-42 Bluetooth Configuration to master RN-42 Bluetooth configuration. Furthermore, the temperature data will be sending on Raspberry Pi 3 as an Internet Gateway then data will be sent to the internet and sensor data will be stored in the MySQL database. Furthermore, Sensor data can be accessed by other computers on the internet network using PuTTY with the Raspberry Pi 3 IP Address 192.168.1.145. Moreover, testing is also done by measuring the Signal power of Wireless Personal Area Network with the Receive Signal Strength Indicator variable, so the Bluetooth signal strength in sending multi-sensor data can be known appropriately.**

Keywords—RSSI; Bluetooth; Raspberry pi 3; Internet Gateway

I. INTRODUCTION

Wireless Sensor Network technology continues to grow rapidly including Bluetooth, one of the advantages of sensor data delivery systems based on Wireless Sensor Network is the Low Power Consumption, currently the technology of Wireless Personal Area Network (WPAN) developed in the world of research by the telecommunications world is Bluetooth Low Energy (BLE), the Bluetooth Low Energy (BLE) specification is (10.1 uA, 3.3 V supply at 120 s interval), therefore the energy needed is the smallest compared to ZigBee (15.7 uA) and ANT (28.2 uA) [1].

In this research using Bluetooth RN-42, one of the advantages of Bluetooth RN-42 is Low power (26 uA sleep, 3 mA connected, 30 mA transmit) but when compared to the energy needed by Bluetooth Low Energy (BLE) adrift to 20 uA, this includes a fairly large value [13], another advantage of Bluetooth RN-42 is compatible with the Arduino Microcontroller, this is because many types of Bluetooth are not compatible with certain types of microcontrollers that are easy to complete coding programs inside Integrated Development of Environment (IDE) [13].

Wireless Personal Area Network (WPAN) 802.15.1 or Bluetooth is very suitable to be used in short distances areas, unlike ZigBee (802.15.4) which has mesh capability so that more sensor nodes as router nodes will minimize the distance of one node to another node so can minimize the use of battery or power [4]. The standard Bluetooth protocol in sending and receiving data is 2.4 GHz. Bluetooth is used as a data sending device for short-range, another advantage of Bluetooth is low-power and low-cost sensors [10].

Bluetooth can send short-range sensor data approximately less than 100 meters. bluetooth is a Radio Frequency (RF) transmission device using serial communication. Bluetooth devices have an address usually presented as the hexadecimal value [10]. Parameters for the quality of sending Radio Frequency on Bluetooth are using the Received Signal Strength Indicator (RSSI). In this research, the measurement of the Received Signal Strength Indicator (RSSI) values on Bluetooth RN-42 will be measured with different distances so that the distance ratio and RSSI value (dBm) can be known [2].

The sensor used for this research is an LM35 temperature sensor and one of the characteristics of this sensor is that it is sensitive to heat, therefore, This sensor can be implemented in various fields. e.g. monitor the temperature in concrete, this is very important because the amount of heat need to manage properly, one method to monitor on the temperature of the concrete during the hardening process, the sensor used is LM35 [6].

The position of Bluetooth compared to other data sending devices is still popular, with short distance prosperity in addition to audio and stereo communication, Bluetooth is also used to support the Internet of Things (IoT) and Machine to Machine (M2M) application using Bluetooth Low Energy (BLE), therefore it is expected that Bluetooth will remain the device for sending data packages to date [5].

II. Related Studies

Artem Dementyev [1] in this research, discussed Power Consumption analysis of Bluetooth Low Energy, ZigBee and ANT sensor nodes in a cyclic sleep scenario. therefore, it was concluded that Bluetooth Low Energy (BLE) is a data sending device that has power consumption lowest, compared to ZigBee and ANT. accordingly the theory and benefit of Wireless Sensor Network, This is very important in considering long life factors on sensor nodes seen from energy use factors.

Guoquan Li [2] in this research examined the Received Signal Strength Indicator (RSSI) by placing the sensor node in the indoor position using the Positioning algorithm approach. therefore, The RF radio used is Bluetooth by using mobile technology development. furthermore, with the Positioning algorithm, the results are better than real-time RSSI values.

Janire Larranaga, Leire Muguira, Juan-Manuel Lopez-Garde and Juan-Ignacio Vazquez [3], take measurements using positioning algorithms, while the parameters used are The RSSI (Received Signal Strength Indicator) I refer to this research, while the device used is ZigBee, accordingly by using the positioning algorithm, we estimate the node position with good resolution (3 m average error).

Manuel Ramos [6] in this research, use an LM35 is one of the sensors used to examine the quality of Concrete during heated conditions.accordingly from the system block diagram, the temperature sensor used is 2-32 LM35 sensors that are connected to the Data Acquisition Module and Personal Computer. Based on the results of an analysis, the more sensors used can significantly provide accurate tracking of the internal temperature of the concrete so that the quality of concrete can be known.

M. Niswar [7] in his study of sending Wireless Sensor Network data using ZigBee, the data sent was Pulse Sensor, his research examined Quality of Service (QoS) when data transmission took place, experiments were carried out at different distances, Packet Loss was obtained when sending 4 sensor nodes (ED) simultaneously to the Coordinator Node, so that only 3 sensor nodes can be accommodated by the Coordinator node without packet loss.

Besides [7], the application of RF Bluetooth signals in the health field was carried out in the research of Ying Zhang, Hannan Xiao, [12], moreover, the use of RF Bluetooth signals in this research is combined with intelligent physiological sensors that involve technology integration RF Bluetooth, hardware and software organization and solution for onboard signal processing.

P. Ferrari [8] in this research, the application of Graphical User Interface (GUI) using Web interfaces is one of the implementations in supporting Internet Of Things applications, P. Ferrari uses Bluetooth-based RF Radio in sending sensor data. Measurement about power dissipation, area coverage, and response time confirm the proposed network feasibility and effectiveness.

III. Methodology

A. The Received Signal Strength Indicator (RSSI)

The Received Signal Strength Indicator (RSSI) is a parameter to measure the quality of Radio Frequency (RF) in this case is we can be measured a Bluetooth Communication Prosperity. The Bluetooth type is the RN-42 Bluetooth module with the Master-Slave Bluetooth configuration. furthermore, RSSI can be determined from A, n and d value. the magnitude of the RSSI value is expressed in decibel milliwatts (dBm), the Received Signal Strength Indicator (RSSI) is used as a determinant of a signal strength parameter. Accordingly, with [3], important parameters used to support the success of the indoor positioning algorithm are RSSI. the formula used in the RSSI calculation is in accordance with equation (3).

$$[P_r(d)] = [P_r(d_0)]_{dBm} - 10\,log\left(\frac{d}{d_0}\right) dBm + X_{dBm} \qquad (1)$$

$$[P_r(d)] = [P_r(d_0)]_{dBm} - 10\,n\,log\left(\frac{d}{d_0}\right) \qquad (2)$$

$$\text{RSSI}(dBm)] = [P_r(d_0)]_{dBm} = A - 10n\log d \qquad (3)$$

$$d = 10^{\left(\frac{A-RSSI}{10^n}\right)} \qquad (4)$$

Parameters description :

- RSSI = Received Signal Strength indicator (dBm)
- d = distances (meter)
- n = path loss exponent (e.g : free space = 2)
- A = Received signal Strength at 1 meter (dBm)
- Pr = Receiver Power (dBm)
- Pt = Transmit Power (dBm)

TABLE I. PATHLOSS EXPONENT VALUE FROM DIFFERENT ENVIRONMENT

No	Environment	PathLoss Exponent, n
1	Free Space	2
2	Urban area cellular radio	2.75 to 3.5
3	Shadowed urban cellular radio	3 to 5
4	In Building line of Sight	1.6 to 1.8
5	Obstructed in building	4 to 6
6	Obstructed in factories	2 to 3

At [7], the distance (m) found without loss packet is 30 meters. Equation (4) is a formula to get a distance value from RSSI, A, and n so that from this formula can be determined value of d (meter). Table 1 Shows the PathLoss Exponent used to calculate the RSSI value and determine the distance, n value varies based on the environment.

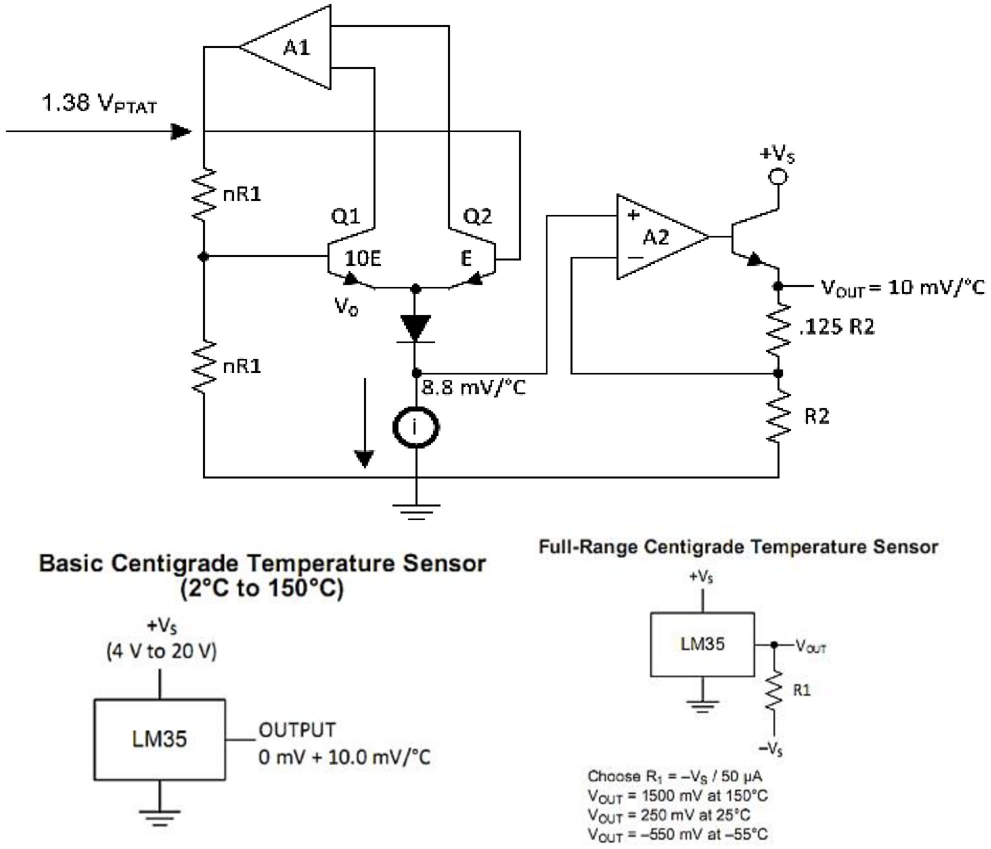


Fig. 1. Schematic of LM35 Temperature Sensor.

B. LM35 Temperature Sensor

The LM35 temperature sensor is used to convert temperature to electrical quantities in the form of voltage. nevertheless, this sensor has high accuracy and easy to design. accordingly, the input pins of LM35 temperature sensors different with DHT11 Temperature and humidity sensors, the difference is in the position of the input data. the theory of LM35 use analog input data from Arduino board and DHT11 use Digital input data from Arduino board, following the schematic of the LM35 temperature sensor like a show at figure 1 [14].

Figure 1 shows the schematic of an LM35 Temperature sensor, in a schematic, describes 3 input/output pins GND, data / Vout and Vs, Vs 4 to 20 Volt Dc, Vout of 10 mV / C^{o}.[14]

LM35 Temperature Sensors work with change a temperature value into a voltage quantity. Accordingly The ideal voltage of LM35 temperature sensor has a temperature ratio of 100°C equivalent to 1 volt. and This sensor has self-heating <0.1 ° C. therefore, LM35 sensor converts the physical temperature to a voltage that has a coefficient of 10 mV / °C, which means a rise in temperature of 1 ° C will increase in temperature by 10 mV.

Figure 2 describes the schematic between Arduino, 2 LM35 Temperature sensors and Bluetooth modules, figure 2 is a Slave module.

C. Analog to Digital Converter (ADC) and Relationship between Temperature and Voltage

Figure 2 show LM35 temperature sensor connected with A0 Arduino board, accordingly temperature data is derived from analog data, therefore LM35 Temperature sensor produces the analog signal data. Microcontroller ATmega 328p have an 8 and 10 Bit ADC and have the same function and number of pins. the type of microcontroller used in this research has a 10 Bit ADC, meaning that the digital data from the conversion of 10 Bit ADC is 2^10 the result is 1024.

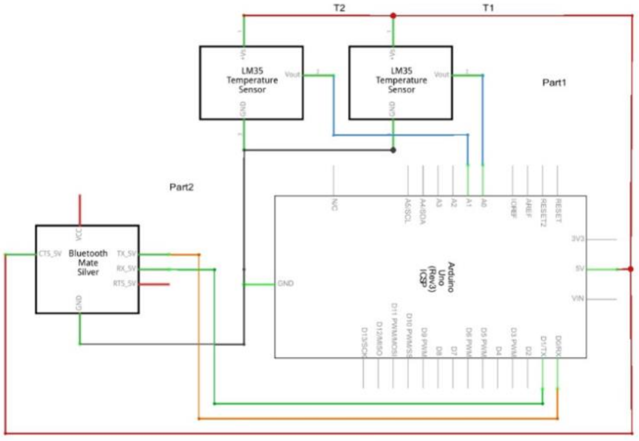


Fig. 2. Temperature Sensor LM35 at Arduino Board.

The conversion results of ADC = (Vin x 1024) / Vref (5V) (5)

for example, Vin is 1000mV, then the conversion result

is ADC = (1000mV x 1024) / 5000mV = 205

Vin= results of ADC x (5/1023) = 1 Volt or 1000 mV

Temperature °C = Vin / 0.01 Volt (10mV / Celsius, LM35) (6)

0.01 volt from the characteristics of LM35 (10mV / Celsius)

For example Vin =1Volt, then T = 1/0.01 =100°C

0 V DC = 0 °C, 10 mV DC = 1 °C, 100 mV DC = 10 °C, 1000 mV DC = 100 °C, 1500 mV DC = 150 °C

D. dBm and mWatt

Power below 1 mW is expressed as a negative dBm value, on the contrary above mW is a positive dBm value. accordingly, dBm to measure signal strength, the logarithmic scale is easier to understand, by measuring where 1 mW (milliwatt) of power is defined as 0 dBm. the normal signal strength ranges from -100 dBm to -50 dBm, even though there are smaller or larger ones than this range, accordingly the theory, this is the normal range. The following is the conversion from dBm to Watts, from Watts to dBm and Milliwatts to dBm [15].

dBm = 30 + Log 10 (Watts) (7)

Watts = 10^((dBm - 30)/10)

milliWatts = 10^(dBm/10)

E. RaspBerry Pi 3 Model B Board

Raspberry Pi 3 is a board that has many advantages compared to the previous RaspBerry version moreover RaspBerry Pi 3 is equipped with On-Board Bluetooth 4.1 Wi-Fi, 4 USB 2 Port, 10/100 LAN Port, 40 Pin Extended GPIO, Micro SD Card Slot, Full-Size HDMI Video Output, CSI Camera port, 3.5 mm 4 pole composite video and audio output jack, Micro USB Power Input. an upgraded switched power source that can handle up to 2.5 Amps, Broadcom BCM2837 64 bit Quad Core CPU at 1.2 GHz, 1 GB RAM, the difference from the previous lies in On Board Bluetooth 4.1 Wi-Fi [16].

Figure 3 is a Raspberry Pi 3 model B, In this research, accordingly the schematic of RaspBerry Pi 3 Model B Board will be connected to the Arduino UNO Board via USB 2.0 Port Cable Arduino, in this case, the prosperity of Raspberry Pi as the IoT devices will be tested, the ability of raspberry pi 3 in sending data to the database is very strong, the database which is MariaDB, remote IP can be done using PuTTY, this is done to work on the Raspberry Pi processor. on the Arduino board, Master Bluetooth node will be used as the recipient of temperature data from the LM35 sensor on the Slave Bluetooth node. the initial experiment is to use LEDs.

More application can be used LEDs to make the indicators, this is important things. The Slave Bluetooth node sends a command to turn on the Blinking LED on the Bluetooth master node. consequently, the data displayed on the master Bluetooth node Serial monitor. the data output at Master Bluetooth is LM35 Sensor node. accordingly, The function of RaspBerry pi 3 is to send real-time data from RaspBerry pi 3 to a MySQL database using the Python programming language, Raspberry Pi 3 Compatible with Python 3 programming language, e.g. a pymysql database library compatible with python 3.

Furthermore, in experimental data using LEDs, Commands in programming languages used using strings or char (characters) by sending L and H values. while in this experiment, the temperature (°C) data using float on the data type [5].

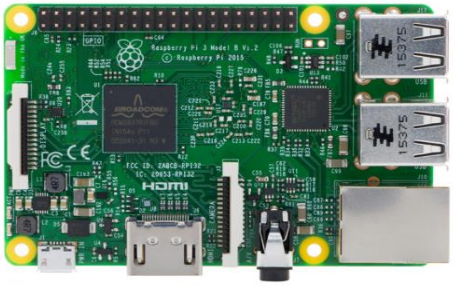


Fig. 3. RaspBerry Pi 3 Model B Board.

F. RN-42 Bluetooth Module

The RN-42 Bluetooth module has been created by a wireless serial communication interface between two devices, e.g. a microcontroller, PC, Smartphone and other modules.on the schematic of the RN-42 Bluetooth module will send the LM35 temperature sensor data. accordingly, the RN-42 Bluetooth breadboard-friendly module is compatible with all 5V and 3.3V microcontroller platforms, e.g. this research use Arduino Microcontroller [13].

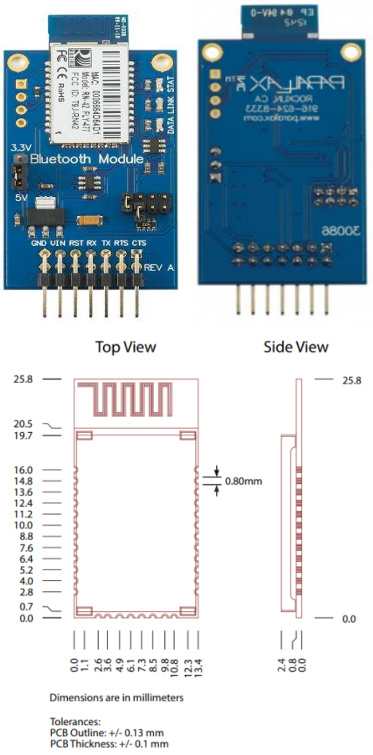


Fig. 4. Bluetooth RN-42 Module and Dimension.

Figure 4 is the module and dimension of the RN-42 Bluetooth module [13], the RN-42 module is compatible with the Arduino IDE programming using C ++. The RN-42 Bluetooth module has 4 pins that are used to communicate, VIN, TX, RX, and GND, the rest are not used. The command to communicate between slave and Bluetooth master is shown in the following program code:

```
//puput MERL Code_Master
String Remote ="0006664FBB11"; // slave
void setup()
{
Serial.begin(9600);
Serial.print("$$$");
delay(1000);
Serial.println("c,"+ Remote +"\r");
delay(1000);
Serial.println("SM,1"); // Master
delay(1000);
Serial.println("---,"+ Remote +"\r");
delay(1000);
}
------------------ Code Program 1 ------------------
```

The master Bluetooth node, number 0006664FBB11 is the MAC ID from the slave Bluetooth node, the MAC ID very important for identity. accordingly 00066660618D is the MAC ID by the master Bluetooth module sent by the Slave Bluetooth module, furthermore, the Slave Bluetooth command is shown in the following program code :

```
//puput MERL Code_Slave
String Remote ="00066660618D"; // master
void setup()
{
pinMode(led,OUTPUT);
Serial.begin(9600);
Serial.print("data dari slave");
delay(100);
Serial.println("c,"+ Remote +"\r");
delay(100);
Serial.println("SM,0"); // Slave
delay(100);
Serial.println("---,"+ Remote +"\r");
delay(100);
}
------------------ Code Program 2 ------------------
```

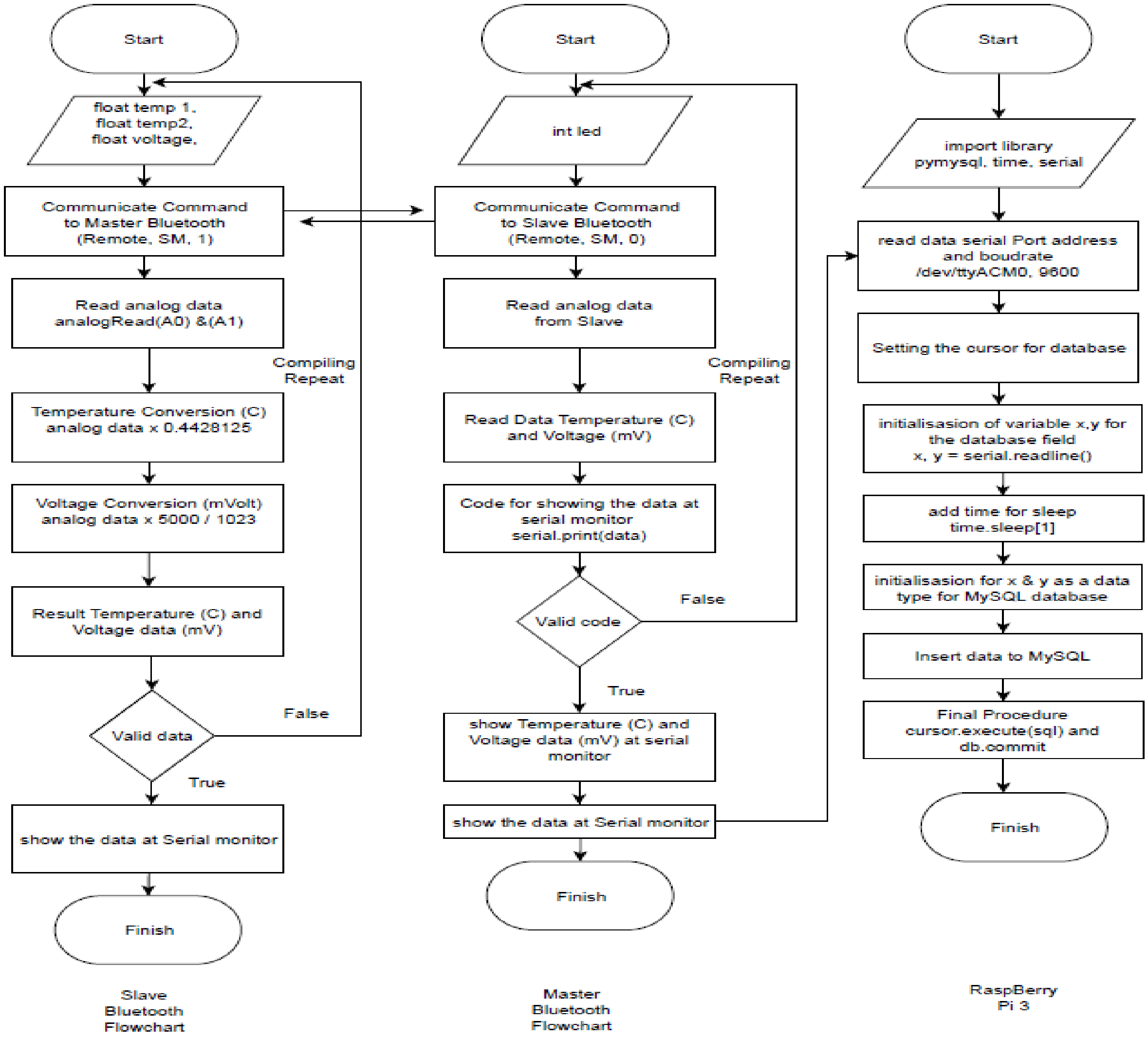


Fig. 5. The Flowchart System.

G. Flowchart System

The method in this research can be seen in the flowchart in figure 5. accordingly there are 3 important parts that are presented in this flowchart and all three are related. The three parts are Bluetooth Slave, Bluetooth master, and RaspBerry Pi 3. Furthermore, Bluetooth Slave is a node unit formed from an Arduino microcontroller, LM35 temperature sensor, and one unit RN-42 Bluetooth module. Same with Bluetooth master. The difference is the function of each node. In Bluetooth slave, the program created is how to send LM35 temperature sensor data to a Bluetooth master, then the master node sends LM35 sensor data to MySQL using compatible python code in the RaspBerry Pi 3.

The process of sending LM35 sensor data to this Bluetooth Master will be analyzed the ability of the reception signal, which needs to be considered is the python program in displaying data from the sensor, then RaspBerry Pi 3 read to the MySQL database. the arrow at the flowchart shows the connectivity between slave and master Bluetooth RN-42 and then show the data at Raspberry Pi.

Figure 5 describes the flowchart system in this research. In the flowchart section 1 explains how the Slave Bluetooth node communicates with part 2 of the flowchart or master Bluetooth node with all configurations used then after successfully and validated correctly, the data sent to part 3 of the flowchart, RaspBerry pi 3 which is then processed using the Python programming language and then will be sent to the MySQL database, there needs to be a library to connect Python and MySQL so that the LM35 sensor temperature data can be successfully sent and stored in the MySQL database and data can be viewed easily at PHPMyAdmin. to access PHPMyAdmin, the default IP Address is 127.0.0.1/ PHPMyAdmin.

Important to configure start and stop on MySQL to know a MySQL running or not. For example # sudo service mysql stop and Sudo mysql_safe –skip-grant-tables & then easy to enter the MySQL without the password.

Figure 6 shows The Hardware Connectivity and Communication Testing. accordingly, temperature data send from the master Bluetooth to Bluetooth slave, if testing the delivery of masters have succeeded with the code in Arduino use C++ language, then the LM35 temperature sensor data sent through Python programming using RaspBerry pi to store the data in the MySQL Database, therefore MySQL database library in Python is needed.

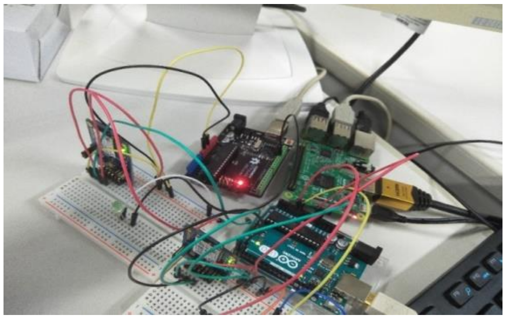


Fig. 6. Hardware Connectivity and Communication Testing.

Figure 6 shows the two RN-42 Bluetooth module, two Arduino microcontroller modules, and one raspberry pi 3 modules each of which is connected.

Figure 7 shows the Design system in this research as a whole. Where during the process described in the flowchart then the data is processed by RaspBerry Pi 3 using the Python programming language then the data is sent to the database and other computers can be read the data in realtime using PuTTY Configuration.

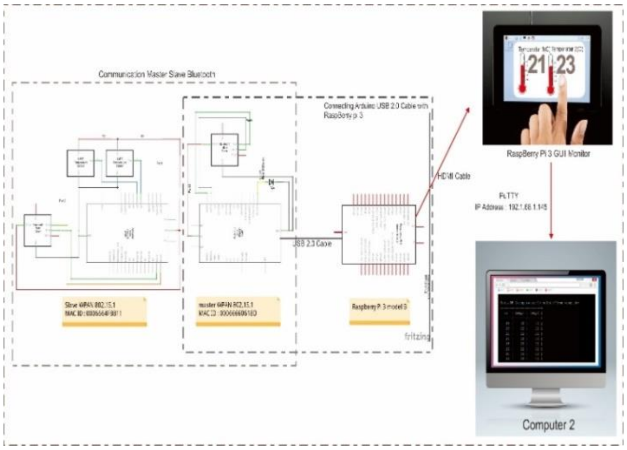

Communication Master Slave Bluetooth
Connecting Arduino USB 2.0 Cable with RaspBerry pi 3
USB 2.3 Cable
Raspberry Pi 3 model B
fritzing
HDMI Cable
21
23
RaspBerry Pi 3 GUI Monitor
PuTTY
Computer 2

Fig. 7. Design System at this Research.

IV. Result and Analysis

A. Voltage and Temperature

On voltage and temperature measurement of an LM35 temperature sensor using an Arduino microcontroller, the value of the input voltage (Vin) is determined from the reading of the Analog-Digital Converter (ADC).

Then the temperature value depends on the input voltage (Vin) value. an increase in the temperature value in ordinarily give impact on voltage value, this condition is shown in figure 8 & 9.

Figure 8 shows the temperature change at a certain time, The graph shows an increase in temperature value consequently give the impact in voltage value, as for the number of sensor used is 1 Temperature sensor.

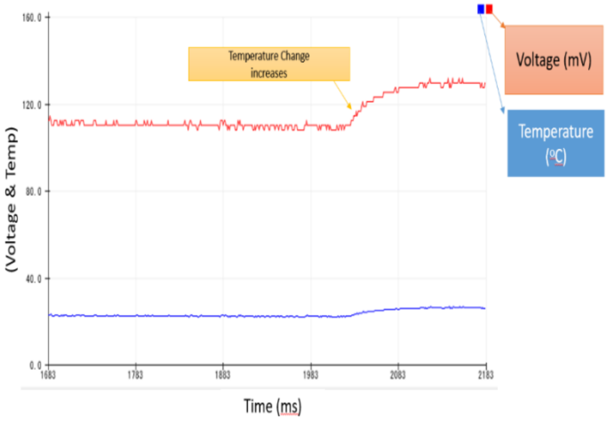


Fig. 8. Voltage and 1 Temperature Sensor Graph.

Figure 9 shows the temperature change at a certain time, The graph shows an increase in temperature value consequently give the impact in voltage value, as for the number of sensors used is 2 Temperature sensors.

```
    float temp1;
    float temp2;
    int led=13;
    void setup() {
    pinMode(led,OUTPUT);
    Serial.begin(9600);
    }
    void loop() {

      temp1=analogRead(A0);
      temp1=temp1*0.4428125;
      Serial.print("1= ");
      Serial.print(temp1);
      Serial.print(" ");
      delay(500);
      temp2=analogRead(A1);
      temp2=temp2*0.4428125;
      Serial.print("2= ");
      Serial.println(temp2);
      Serial.print(" ");
      delay(500);
}
--------------- Code Program 3 -----------------
```

After the two LM35 temperature sensors that are initialized in the code program 3, furthermore continue to proceed temperature to voltage conversion.

```
    //Puput Code ~ Voltage Temp Convert
    //temperature voltage (T): 0-500
    // voltage input(v in) : 0-1024
    //T=(vin*500)/1024;
    float data,suhu,vref;
    float voltage;
    String Remote ="0006664FBB11";
    const int pSuhu=A0;
    void setup()
    {
      pinMode(pSuhu,INPUT);
      Serial.begin(9600);
      Serial.print("$$$");
      delay(1000);
          Serial.println("c,"+ Remote +"\r");
                     delay(1000);
      Serial.println("SM,1"); // master
      delay(1000);
         Serial.println("---,"+ Remote +"\r");
                     delay(1000);
      }
    void loop()
    {
      data=analogRead(A0);
      data=data*0.4428125;
      voltage=data*5000/1023;
      Serial.print(" ");
      Serial.print(voltage);
      Serial.print(" mV");
      Serial.println();
      Serial.print(data);
      Serial.print(" C");
      delay(100);
    }
--------------- Code Program 4 -----------------
```

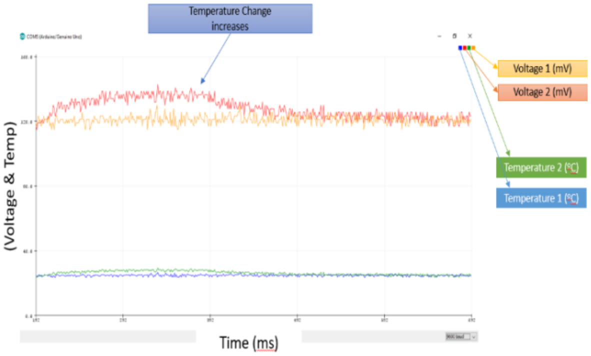


Fig. 9. Voltage and 2 Temperature Sensor Graph.

Furthermore, the Calculation of temperature and voltage values can be seen in equation (6). C ++ language is a programming language in the Arduino IDE to represent temperature sensors, furthermore can be seen in program code 3 and 4.

B. Receiver Signal Strength Indicator (RSSI)

In Waltenegus's Dargie and Christian Poellabauer book references [11], says the method of received signal strength (RSS) is a signal that decays or decreases the distance traveled. therefore, the features found in wireless devices are received signal strength (RSSI) which can measure the amplitude of the incoming radio signal.

The one of analyzing indicator transmitting data on Master and Slave Bluetooth RN-42 is the Received Signal Strength Indicator (RSSI). to get the RSSI value, it is necessary to do the calculation as in equation 3. The graph in figure 10 is the measurement result of the signal strength generated parameter (dBm).

The level of packet loss data is also influenced by RF around the data packet delivery, therefore it is necessary to do data transmission without interruption Another RF, on [9], the level (Packet Error Rate) PER and (bit error rate) studied at the time of sending ZigBee RF with interference is examined some Bluetooth networks.

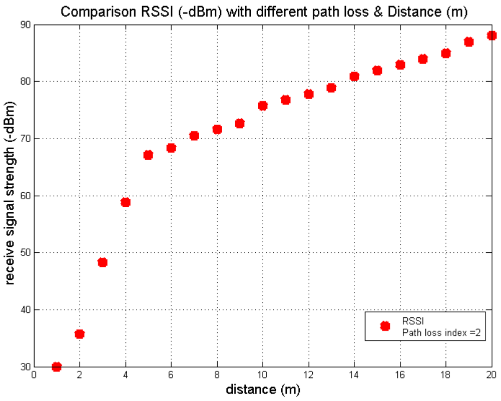


Fig. 10. RSSI WPAN 802.15.1 (Bluetooth RN-42) Graph with n = 2 (Free space).

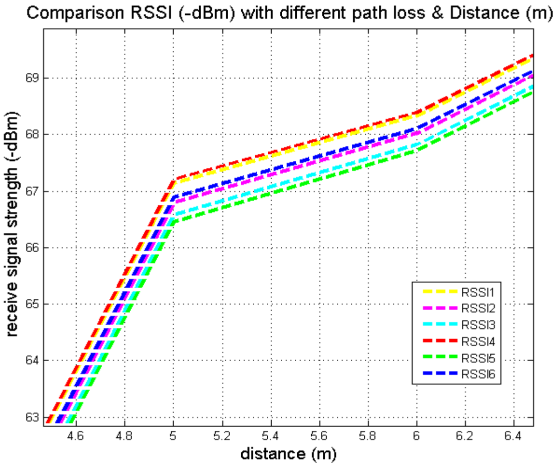


Fig. 11. Comparison RSSI WPAN 802.15.1 (Bluetooth RN-42) with different environment.

Table 1 explains about Pathloss Exponent Value from the different Environment. the value of n (Pathloss Exponent) will determine the value of RSSI. for example, the RF Bluetooth RN-42 signal strength at 1 m is -30 (dBm) with the path loss index n = 2 (free space) . therefore the RSSI (dBm) = A - 10.n log d, then RSSI (-dBm) = 30 - 10.2 log 1, then RSSI value is (dBm) -30 dBm.RSSI is obtained from the calculation of equation (3).

The graph in figure 10 is a comparison of Bluetooth signals (dBm) and distance (m). in figure 10, the distance in the experiment is 20 m, the further of the distance between Bluetooth End Device and the coordinator node consequently the value of RSSI shows a decreased strength of the signal (dBm).

The graph in figure 11 is a comparison of Bluetooth signals (dBm) and distance (m).in figure 11, the distance in the experiment is 20 m but displayed in more detail, the number of RSSI tested is 6 from the different Exponent or environment (n) Path Loss values.

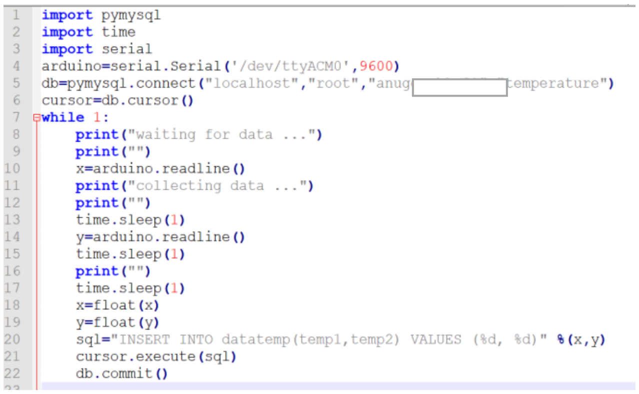
The further the distance between the Bluetooth End Device and coordinator node, as a result, decreased signal strength (dBm).

C. *Sending the Temperature Data to MySQL Database*

Raspberry pi 3 uses Python 3 to programming and sends sensor data to the MySQL database.

Python requires the pymysql library for its MySQL database, beside pymysql library, time and serial library is required. Furthermore use the MySQL command to connect to the localhost network, root folder, username and password.Then analyze the variables that will be displayed for example x = Arduino.readline (), y = Arduino.readline (), then display (x, y) with the data type int or float, for example, float x, y.

to enter the temperature sensor data into the database, we use the INSERT command into the table name, in this case, the data temp table. as shown in figure 12.



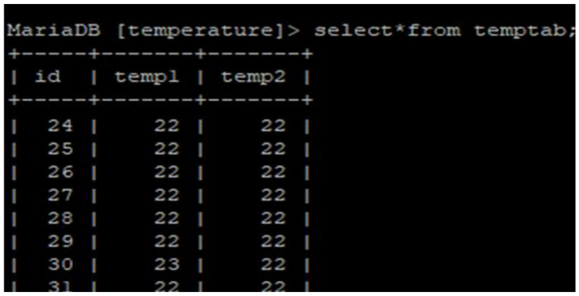
```
import pymysql
import time
import serial
arduino=serial.Serial('/dev/ttyACM0',9600)
db=pymysql.connect("localhost","root","anug[redacted]temperature")
cursor=db.cursor()
while 1:
    print("waiting for data ...")
    print("")
    x=arduino.readline()
    print("collecting data ...")
    print("")
    time.sleep(1)
    y=arduino.readline()
    time.sleep(1)
    print("")
    time.sleep(1)
    x=float(x)
    y=float(y)
    sql="INSERT INTO datatemp(temp1,temp2) VALUES (%d, %d)" %(x,y)
    cursor.execute(sql)
    db.commit()
```

Fig. 12. Phyton Code to sending the data temperature to MySQL database.

Finally, the puTTY command to connect the RaspBerry pi 3 device to another device with an IP address: 192.168.1.145, this the IP Address of the Raspberry Pi 3 device. furthermore, after doing a series of commands, finally, it can enter MySQL is possible.

Basic commands on MySQL database, for example, to viewing tables; show databases, select the databases; use database_name, show the tables; show tables, Look the tables Description or structure; desc tables_name and look the entry data in the tables; select * from tables_name.

Figure 13 shows the output from delivery data of LM35 temperature sensor in real time. accordingly the MySQL Table, There is data 2 temperature that can be monitored assuming that the two temperature sensors are placed in different places later.



```
MariaDB [temperature]> select*from temptab;
+-----+-------+-------+
| id  | temp1 | temp2 |
+-----+-------+-------+
|  24 |    22 |    22 |
|  25 |    22 |    22 |
|  26 |    22 |    22 |
|  27 |    22 |    22 |
|  28 |    22 |    22 |
|  29 |    22 |    22 |
|  30 |    23 |    22 |
|  31 |    22 |    22 |
```

Fig. 13. Output data Temperature sensor in Database.

ACKNOWLEDGMENT

Thanks to Professor Akio Kitagawa a helped correct this text and has prepared all the tools, nevertheless, the help is very useful for the author. furthermore, thanks to all researchers a helped in completing the research on Wireless Sensor Network using Bluetooth and implementation at Internet of Things (IoT) application.

V. CONCLUSIONS AND SUGGESTION

The theory of the RSSI value equation with a change in the value of n (Path Loss Index) has produced a different signal strength value, the greater the value of n, the greater the signal strength so that the signal strength weakens accordingly equation 3.

From this research, successfully storing data to the database using the Python Language command in figure 12. The temperature data table output can be seen in figure 13. localhost / phpmyadmin used to manage the databases, furthermore interface development is needed by creating a Graphical User Interface (GUI) by the PHP, HTML, Javascript and JASON programming language to produce the real-time graph so that it is more user-friendly.

REFERENCES

[1] Artem Dementyev, Steve Hodges, Stuart Taylor, Joshua Smith, "Power consumption analysis of Bluetooth Low Energy, ZigBee and ANT sensor nodes in a cyclic sleep scenario", IEEE International Wireless Symposium (IWS), 2013.

[2] Guoquan Li, Enxu Geng, Zhouyang Ye, Yongjun Xu, Jinzhao Lin and Yu Pang, "Indoor Positioning Algorithm Based on the Improved RSSI Distance Model", Sensors MDPI Journals, Published: 27 August 2018.

[3] Janire Larranaga, Leire Muguira, Juan-Manuel Lopez-Garde and Juan-Ignacio Vazquez “An Environment Adaptive ZigBee-based Indoor Positioning Algorithm”, 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Zürich, Switzerland, 15-17 September 2010.

[4] Kazem Sohraby, Daniel Minoli, Taieb Znati, “Wireless Sensor Networks Technology, Protocols, and Application”, Wiley, 2007.

[5] Kuor-Hsin Chang, "Bluetooth: a viable solution for IoT? [Industry Perspectives]", IEEE Wireless Communications, Volume: 21 , Issue: 6 , December 2014.

[6] Manuel Ramos "Characterization of LM35 Sensor for Temperature Sensing of Concrete" Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 , March 15 - 17, Hong kong Vol II, IMECS 2017.

[7] M. Niswar, A. A. Ilham, E. Palantei, R. S. Sadjad, A. Ahmad, A. Suyuti, Indrabayu, Z. Muslimin, T. Waris, and Puput Dani Prasetyo Adi, “Performance evaluation of ZigBee-based wireless sensor network for monitoring patients’ pulse status,” in Proceedings - 2013 International Conference on Information Technology and Electrical Engineering: “Intelligent and Green Technologies for Sustainable Development”, ICITEE 2013, 2013.

[8] P. Ferrari, A. Flammini, D. Marioli, E. Sisinni, A. Taroni, " A Bluetooth-based sensor network with Web interface", IEEE Transactions on Instrumentation and Measurement Volume: 54 , Issue: 6 , Dec. 2005.

[9] Soo Young Shin, Hong Seong Park, Sunghyun Choi, Wook Hyun Kwon, "Packet Error Rate Analysis of ZigBee Under WLAN and Bluetooth Interferences", IEEE Transactions on Wireless Communications, Volume: 6 , Issue: 8 , August 2007.

[10] Sparkfun, Bluetooth basics, Serial Communication and Hexadecimal learn.spurkfun.com/tutorials/Bluetooth-basics/all, 2017.

[11] Waltenegus Dargie and Christian Poellabauer “Fundamentals of Wireless Sensor Networks Theory dan Practice” Wiley Series on Wireless Communications and Mobile Computing, USA, 2010.

[12] Ying Zhang, Hannan Xiao, "Bluetooth-Based Sensor Networks for Remotely Monitoring the Physiological Signals of a Patient", IEEE Transactions on Information Technology in Biomedicine Volume: 13 , Issue: 6 , Nov. 2009.

[13] https://www.mouser.com/datasheet/2/268/rn-42-ds-v2.32r-268826.pdf access date : 21 Oktober 2018.

[14] http://www.ti.com/lit/ds/symlink/lm35.pdf access date : 12 November 2018.

[15] https://ww3.minicircuits.com/app/AN40-012.pdf access date : 24 September 2018.

[16] https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/ access date : 17 Agustus 2018.

Underwater Optical Fish Classification System by means of Robust Feature Decomposition and Analysis using Multiple Neural Networks

Mohcine Boudhane[1], Taoufiq Belhoussine Drissi[3]
GITIL laboratory
Hassan II University, FSAC
Casablanca, Morocco

Benayad Nsiri[2]
Research Center STIS, M2CS
Mohammed V University, ENSET
Rabat, Morocco

***Abstract*—Live fish recognition and classification play a pivotal role in underwater understanding, because it help scientists to control the subsea inventory in order to aid fishery management. However, despite technological progress, fish recognition systems still have many limitations on observing fish. Difficulties in visualizing optical images can arise due to external attenuation, scattering properties of water. Optical underwater imaging systems can also have detection problems such as changing appearance/orientation of objects, and changes in the scene. In this paper, we propose a new object classification system for underwater optical images. The proposed method is based on robust feature extraction from fish pattern. A specific pre-processing method is used in order to improve the recognition accuracy. A mean-shift algorithm is charged to segment the images and to isolate objects from background in the raw images. The training data is processed by Principal component analysis (PCA), where we calculate the prior probability inter-features. The decision is given using a combined Bayesian Artificial Neural networks (ANNs). ANNs will calculate non linear relationship of the extracted features, and the posterior probabilities. These probabilities will be verified in the last step in order to keep (or reject) the decision. The comparison of results with state of the art methods shows that the proposed system outperforms most of the solutions in different environmental conditions. The solution simultaneously deals with artificial and reel environment. The results obtained in the simulation indicate that the proposed approach provides a good precision to make distinguish between different fish species. An average accuracy of 94.6% is achieved using the proposed recognition method.**

Keywords*—*Fish recognition; Optical image analysis; scene understanding; principal component analysis; non-linear artificial neural networks

I. Introduction and Related Work

Oceans and seas are very fragile environments that harbor millions of plant and animal species. Morocco has a coastline stretching over 3,500 km of coastline (in the Atlantic Ocean and the Mediterranean Sea), a maritime area of about 1.2 million km^2, a fishing potential estimated by FAO at around 1.5 million tons (renewable every year) [1]. However, some species between them, victims of intensive fishing, are threatened with extinction [2]. Today, the exploitation of these resources is an obligation. Nowadays, the underwater ecosystem can be protected by utilizing different monitoring systems. Some of these systems consider the automatic observation and visualization of fish [3]. Automatic fish monitoring use different sensors such as cameras and sonars, the detection and classification of different species is done automatically instead of manual annotation. Computer vision and pattern recognition techniques offer powerful methods, which can tackle with the uge collected sensors data. Actually, the most underwater monitoring systems are based on optical images (captured by camera) and the exploitation of data. The big advantage of the use of cameras is a low cost in data gathering.

A major difficulty in processing underwater images is the attenuation of light (fig. I). This last reduces visibility range to about twenty meters in clean water and less than five meters in turbid water. In addition, images captured underwater are suffered due to many problems: first the rapid attenuation of light requires the attachment of a light source to the vehicle providing the lighting necessary [4], [5], [6]. Unfortunately, artificial lights tend to illuminate the scene in a non-uniform way producing a bright spot in the center of the image and the poorly lit surrounding area. Furthermore, many technologies have been built in this context in order to develop sophisticated systems in underwater fish detection and recognition [7], [8], [9], [10]. Authors in [11] use an automatic computer assisted by underwater video analysis for long term observation. [12] develop a fish detection method that involves the training of a classifier based on features extracted from fish and samples of other object types. Statistics about specific oceanic fish species distributions, namely discrimination and independence are given in [13] in order to aid in the feature selection process.

Choosing which features are to be used in the identification process has a major influence on the results of the study. In [14] and [15], the authors used shape and texture as features for the classification, which were derived from an active appearance model. The Principal Component Analysis (PCA) is used for fish cataloging. The classification accuracy obtained was about 76%. However, this method deal with very limited fish species group and its extracted features use only shape and texture. McGrath et al. [16] developed a fish identification system where the features are based on color, texture, and shape taken from the video sequences of four species. K-nearest neighbor is used on feature selection and fish classification. Correct classification rate about 70.6% is achieved. However, this method is limited in use as the images were taken on dead fish. Therefore, it is difficult to apply it in a real underwater environment and in real-time. Spampinato *et.al* [17] acquire underwater live fish recognition using a Balance-Guaranteed

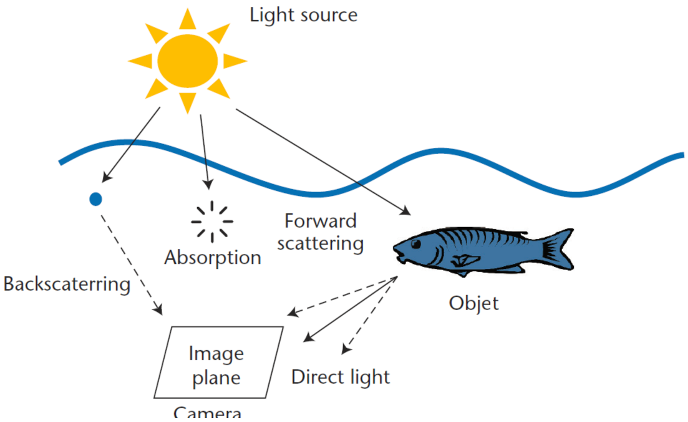


Fig. 1. Underwater environment affect seriously to the optical images.

Optimized Tree. This method extracts many features, which sum up a combination of color, shape, and texture properties from fish pattern. Those features, however, do not perform well for noisy images. The method obtain 4% improvement of the average recall (AR) compared to the flat SVM classifier. Qin *et.al*[18] propose a system based on two feature extraction techniques: SIFT and Principal PCA, and two immunological algorithms: Artificial Immune Network and Adaptive Radius Immune Algorithm. The system achieves a 92% of success rate for six species of dead fish.

In all previous works, the accuracy is based on the proportion of correct recognitions while robustness means recognizing fish in a complex environment. In this paper, we present a new method for fish recognition using cameras. We focus our study to offer the submarine biologist methods to better exploration of marine resources without prior knowledge of underwater environment. The proposed approach, instead of building techniques in order to perform the classifiers structure itself, we consider it as a blackbox and focus on the extraction of robust features.

In next, this article is organized as follows. In Section 2, system overview is described where we discuss each step of the approach under investigation. In Section 3, experimental results are shown, and Section 4 concludes this work.

II. System Overview

The main goal of fish classification is to derive methods for the cataloging of species underwater. Given a set of data from fish $\mathbf{D} = \{d_1, ..., d_n\}$, we look for its categories $\mathbf{c} = \{c_1, ..., c_m\}$. A raw data is represented by a feature vector, which will predict the fish category (Fig. 2).

The method is divided into two modes: offline and online. The role of the offline mode is to calculate the prior probabilities of the training data in order to make a preliminary distinction between species according to each feature. Offline mode is described in the next subsection The goal of the online mode is to classify fish from the scene. In this mode, the system selects the detected fish and recognizes them in real time. The online mode is discussed in Section 2.2.

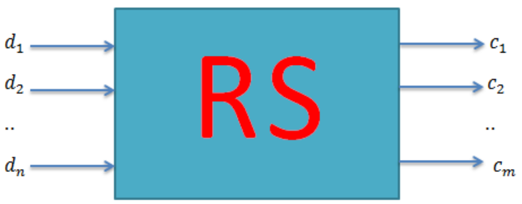


Fig. 2. RS (Recognition system).

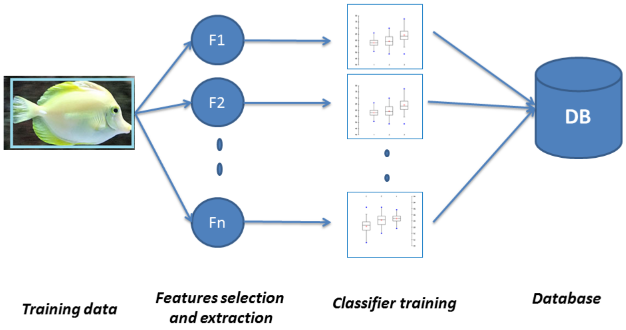


Fig. 3. Training data processing. The training data is made offline as pretreatment stage. In this stage, the prior probabilities according to each feature are calculated. $F_0, F_1, \cdots F_n$ are the features used in the offline mode. For example the l̈ength of fishïs one of the features used.

A. The Offline Mode

1) Initial probability: In this stage (Fig. 3), the dataset has to be trained. The aim of this step is to calculate the prior probability for fish species according to their features. If we consider $\mathbf{c}$ to be a set of fish categories, and m the number of categories, the initial probability of each category is defined as follows:

$$P(c_j) = \frac{1}{m} \tag{1}$$

with

$P(c_j)$ is the probability for the j-th category;

$\mathbf{c}$ = { $c_1, c_2,$ ... c_m }

c_j = $c_1, c_2,$ ... c_m; ($j \in [1, m]$)

$\sum_{j=1}^{m} P(c_j) = 1$

2) Prior probability: Unfortunately, in the practice, the previous probability is not uniform. Accordingly, the probability in (1) cannot help us to make a preliminary differentiation between species in the dataset. For this purpose, we compare the fish features according to their categories in order to find differences between them. A feature denotes a certain attribute

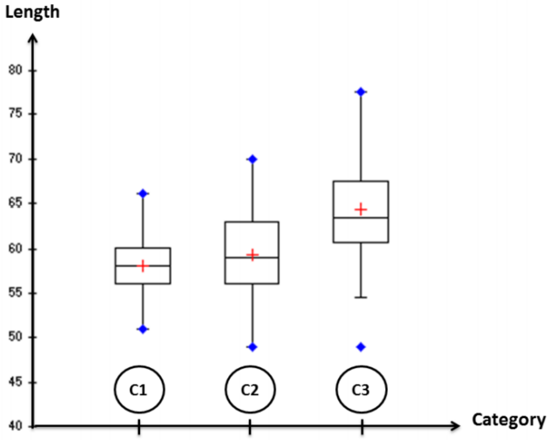


Fig. 4. box plot of the training data. The following example compares the length of three different fishes: c1-Salmon, c2-Blowfish, and c3-Parrotfish. After careful analysis, we reached the conclusion that the third specie surpasses in length the other two. Therefore, in the identification process, when we have to consider a big fish, for example, there is a higher probability it would belong to the third category.

that is considered important to describe a fish category to other one. For example: shape, length, texture, number of tails,... are some kind of characteristics (features).

In fact, the main objective of this stage is to find relevant data in many different features that deal with large quantities of information. In other words, this part take charge to calculate the probabilities of the training data according to each feature F_i, in order to make prior distinction between fish species. For example, Fig. 4 illustrates the results of a comparative analysis, based on length, of three types of fish, undertaken with the aid of box plots. Used to show overall patterns of response for a dataset, box plots provide a useful way to visualize the characteristics of a large group of different fish. The fish categories represented in Fig. 4 are: Salmon (1), Blowfish (2), and Parrotfish (3). Note that the red marked symbols illustrate the median values of each category.

As you can see in Fig. 4, the population of the third category surpasses the one in the first and second category. Moreover, the second category is nearly as big as the first one. Accordingly, the prior probability in (1) will be changed. Precisely, in case the detection process contains large fish (for example an individual fish measuring more than 60 cm), the prior probability of the categories in relation with the length will be as

$$P(c_3|length) > P(c_2|length) > P(c_1|length) \qquad (2)$$

This process is applied to all features in our fish dataset according to their categories. In next, the following subsection will deal with the online mode.

B. The Online Mode

In order to be able to identify fish using the online mode, a six stage process is proposed, as it can be seen in Fig. 5.

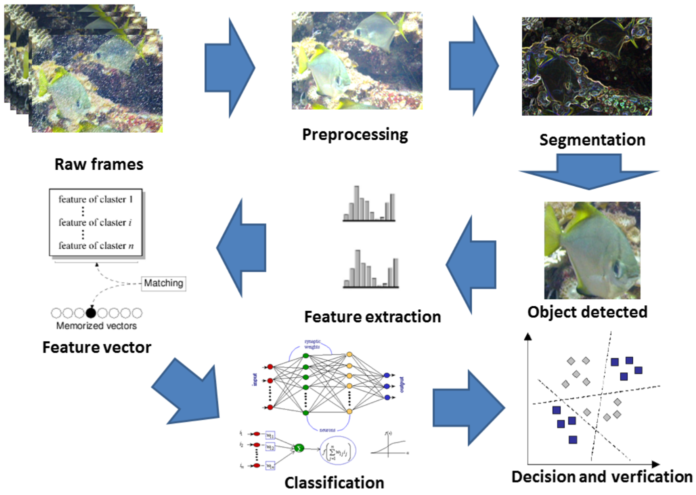


Fig. 5. Online system architecture.

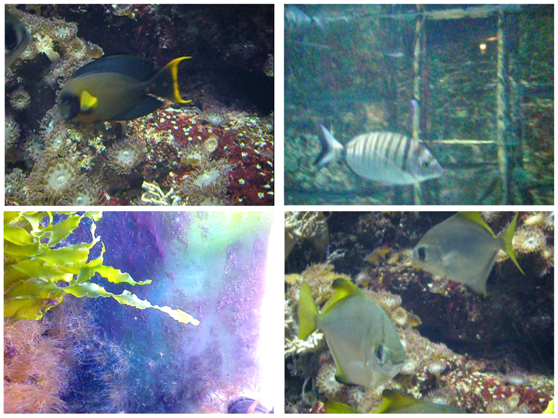


Fig. 6. Example of scenes used in the classification.

Further on, we will divide this unit into four sections. The first one, which we will entitle subsection 2.2.1., deals with the raw images recorded from the scene. These images will be preprocessed and further on segmented in order to separate the fish body from the rest of the image. subsection 2.2.2, on the other hand, deals with the preprocessing stage, while the segmentation one is discussed in subsection 2.2.3. When the fish is detected (in subsection 2.2.4), features of interest will be extracted. Then, feature vectors are generated. This process is described in Section 2.2.5. In last, subsection 2.2.6, stands as the classificatory one, when the cataloging, through a fish division (or partition) technique, using mixed neural networks, takes place.

1) Data acquisition: The underwater platform is designed to detect and record optical signals over an extended deployment. The system is also able to take in the input images and video sequences. Fig. 6 illustrates some examples of raw scenes.

2) Preprocessing: Image analysis requires to be preprocessed regardless of the level of noise that the take displays. This noise should be eliminated. The objective of this subsection is to reduce the relevant noise, by ameliorating the visual appearance of the image. Furthermore, when looking at the underwater images, we could find two major problems: smoothness/noise. For this reason, we use a novel method of image denoising and enhancement suggested in [19] in order to improve the visibilty in raw images. In [19], authors model underwater environment by two overlapping processes: a Poisson distribution, and Gaussian mixture GM distribution. It's called Poisson-Gauss mixture distribution. The distribution was defined as (in [19]):

$$p(z|\mathbf{y}) = \sum_{k=0}^{\infty} \left(\frac{x^k}{k!} e^{-x} \sum_{m=1}^{M} \alpha_m \cdot f(\mathbf{y}, \mathbf{C}_m, \mu_m) \right),$$

where

$$\begin{cases} \lambda : a > 0 \text{ real number.} \\ z : \text{Event.} \\ x : \text{Realization of Poisson-distribution.} \\ \mathbf{y} : \text{the realization of Gaussian-distribution } f. \\ m : \text{number of Gaussians distributions.} \\ C_i : \text{covariance matrix of the } i\text{-th Gaussian.} \\ \mu_i : \text{mean of the } i\text{-th Gaussian.} \\ \alpha_i : \text{the mixture coefficient.} \end{cases}$$

3) Segmentation: Mean shift is introduced by Fukunaga and Hostetler [20], this procedure, which stands as a powerful, non-parametric, iterative algorithm, was developed in order to be applied in many fields of computer vision. Its purposes are various. Mean shift associates these segments with the nearby pixels of the dataset probability density function. For each segment, it defines a window around it, and then it computes the mean of the data points. Then it shifts the center of the window to the mean and repeats the algorithm until it converges. After each iteration, the window shifts to a denser region of the image. As a result, mean shift segments images into different regions. Thereafter, the regions corresponding to fish will be extracted.

4) Feature extraction: Feature extraction is a type of dimensionality reduction that efficiently represents appealing parts of an image as a compact feature vector. This approach is useful when the size of the image is large, and a reduced feature representation is required to rapidly complete tasks as image matching and retrieval. Feature detection and extraction are often combined to solve common computer vision problems, as well as object detection, content-based image retrieval, face detection and recognition, and objects classification.

Challenges: Automatic fish recognition is a difficult undertaking. In over thirty years of research in computer vision, progress has been limited. The main challenge is the amount of variation in visual appearance. A feature detector must take into account all specific characteristics of each category, and also with the specificity of visual imagery that exists underwater. For example, fish varies in size, shape, color, and in small details—such as the shape of the head, texture, and the position or number of fins. Furthermore, the lighting, surrounding scenery, algae, and an object's position affect its appearance. A feature detection algorithm must also distinguish fish species from all other visual patterns that can occur.

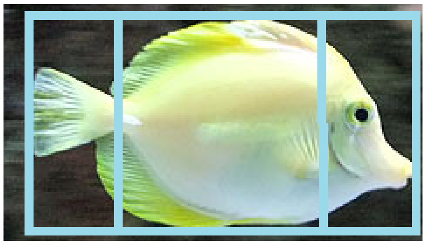


Fig. 7. Fish decomposition. The image is divided into three parts. The first part illustrates the head, the second one represents the body, and the third one the tail. The size of the body is twice the size of the head, and correspondingly the tail.

Formulation: The purpose of this stage is to extract as much information as possible from the data obtained in the previous stage. These features are expected to characterize different properties of structures and objects in each source of data. After feature extraction, a large amount of valuable information is obtained.

In this stage, the model traces useful object parts in each testing image. We split the fish body into three categories: head, middle, and tail. Fig. 7 depicts an example of this distribution. Size, orientation, shape, color, texture, and specific features of each part are extracted as a main group of features. As its name indicates it, the particular features represent a specific characteristic of a certain part of the fish body. In other words, it defines a special feature related only to a specific part of the fish's body. For example, the "position of the mouth" is a particular characteristic which is situated only in the head part. It can be terminal, inferior or superior. Therefore, the feature vectors will not be identical. These categories are later on used as class labels in the classification stage. In addition, we will study each part of the fish body. There are four major scenarios (or parts) that we desire to study: a whole body, fish tail, middle part, and fish head.

a. First scenario: *The whole body*

In the first scenario, we have set out to determine a large set of features (Fig. 8). For each fish species, we have gathered more than fifty different characteristics that are divided into four groups. These groups and their corresponding numbers of features are the following:

$$\begin{cases} \text{Size (8 features)} & \textbf{e.g: } \textit{Area, length.} \\ \text{Shape (20 features)} & \textbf{e.g: } \textit{Moment, aspect ratio.} \\ \text{Color (9 features)} & \textbf{e.g: } \textit{luminance, chrominance.} \\ \text{Texture (16 features)} & \textbf{e.g: } \textit{Inertia, energy.} \end{cases}$$

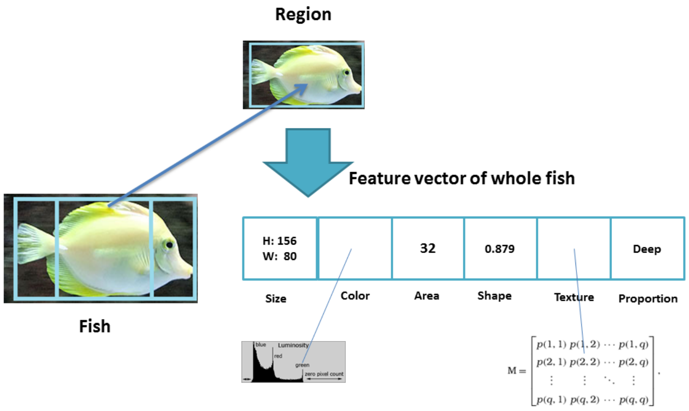


Fig. 8. Feature extraction from a whole body of fish. This example shows a part of the feature vector in accordance with the whole body of the fish. Most of the features extracted here, are also used in the other scenarios.

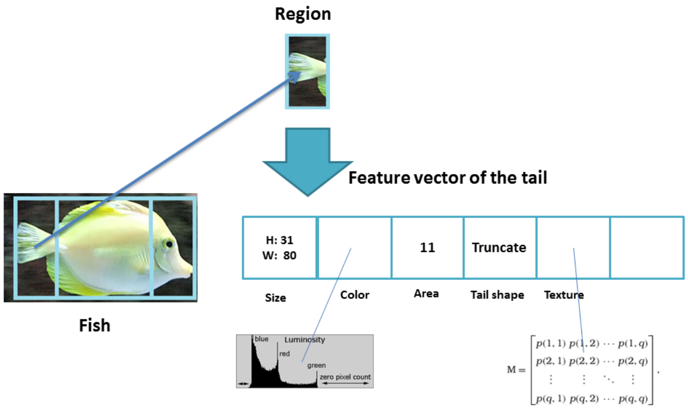


Fig. 9. Feature extraction from fish tail. We note that the feature vector does not resemble the others.

The gathering of these features is performed in all scenarios. In the other scenarios, we will approach only the extra-features.

b. Second scenario: *Fish tail*

This section refers to the “fish tails” particularities, such as the shape. There are different kinds of tail shapes: heterocercal, forked, lunate, emarginated, truncated, rounded, or pointed. These details play a critical role in the differentiation of species, and it can improve the final decision. Fig. 9 shows an excerpt from the feature vector for the tail.

c. Third scenario: *Middle part*

The middle part is very important because it contains the texture and some fish fins. In this part, we will focus our attention on the number of fins, visual appearance, and texture. The features discussed in the first scenario will be used as well. Fig. 10 shows a part of the feature vector concerning the middle part.

d. Fourth scenario: *Fish head*

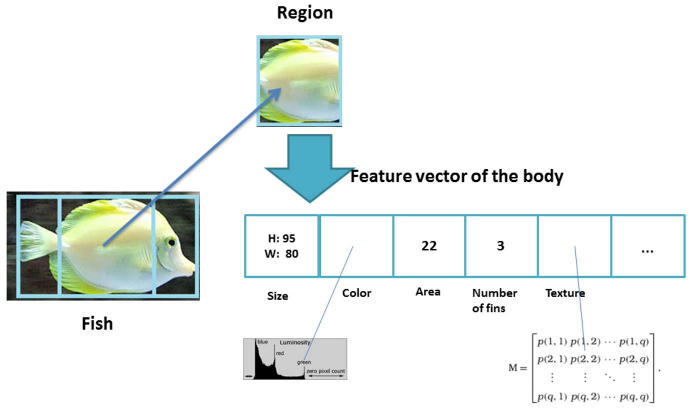


Fig. 10. Feature extraction from the middle part.

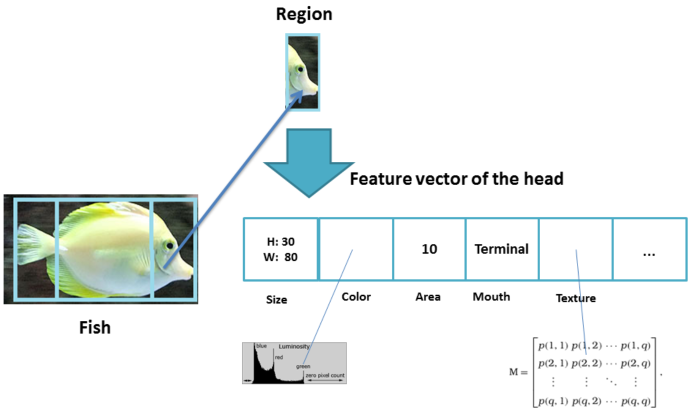


Fig. 11. Feature extraction from the head.

The head of the fish displays specific features, as well. For example, “the position of the mouth” can be: terminal (mouth oriented in the middle), superior (mouth oriented upwards), or inferior (mouth oriented downward). Fig. 11 illustrates an excerpt of the feature vector of the head.

5) Fish identification: It is performed using Artificial Neural networks (ANNs) [21]. Ann defines interaction between elements (nodes or neurons). A special function is responsible to compute the output of nodes. In the practice, ANNs are composed of three main layers, namely, the input, hidden and output layer (Fig. 12):

- **Input layer:** They take as input n real values $I_1, I_2, ..., I_n$. These values represent feature vector of part of fish.

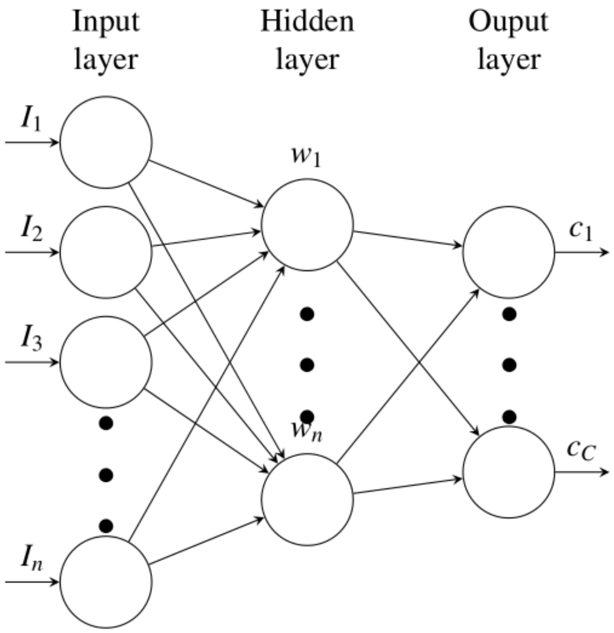


Fig. 12. ANN architecture.

- **Hidden layer:** It is the feature combination function which calculates the summation $\sum_{i=1}^{n} I_i w_i$ where w_i is the input weight $w_1, w_2, ..., w_n$.
- **Output layer:** It's the activation function, in which we compute the posterior probabilities of fish categories.

layers are composed by one or more nodes. It's symbolized as small circles. The flow of information, from one node to the other, is figured by oriented lines.

III. Classification Results

The classification was carried out using the ANNs combined artificial neural network. Fig. 13 shows the processing operation for input features according to fish sub-divisions, where each body part of a fish is treated independently. In which each region will be analyzed according to the input values obtained by the feature vector of each part (cluster). Subsequently, the resulting probabilities obtained by each of them will be combined to have a value. The latter will be considered as the overall decision that presents the final decision.

The main purpose of this section is to extract pertinent information from each body part in order to obtain final decision by combining each ANN output (final probability of each Neural network). The ANN model can be represented as follows:

$$y_t^* = w_1 y_1 + w_2 y_2 + \cdots + w_n y_n. \tag{3}$$

where

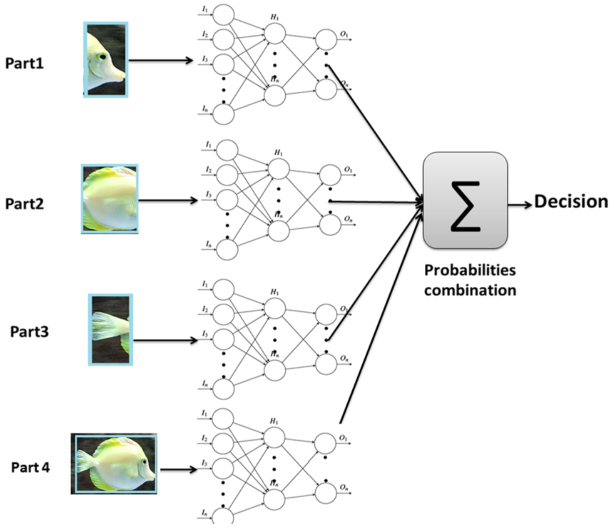


Fig. 13. The proposed recognition system architecture.

TABLE I. Hypothesis according to the feature "*shape of the head*".

Hypothesis **k**	Shape of tails **Hypothesis** H_k
1	Heteroceral
2	Forked
3	Lunate
4	Emarginate
5	Truncate
6	Rounded
7	Pointed

- t is the aggregated single network model prediction.
- w_k is the combination weight of the k-th network.

In order to build data fusion of several ANNs, we define a combination function after getting the output posterior probabilities (from each ANN) (in Fig. 13). It can be drawn from the formula of Bayesian probability combinations.
Let us suppose that we have H_1, H_2, $\cdots$ H_K in the Bayesian inference process, where H_K represents hypothesis that can illustrate an observation E (or an event). Thus, the posterior probability of hypothesis H_k being true, given the evidence E, $P(H_k|E)$, can be defined as in (4).

$$P(E|H_j) = \frac{P(H_k|E)P(H_j)}{\sum_{j=1}^{K} P(H_j|E)P(H_j)} \tag{4}$$

where $P(H_j)$ is the prior probability of hypothesis H_j being true and $P(E|H_j)$ is the probability of observing evidence E given that H_j is true.

Table I and Table II show hypothetical examples based on the "head shape" of the fish. In these examples, we note that the feature "shape of the head" is symbolized as f_0, and we

TABLE II. CATEGORIES AND THEIR INITIAL PROBABILITIES.

Category i	Name of category C_i	Initial probability $P(C_i)$
1	Salmon	$\frac{1}{3}$
2	Blowfish	$\frac{1}{3}$
3	Parrotfish	$\frac{1}{3}$

investigate its values in relation to the tail's appearance, which, in these situations, is "forked" (according to the data gathered in Table I, it means $f_0 = 2$). Equation number (5) shows the application of calculation (4) using the example above. This equation is very useful when building the probability of each part.

$$P(f_0 = 2|Salmon) = \frac{P(Salmon|f_0 = 2)P(f_0 = 2)}{P(Salmon)} \quad (5)$$

In this approach, the conclusion is based on the values of the posterior probability of individual networks. In one hand, the maximum posterior probability is attributed a weight of 1, and in the other hand the other networks are attributed weights of 0. Given raw fish, we split them into a set of parts, each consisting of a group of variables. We then consider these parts as statistically independent. Bearing this assumption, the posterior probability of a network is calculated in the equation (6) ([22]).

$$P_{fish} = \prod_{i=1}^{N} P(part = i) \quad (6)$$

where

$$\begin{cases} P(part = i) & \text{or } P(part_i) \text{ is a global probability of the } i^{th}\text{part.} \\ N & \text{is the number of fish parts.} \end{cases}$$

We assume the same prior probabilities for all the parts that make out the body of fish. The combination intends to know the probability of the parts of fish, given the observation on categories $\mathbf{c}_j$ as $P(part_i \mid c_1, c_2, \cdots, c_m)$ $i \in [1, N]$. We presume that $part_i$ is independent according to its categories, so the posterior probability is defined as:

$$P(part_1, \cdots part_N|c_1, \cdots, c_m) = \prod_{i=1}^{N} P(part_i|c_1, c_2, \cdots, c_m). \quad (7)$$

We assume that the categories are strictly independent, and then we obtain for each i:

$$P(part_i \mid c_1, \cdots, c_m) = P(part_i \mid c_1) \text{ x } \cdots P(part_i \mid c_m). \quad (8)$$

$$P(part_i \mid c_1, \cdots, c_m) = \prod_{j=1}^{m} P(part_i \mid c_j). \quad (9)$$

After simplification, we obtain:

$$P(part_i \mid c_1, \cdots, c_m) = \prod_{j=1}^{m} \frac{P(c_j \mid P(part_i) \text{ x } P(part_i)}{P(c_j)} \quad (10)$$

Based on precedent equations, we deduce that the global probability of the whole fish can be given as

$$P(fish \mid c_1, \cdots, c_m) = P(part_1, \cdots part_N \mid c_1, \cdots, c_m). \quad (11)$$

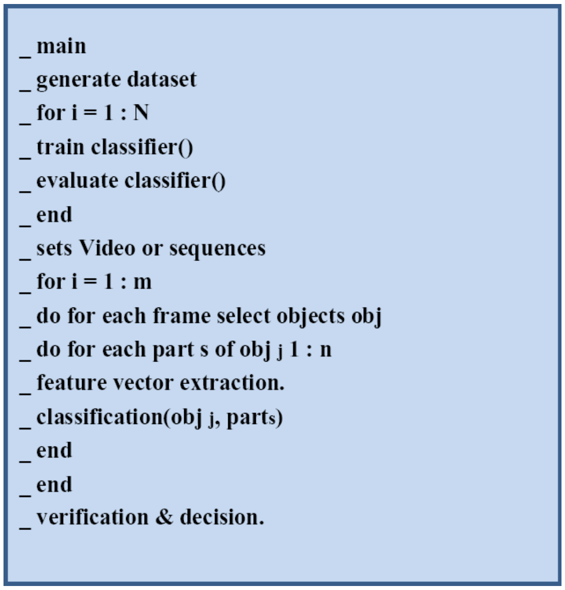
and from (7) and (10), we conclude:

$$P(fish \mid c_1, \cdots, c_m) = \prod_{i=1}^{N} \prod_{j=1}^{m} \frac{P(c_j \mid P(part_i) \text{ x } P(part_i)}{P(c_j)} \quad (12)$$

$$\text{Choose } c_j \text{ if : } \begin{cases} j \in [1, m]. \\ c_j = \max_{j \in [1,m]} P(fish \mid c_j) \end{cases}$$

- $P(fish \mid c_1, \cdots, c_m)$ is the probability of the target being one of the categories (c_j), given the observation from the whole fish.
- $P(part_i \mid c_1, \cdots, c_m)$ is the probability of the target being one of the categories (c_j), given the observation from the part i.
- $P(part_i)$ is the probability of the target being of the part i.
- $P(c_j)$ is the probability of the target being one of the categories.

In addition, two datasets sequences are randomly picked from each species and altogether they are compiled of several images. When running the classification algorithm, a set of fish categories can be recognized. The algorithm used in the proposed approach is described in algorithm 1.



```
_ main
_ generate dataset
_ for i = 1 : N
_ train classifier()
_ evaluate classifier()
_ end
_ sets Video or sequences
_ for i = 1 : m
_ do for each frame select objects obj
_ do for each part s of obj j 1 : n
_ feature vector extraction.
_ classification(obj j, parts)
_ end
_ end
_ verification & decision.
```

TABLE III. CLASSIFICATION ACCURACY

c_i	Classification Accuracy
Salmon 1	99.44%
Blowfish 2	95.51%
Parrotfish 3	99,7%

Algorithm 1: The proposed pseudo code algorithm.

The application of classification is usually a trade-off between low numbers of false alarms (false positives FP) and high numbers of correct detections (true positives TP). According to statistics, the numbers of true and false positives are described using the following formula:

$$\textbf{Detection rate} = \frac{1}{l}\sum_{i=1}^{l}\frac{TP_i}{TP_i + FN_i}. \qquad (13)$$

where l is the number of classes.

Different combinations of both measures can be plotted as ROC (Receiver Operating Characteristic). The number of true negative TN is not exactly defined. The amount of false detection can be produced. Furthermore, it is often not possible to estimate the correct number of background objects. For this reason, another measure is used to describe the number of false alarms:

$$\textbf{False detection rate} = \frac{1}{l}\sum_{i=0}^{l}\frac{FP_i}{FP_i + TP_i}. \qquad (14)$$

A total accuracy, which is defined as the percentage of correctly classified fish. Usually, accuracy is represented as a real value between 0 and 1.

$$\text{Accuracy} = \frac{\text{Number of correct decisions}}{\text{Number of total decisions}} \qquad (15)$$

The first evaluation only assesses the performance of our fish detection system, which consists of two classifiers. When evaluating each class separately, the classification performance for each class could be concluded:

The individual class precision/recall is represented in Fig. 14. The suggested approach achieves a good accuracy in fish distinction. As shown in Table. III, an accuracy average of 94.6% (dataset1: 99.4% and dataset2: 89.9%), and the matching speed of 0.26 seconds for several test sample images have been obtained by the proposed approach.

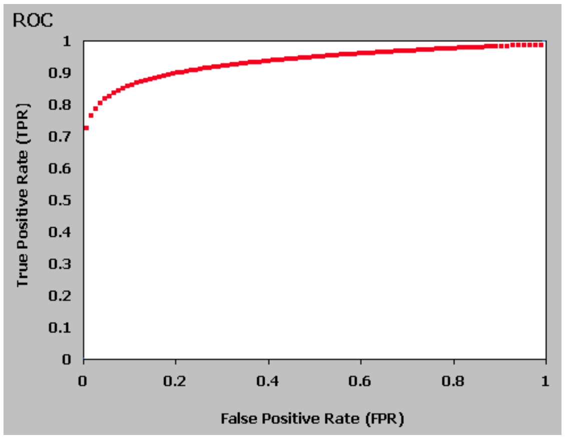


Fig. 14. ROC diagram of our classification system.

TABLE IV. CLASSIFICATION RESULTS OF THE PROPOSED METHOD COMPARED TO: ADA-BOOST (PROPOSED BY *Hsu et al.*[23]), AND HIERARCHICAL TREE (PROPOSED BY *Fan et al.*[24])

Method	d.set 1	d.set 2	Overall
(Ada-boost Hsu *et.al*[23])	87.3%	71.2%	79.2%
(Hierarchical tree Fan *et.al*[24])	95.7%	83.6%	89.7%
The proposed approach	**99.4%**	**89.9%**	**94.6%**

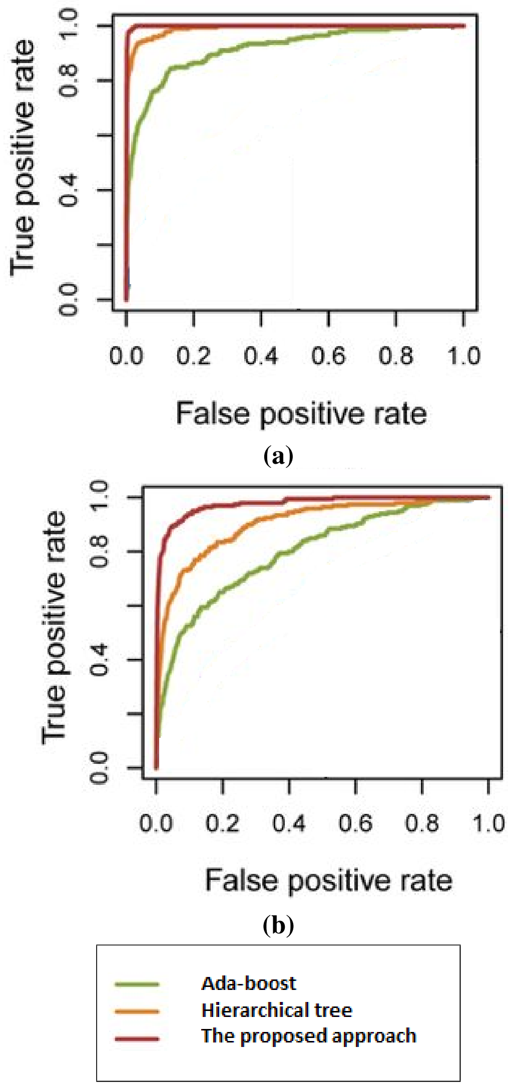
IV. DISCUSSION

In order to measure the robustness of our method, the proposed approach is tested in two environmental conditions: clair and turbid environment. The proposed approach is also compared with two methods: Ada-boost (proposed by Hsu et.al[23]) and hierarchical tree (proposed by Fan et al. [24]), and applied to the same data. As shown in Table IV, we note that our method is about 5% better than hierarchical tree. The average was about 94.6% using the proposed approach against 89.7% by hierarchical tree. The Ada-boost method achieves 79.2% accuracy. Furthermore, in many applications, the ROC curve provides more interesting information about the quality of learning than just the error rate. Fig. 15 (a) and (b) show ROC diagram of the three methods on the database 1 and 2. The curves in red represents the proposed method, in orange the hierarchical tree, and in green the Ada-boost method, respectively. As shown in Fig. 15, curves in "green" and "orange" are lower than the "red" one. It means that the number of true positive is important compared to the others methods. As results,we conclude that the proposed approach outperforms these two methods.

**d.set1: dataset 1 "clair environement",*
**d.set2: dataset 2 "turbid environement".*

V. CONCLUSION

The present paper described an algorithm employed in fish classification based on robust fish division/sectioning/partition and feature extraction. Our goal was to develop a system that



(a)

(b)

Fig. 15. ROC diagram, (a): using dataset1, (b): using dataset2.

detects and recognizes an array of different types of fish in images and videos, including various structures in underwater images. Each segment (or part) forms a sub-classifier and represents various local properties of the fish. Bayesian mixed ANNs are used to classify each part. Probabilities are given by each ANN and they were combined and recalculated in order to reach a conclusion. Experimental results show a fine accuracy in the fish classification system. We note that the observed precision of 94.6% may be considered a satisfactory estimation, as it provides a reasonable approximation of fish classification, the varying environmental conditions in an open unconstrained space and the changeable status of the sensors used. However, the use of static features could limit the application of the proposed algorithm. In the future work, we plan to build a novel fish recognition system using deep learning methods, that can provide more dynamicity and flexibility in the feature selection process.

REFERENCES

[1] OCEANA, ”Oceans in danger”, PLAZA ESPAÑA-LEGANITOS, 47 28013 MADRID SPAIN, **Link:** https://eu.oceana.org/sites/default/files/reports/oceans_in_danger.pdf.

[2] Gilles Boeuf, ”La biodiversité, de l’océan à la cité”, Paris, Fayard, Collège de France, coll. Leçons inaugurales, 2014, pp:84, ISBN:978-2-213-68148-1.

[3] J. Jaffe ”Underwater Optical Imaging: The Past, the Present, and the Prospects”, IEEE Journal of Oceanic Engineering, Vol:40, Issue:3,(2015), pp:683-700.

[4] Yu.I. Troutskaya; V.V. Bakhanov; S.A. Ermakov, ”Underwater wave phenomena in the upper ocean and their surface manifestation: Role of nonlinearity”, (2006), IEEE US/EU Baltic International Symposium, Year:2006, pp:1-6.

[5] Donna M. Kocak, Frank M. Caimi ”The Current Art of Underwater Imaging - With a Glimpse of the Past and Vision of the Future” Marine Technology Society Journal. 39(3):5-26.

[6] Joaquín Aparicio, Ana Jiménez, Jesús Ureña, Fernando J. Álvarez, ”Realistic modeling of underwater ambient noise and its influence on spread-spectrum signals”, OCEANS 2015 - Genova, pp. 1-6, 2015.

[7] Xin Sun, Jianping Yang, Jianping Yang, Changgang Wang, Changgang Wang, Junyu Dong, Junyu Dong, Xinhua Wang, Xinhua Wang, ”Low-contrast underwater living fish recognition using PCANet”, Proc. SPIE 10615, Ninth International Conference on Graphic and Image Processing (ICGIP 2017), 106150Y (10 April 2018); doi: 10.1117/12.2302695; https://doi.org/10.1117/12.2302695

[8] P.x. Huang, B.J. Boom, R.B. Fisher, (2012), ”Underwater Live Fish Recognition using a Balance-Guaranteed Optimized Tree”, Asian Conference on Computer Vision ACCV 2012: Computer Vision – ACCV 2012 pp 422-433.

[9] Huang P.X., Boom B.J., Fisher R.B. (2013) Underwater Live Fish Recognition Using a Balance-Guaranteed Optimized Tree. In: Lee K.M., Matsushita Y., Rehg J.M., Hu Z. (eds) Computer Vision – ACCV 2012. ACCV 2012. Lecture Notes in Computer Science, vol 7724. Springer, Berlin, Heidelberg

[10] Stefan Wender, Klaus C. J. Dietmayer. ”A Feature Level Fusion Approach for Object Classification”, 2007, IEEE Intelligent Vehicles Symposium, 2007, ISSN: 1931-0587.

[11] Jacopo Aguzzi, Antoni Mànuel, Fernando Condal, Jorge Guillén, Marc Nogueras, Joaquin del Rio, Corrado Costa, Paolo Menesatti, Pere Puig, Francesc Sardà, Daniel Toma, Albert Palanques, ”The New Seafloor Observatory (OBSEA) for Remote and Long-Term Coastal Ecosystem Monitoring ”, Sensors 2011, 11, 5850-5872; doi:10.3390/s110605850.

[12] M. Boudhane, B.Nsiri, H. Toulni, ”Optical fish classification using statistics of parts”, (2016) International Journal of Mathematics and Computers in Simulation, vol:10, pp:8-12, ISSN: 1998-0159.

[13] Lian Li ; Jinqi Hong, ”Identification of fish species based on image processing and statistical analysis research,” in Proc. IEEE IITA’08, Qingdao, China, (2008), pp:346-50.

[14] R. Larsen, H. Olafsdottir, and B. Ersboll. ”Shape and texture based classification of fish species”. In Proceedings of the 16th Scandinavian Conference on Image Analysis, SCIA ’09, pages 745–749, Berlin, Heidelberg, 2009. Springer-Verlag.

[15] M. T. A. Rodrigues, F. L. C. Padua, R. M. Gomes and G. E. Soares, ”Automatic fish species classification based on robust feature extraction techniques and artificial immune systems,” 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), Changsha, 2010, pp. 1518-1525.

[16] R. McGrath, (2011), ”Learning to recognise fish”. thesis University of Edinburgh,2011.

[17] C. Spampinato, D. Giordano, R. Di Salvo, Y. Chen-Burger, Robert B. Fisher, G. Nadarajan ”Automatic fish classification for underwater species behavior understanding”. In Proceedings of the first ACM international workshop on ARTEMIS ’10, pp45–50, New York, USA, 2010.

[18] Qin H., Li X., Liang J., Peng Y., Zhang C. (2016). DeepFish: Accurate underwater live fish recognition with a deep architecture. Neurocomputing, vol:187, pp:49–58. doi:10.1016/j.neucom.2015.10.122

[19] M. Boudhane, S. Badri-Hoeher, B. Nsiri ”Optical fish estimation and detection in noisy environment” IEEE Oceans - St. John’s, 2014. pp1-6.

[20] K. Fukunaga, L. Hostetler, (1975), ”The estimation of the gradient of a density function, with applications in pattern recognition”, IEEE Trans. Information Theory, vol:21, number:1, pp:32-40, 1975.

[21] E.A Wan, ”Neural network classification: a Bayesian interpretation” IEEE Transactions on Neural Networks (1990), (Volume:1, Issue:4), pp:303-305.

[22] S. Kiartzis; A. Kehagias; G. Bakirtzis; G. Bakirtzis; V. PETRIDISV; ”Short term load forecasting using a Bayesian combination method”, (1997), International Journal of Electrical Power and Energy Systems 19(3), pp:171-177, DOI:10.1016/S0142-0615(96)00038-5

[23] Ya-Wen Hsu ; Yue-Sheng Ciou ; Jau-Woei Perng, (2017) ”Object recognition system design in regions of interest based on AdaBoost algorithm” 20th International Conference on Information Fusion, Year:2017, pp:1-5.

[24] Jianping Fan ; Tianyi Zhao ; Zhenzhong Kuang ; Yu Zheng ; Ji Zhang ; Jun Yu ; Jinye Peng (2017) ”HD-MTL: Hierarchical Deep Multi-Task Learning for Large-Scale Visual Recognition”, IEEE Transactions on Image Processing, Year:2017 , Vol:26 , Issue:4, pp:1923-1938.